Title

Genomic analysis reveals geography rather than culture as the predominant factor shaping genetic variation in northern Kenyan human populations

Running title

Genetics and culture of Kenyan pastoralists

Authors

Angela M. Taravella Oill[1,2], Carla Handley[3], Emma K. Howell[1,2], Anne C. Stone[2,3,4], Sarah Mathew[3,4*], and Melissa A. Wilson[1,2*]

*Co-corresponding authors

Affiliations

1. School of Life Sciences, Arizona State University, Tempe, AZ 85287 USA

2. Center for Evolution and Medicine, Arizona State University, Tempe, AZ 85287 USA

3. School of Human Evolution and Social Change, Arizona State University, Tempe, AZ 85287 USA

4. Institute of Human Origins, Arizona State University, Tempe, AZ 85287, USA

Abstract

Objectives: The aim of this study was to characterize the genetic relationships within and among four neighboring populations in northern Kenya in light of cultural relationships to understand the extent to which geography and culture shape patterns of genetic variation.

Materials and Methods: We collected DNA and demographic information pertaining to aspects of social identity and heritage from 572 individuals across the Turkana, Samburu, Waso Borana, and Rendille of northern Kenya. We sampled individuals across a total of nine clans from these four groups and, additionally, three territorial sections within the Turkana and successfully genotyped 376 individuals.

Results: Here we report that geography predominately shapes genetic variation within and among human groups in northern Kenya. We observed a clinal pattern of genetic variation that mirrors the overall geographic distribution of the individuals we sampled. We also found relatively higher rates of intermarriage between the Rendille and Samburu and evidence of gene flow between them that reflect these higher rates of intermarriage. Among the Turkana, we observed strong recent genetic substructuring based on territorial section affiliation. Within ethnolinguistic groups, we found that Y chromosome haplotypes do not consistently cluster by natal clan affiliation. Finally, we found that sampled populations that are geographically closer have lower genetic differentiation, and that cultural similarity does not predict genetic similarity as a whole across these northern Kenyan populations.

Discussion: Overall, the results from this study highlight the importance of geography, even on a local geographic scale, in shaping observed patterns of genetic variation in human populations.

Keywords

Africa, Kenya, geography, culture, social organization, genetic structure, genetic F_{ST}, cultural F_{ST}

Introduction

Among human populations, both geography and culture contribute to modifying patterns of genetic variation. Gene flow can be constrained by geographic distance (Wright 1943). In humans, it is commonly observed that as geographic distance between populations increases, genetic similarity decreases (e.g.,(Manica, Prugnolle, & Balloux, 2005; Novembre et al., 2008; Ramachandran et al., 2005). In addition to geography, genetic variation and population structuring are also influenced by cultural factors, like language (e.g., (Hunley et al., 2008; Nettle & Harriss, 2003; Pagani et al., 2012; Sun et al., 2013; Xu et al., 2010)) or social organization (Bose, Platt, Parida, Drineas, & Paschou, 2021; Chaix et al., 2007; Heyer et al., 2009; Marchi et al., 2017). Like geographic distance, linguistic distance has been shown to correlate with genetic distance (Cavalli-Sforza, Piazza, Menozzi, & Mountain, 1988; Nettle & Harriss, 2003).

Large scale research efforts have aimed to curate genetic variation across some African populations to understand population history and human health and disease (e.g., (1000 Genomes Project Consortium et al., 2015; Choudhury et al., 2020; Gurdasani et al., 2015; Mulindwa et al., 2020; Tishkoff et al., 2009)). However, with more than 2,000 ethnolinguistic groups across the continent, there is still much to learn about the determinants of substructure both among and within individual ethnolinguistic groups at smaller geographic scales. For example, do cultural boundaries play an important role in shaping gene flow among neighboring groups, or does geography primarily shape patterns of genetic variation, even on a local scale?

The Turkana, Samburu, Waso Borana, and Rendille are neighboring pastoral ethnolinguistic groups inhabiting the semi-arid northern region of Kenya**.** They herd cattle, camel, sheep and goat, and migrate over varied distances to access pasture and water. There is intense competition for scarce dry season grazing and water resources in this region, and armed livestock raiding occurs especially between communities belonging to different ethnolinguistic groups (Handley & Mathew, 2020). Marriage across ethnic boundaries particularly between the Rendille

and Samburu has been noted (Spencer, 2012). Nomadic pastoralism in all four populations involves lots of movement of people over large distances, including into one another's territory. While there are no cultural prescriptions against intermarriage between these groups, historical and contemporary livestock raiding and resource competition often lead to hostile relationships between these groups. It is currently unclear whether these tensions contribute to isolation among these groups, or whether gene flow occurs despite the socio-political barriers.

All four groups have lineage-based divisions where individuals are organized into clans and, for some groups, clans are further grouped into either moieties or phratries (**Figure S1**). The Turkana, Samburu and Rendille are exogamous at the clan level while the Borana are exogamous at the moiety level, so for these groups, individuals generally marry outside of their birth clan. In addition to lineage-based divisions, the Turkana are unique from the other three groups in that they also have territory-based divisions that cross-cut clan-level organization. There are no marriage restrictions at the territorial level. This territorial division provides an opportunity to investigate whether territory-based division impacts substructuring in the Turkana. Detailed descriptions of the social organization of these populations can be found in (Handley & Mathew, 2020).

Each of these groups has a patrilineal descent system where an individual’s natal clan affiliation typically follows that of their father. There are, however, situations in which children do not take on the clan identity of their biological father. When the biological father is not officially married, i.e., has yet to pay a bride price to the family of the child’s mother at the time of birth, the child remains affiliated with the natal clan, which is the child’s mother’s and maternal grandfather’s clan. Households with few children or that have relatively higher material security may adopt a child, in which case the child takes the clan identity of the adoptive father. Therefore it is unclear how cultural systems of patrilineal descent shape patterns of male-specific genetic variation within these groups.

In a previous study, two members of this research team sampled 750 individuals from nine clans across these four ethnic groups, and in the Turkana additionally included three territorial sections, to obtain data on cultural beliefs and norms, and to quantify levels of cultural differentiation (cultural F_{ST}) among these groups (Handley & Mathew, 2020). Cultural F_{ST} is the proportion of the total variation in cultural traits that lie between populations (Bell, Richerson, & McElreath, 2009; Handley & Mathew, 2020) and can provide a quantitative measure of cultural similarity between groups. For the current study, we sampled individuals from these same nine clans and three Turkana territorial sections, which allows us to examine the relationship between genetic and cultural differentiation.

To form a better understanding of the population structure among the Turkana, Samburu, Rendille, and Waso Borana ethnolinguistic groups and how geography and culture contribute to shaping genetic variation in northern Kenya, we worked with these local groups to obtain genetic samples from 572 individuals across all four populations (**Table 1**). We were able to successfully genotype 376 of the 572 individuals on Illumina's Multi-Ethnic Global Array (**Table S1**). For all samples, we additionally collected culturally relevant demographic information that included natal and post-marital affiliations and spoken languages for themselves, parents, and grandparents. For married men, we additionally collected demographic information for their spouse(s) (e.g. spouse's ethnic group and natal clan affiliation, etc.). We report here that geography predominately shapes genetic variation within and among human groups in northern Kenya. Specifically, we found a clinal pattern of genetic variation that mirrors the overall geographic distribution of the individuals we sampled. We found evidence of gene flow and relatively higher rates of intermarriage between the Samburu and Rendille than between any other pair of groups in our sample. We further observed strong recent genetic substructuring among the Turkana, based on territorial section affiliation, that did not affect the between-ethnolinguistic group comparisons. Within ethnolinguistic groups, we found that male Y chromosome haplotypes do not consistently cluster by natal clan affiliation. Finally, we found that ethnolinguistic groups that are

geographically closer have lower genetic differentiation, and that cultural similarity (estimated via cultural F_{ST}) does not predict genetic similarity as a whole across these four northern Kenyan populations. Overall, despite cultural and linguistic differences, our analysis suggests that geography is the main driving force of genetic variation, even on a very local geographic scale.

Ethnolinguistic group	**Language genus[a]**	**Spoken language**	**Social organization**	**Approx. population in Kenya[b]**	**Clans sampled (sample size)[d]**
Borana	Lowland East Cushitic	Southern Oromo	2 exogamous moieties 17 clans	276,236[c]	Noonituu (40) Warrajidaa (40) Other (30)
Rendille	Lowland East Cushitic	Kirendille	2 phratries 9 exogamous clans	96,313	Ldupsai (43) Saale (45) Other (22)
Samburu	Nilotic	Northern Maa	2 phratries 8 exogamous clans	333,471	Lpisikishu (40) Lukumai (42) Other (27)
Turkana	Nilotic	Kiturkana	18 territorial sections (TS) 24 exogamous clans cross-cutting the TS	1,016,174	Kwatela TS: Ngisiger (24), Ngipongaa (21), Ngidoca (23) Ngiyapakuno TS: Ngisiger (24), Ngipongaa (26), Ngidoca (22) Ngibochoros TS: Ngisiger (21), Ngipongaa (22), Ngidoca (21) Other (38)

Table 1. Background information on study populations and collected samples. We sampled a total of 572 individuals in northern Kenya. Here, we describe general background information on the four study populations, including information on language, social organization, and population sizes in Kenya. This table was adapted from (Handley and Mathew 2020) but altered to reflect sample sizes collected in the current study. [a]Language information reported here is based on assignments from the *World Atlas of Languages (WALS) Online*. [b]Population sizes reported here were obtained from the 2019 census report of the *Kenya National Bureau of*

Statistics. [c]Borana extend into Ethiopia so their total population exceeds the numbers living in Kenya. [d]Sample numbers are based on natal affiliations. Since we opportunistically sampled in these regions, we also sequenced individuals beyond the targeted clans and territories and these samples are marked as "Other" here. One individual was from the Gabbra ethnic group and not included in this table.

Materials and Methods

Community engagement and ethics

Both SM and CH have worked in Northern Kenya for over a decade and have established and maintained a strong relationship with the local communities. Research with the Turkana, Borana, Rendille, and Samburu has expanded to include genetic analyses and great care has been taken to ensure ethical informed consent, data collection, outreach communication, and data sharing.

Subsequent to obtaining the appropriate research permitting through Kenya's National Commission for Science, Technology and Innovation (NACOSTI), yet prior to commencing any data collection, the field teams spent a considerable period of time with each local community and its leaders to sensitize individuals to the purpose and process of collecting genetic samples for this study. Other than SM and CH, field teams were composed solely of individuals from the participant communities, many of whom had been working with SM and CH for several years. Research assistants (RAs) and guides worked within their own ethnic groups, therefore having one team per group, and all information was presented to participants using local languages. As a key goal of any informed consent process, potential subjects must demonstrate sufficient comprehension of the methods and underlying scientific principles on which to base their decision to participate. With literacy rates for northeastern Kenya estimated below 10% of the adult

population (Kebathi, 2008) explaining fundamental concepts regarding DNA, genetic data sample collection, and data sharing was of paramount importance. For each area surveyed, teams met with the responsible county/deputy commissioners, local/assistant chiefs, and/or community elders councils to explain the purpose of the study and to obtain permissions from the appropriate bodies. As data collection began, RAs discussed the study, its purpose, methodologies, and underlying scientific principles to each eligible participant and provided ample time for participants to ask questions and address any concerns. At this time, there was an estimated drop-out rate of 20% of eligible participants. For those remaining, we transitioned to the formal consent process, where the objectives, methods, and benefits of the study would be repeated before asking a subject to sign or mark the consent form. At this stage, there was an additional estimated 15% drop-out of eligible participants. Those who agreed to take part in the study were provided with the contact information for both local and foreign research team members in the event that s/he should choose to be removed from the study at any future point. Furthermore, despite the common practice in many locations of husbands granting permission for themselves and for their wives, permissions were obtained explicitly and directly from all female participants. However, the research teams made efforts to avoid households where directly soliciting female participation could transgress cultural norms and inadvertently introduce additional domestic concerns.

Once obtaining consent for participation, research assistants demonstrated the cheek cell swab collection procedure, using a clean swab on themselves to scrub the inside of their own cheeks. Our initial intention was for RAs to swab the participants' mouths; however, we found that participants felt more comfortable being in control of the process to swab their own mouths. This required oversight from the RAs to ensure that the swab was oriented in the appropriate direction, did not come into contact with foreign bodies, and that enough pressure and effort were applied during the collection. Furthermore, we requested that participants rinse their mouths with water if they recently had been chewing tobacco or other organic products. Satisfactory swabs were handed to the RAs to seal within the collection tube and returned to CH for cool bag storage. After

sample collection, participants were asked to respond to a 10-15 minute survey, developed in ONA and implemented in the field through Online Data Kit (ODK) using handheld Samsung tablets. The survey requested permission to record the GPS locations of participants along with questions regarding the biological and cultural kinship lineages of the participants with a resolution to both maternal and paternal grandparents, along with languages spoken within each family household.

In 2018, AMTO traveled to Turkana to present outreach materials and discuss initial findings. We worked with VizLab graphic design studio at ASU to make a series of images to help explain what DNA is, how to get DNA from a cell, what can be learned about people and human history with DNA, and preliminary results from this genetic study (**Figure S2**). Along with the images, we created a script to explain, in layman's terms, what the images mean; we also had questions to ask at the end of the presentation to make sure participants were following and understanding what we were demonstrating to them. We worked with the local field assistants to translate the script into the local language and to present the script to the community. We presented to a total of 6 settlement areas across 3 territorial sections in Turkana County, Kenya. Overall, the presentations were well received by the communities, and people expressed interest in the results. Some people also expressed their excitement about wanting to know what else we would find from their DNA. Additional dissemination from our group will occur in the near future as it becomes safer to travel.

The Turkana, Rendille, Samburu, and Borana are small-scale pastoral populations in northern Kenya. We are therefore taking measures to ensure the protection of these groups by providing the genetic data generated here as controlled access while maintaining appropriate standards of data access. The genomic data generated here will be available through dbGap.

Sampling and sequencing

We collected DNA and demographic information from a total of 572 individuals from Turkana, Samburu, Rendille, and Waso Borana and successfully genotyped 376 of these individuals on Illumina's Multi-Ethnic Global Array (**Table 1; Table S1**). For each ethnolinguistic group, we sampled individuals from at least two clans. Data collection occurred across northern Kenya from October 2016 – October 2017. For each participant, a DNA sample was taken in the form of saliva or cheek swab; the saliva was collected in an Oragene OG-500 DNA collection kit. In addition to collecting a DNA sample, a questionnaire was administered to each participant to acquire demographic information; this information included, for example, natal and post-marital clan affiliation, and spoken languages. DNA was extracted using a phenol-chloroform extraction method for the samples collected from cheek swabs. DNA for the samples collected with the Oragene OG-500 DNA collection kit was extracted at Yale Center for Genomic Analysis. The extracted DNA was then quantified on both a Qubit and Nanodrop. Each sample's extracted DNA was then diluted to at least 35 ng/ul in a volume of 40 ul and sent to Langebio-Cinvestav sequencing facility in Mexico for SNP genotyping on Illumina's Multi-Ethnic Global Array.

Quality control and filtering

We received the SNP genotype data in the form of a raw plink file. The coordinates were mapped to the human reference genome hg19. Initially, there were a total of 1,779,819 markers genotyped on the array. Sites with no valid mapping for the probe or with more than 1 best-scoring mapping for the probe were removed from our analyses. Additionally, we removed any sites marked as insertions or deletions. There were 27,089 duplicated variants in this file; duplicated variants have the same chromosome number and position and can have the same or different allele codes. Duplicated variants with the same chromosome, position, and allele codes were merged, while duplicated variants with the same chromosome and positions, but different allele

225 codes were removed. We merged the duplicated sites using the '--merge-equal-pos' flag and the
226 default merge mode - which ignores missing calls and sets mismatching genotypes to missing -
227 using PLINK v1.9 (Chang et al., 2015). There were a total of 1,715,718 sites after filtering.

228 As an additional quality control measure, we inferred the sex chromosome complement of
229 each individual and compared this information with reported sex information. Two approaches
230 were used to infer the sex chromosome complement of each individual, one approach based on
231 the X chromosome inbreeding coefficient (F) and the other approach based on the number of Y
232 chromosome genotype counts. Since genetic males are expected to have one X chromosome,
233 they should not have any heterozygous sites on the X chromosome (minus the pseudoautosomal
234 regions - PARs) and therefore an inbreeding coefficient equal to 1. We used the "--check-sex
235 ycount" flag in PLINK v1.9 (Chang et al., 2015) to calculate the X chromosome inbreeding
236 coefficient and the number of Y chromosome genotype counts using. The PARs were excluded
237 from this calculation.

238 As per the PLINK documentation recommendations, individuals with an X chromosome
239 inbreeding coefficient greater than 0.8 were considered male, while individuals with an X
240 chromosome inbreeding coefficient less than 0.2 were considered female. The expectation for
241 genetic females is that they have no genotype calls on the Y chromosome - since genetic females
242 are expected to be XX - however, all the female individuals in our data set had genotype calls on
243 the Y chromosome. We, therefore, visualized the X chromosome inbreeding coefficient and non-
244 missing Y chromosome genotype counts together to see the distribution of these values in males
245 and females and to identify any individuals that did not cluster with expected male and female
246 values. We removed individuals that had discrepancies between these two metrics (**Figure S3**).
247 A total of 10 individuals were removed.

248 Identity by descent (IBD) was calculated across the autosomes to identify and remove
249 related individuals. Prior to running the IBD analysis, we filtered sites with missing data across
250 samples greater than 5% (--geno 0.05 flag in PLINK), sites with Hardy-Weinberg equilibrium p-

value threshold less than $1x10^{-50}$, and pruned sites for linkage disequilibrium (50 kb window size, 10 kb variant step size, 0.2 r^2 threshold). Filtering and IBD were calculated, using PLINK v1.9 (Chang et al., 2015), for all pairwise combinations of samples in our data set and output pairs of individuals with more than 18% IBD. We removed two samples with 100% IBD that were not replicate samples (the same sample sequenced twice), as these may reflect contamination. We had two replicates in this data set and removed one sample from each replicated pair. In the cases where there were clusters of individuals related, we removed the individuals who were related to many other individuals, to minimize the number of individuals to remove. In cases where just two individuals were related, we attempted to remove roughly equal numbers of males and females when possible. A total of 67 samples were removed, leaving 301 individuals.

We next performed an initial principal components analysis (PCA) on the 301 samples using smartpca, a program within the EIGENSOFT v6.0.1 software package (Price et al., 2006). We identified a total of 4 outlier samples; these samples were removed from subsequent analyses (**Figure S4**).

After individual filtering, we performed site filtering on the autosomes, Y chromosome, and mitochondrial DNA. For the autosomes, we removed sites with more than 5% missing data across individuals at a given site ("--geno 0.05" flag in PLINK/ 95% call rate filter), removed sites that deviate from Hardy-Weinberg equilibrium (p-value threshold of $1x10^{-50}$), and performed linkage disequilibrium pruning (50 kb window size, 10 kb variant step size, and 0.2 r^2 threshold). For the Y chromosome and mitochondrial DNA, we removed sites with heterozygous calls, removed sites with more than 5% missing data across individuals at a given site, and removed sites that deviate from Hardy-Weinberg equilibrium with a p-value threshold of $1x10^{-50}$. After filtering, 516,821, 3,295, and 811 sites remained on the autosomes, Y chromosome, and mitochondrial DNA, respectively.

Population genetic analyses

To explore the genetic structure within and among northern Kenyan populations, we ran PCA using smartpca (Price et al. 2006) and ADMIXTURE (Alexander, Novembre, & Lange, 2009) on the autosomes for all unrelated samples. We ran ADMIXTURE for K = 2 - 5 with a total of 10 replicates for each K value. For streamlined post analysis and visualization of the different replication runs and K values from the ADMIXTURE analysis, we used pong (Behr, Liu, Liu-Fang, Nakka, & Ramachandran, 2016), an algorithm for processing and visualizing membership coefficient matrices. Pong finds the best alignment across all runs within and across the different K values and identifies modes among all runs for each K. We used the best alignments across all runs within and across the different K values for visualization in this manuscript.

To quantify genetic differentiation within and among northern Kenyan populations, we calculated Hudson's F_{ST} (Hudson, Slatkin, & Maddison, 1992) using the estimator derived in (Bhatia, Patterson, Sankararaman, & Price, 2013). F_{ST} was calculated on the autosomes for ethnolinguistic groups and Turkana territorial sections. Results were visualized in R (R. C. Team & Others, 2013; R. Team & Others, 2015), using the visualization package, ggplot2 (Wickham, 2011).

To test whether genetic F_{ST} was different between Ngibochoros and at least one of the ethnolinguistic groups and Kwatela and/or Ngiyapakuno and the same ethnolinguistic group, we performed a series of permutations. We randomly shuffled samples from two of the territorial sections, calculated F_{ST} between the territorial section and ethnolinguistic group, and then calculated the test statistic which was the absolute difference between F_{ST} for each territorial section. We repeated this 1,000 times and calculated the p-value. This was done for all combinations.

To visualize the relationships among haplotypes within each ethnolinguistic group, we generated haplotype networks for the Y chromosome and mitochondrial DNA. SNPs in our data

set were first set to the forward orientation. Sites on the opposite strand were identified using snpflip (https://github.com/biocore-ntnu/snpflip) and flipped using PLINK v1.9 (Chang et al., 2015). Snpflip uses information from a reference genome fasta sequence and the bim PLINK file to identify SNPs on the reverse strand. The GRCh37 GENCODE reference genome was used (ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_30/GRCh37_mapping/GRCh37.primary_assembly.genome.fa.gz). Any sites unable to be flipped due to ambiguity in whether the site was on the reverse strand were removed. A total of 2,807 and 806 sites remained on the Y chromosome and mitochondrial DNA, respectively. PLINK files were then converted to VCF format using PLINK v.1.9 (Chang et al. 2015). VCF files were converted to FASTA file format using python and haplotype networks were constructed in R (R. C. Team & Others, 2013; R. Team & Others, 2015) using pegas (Paradis, 2010) and ape (Paradis & Schliep, 2019) packages.

Quantification of intermarriage

To quantify the amount of intermarriage in the Turkana, Samburu, Rendille, and Waso Borana, we used questionnaire information collected here and from (Handley & Mathew, 2020). The questionnaire information we collected in this study requested spouses' ethnolinguistic group information only for the married men, while the questionnaire information from (Handley & Mathew, 2020) had spouse information for both married men and women that were sampled. In these groups, men may marry more than one wife, so rates of marriages were based on the total number of marriages rather than the number of individuals. For each ethnolinguistic group, we calculated the percentage of marriages both within the same ethnolinguistic group and from different ethnolinguistic groups. This was calculated separately for men and women.

Correlations between genetic and cultural differentiation

To test whether measures of genetic similarity can predict cultural similarity, we performed a series of Pearson's correlations between genetic F_{ST}, cultural F_{ST}, geographic distance, and linguistic distance. We used cultural F_{ST} values and linguistic distances that were previously calculated among these ethnolinguistic groups (Handley & Mathew, 2020). Briefly, cultural F_{ST} is a measure of cultural similarity between two groups; a low cultural F_{ST} indicates two groups are more culturally similar while a higher cultural F_{ST} indicates two groups are less culturally similar (Bell et al., 2009). For each group, language, genus, and family information were acquired from The World Atlas of Language Structures (WALS) database ("WALS Online - Home," n.d.). Using this information linguistic distances were categorized as follows: a score of 0 for groups that speak the same language (same language, same genus, same family), a score of 1 for groups that speak different languages from the same language genus (different language, same genus, same family), a score of 2 for groups that speak different languages from different genus but within the same family (different language, different genus, same family), and finally, a score of 3 for groups that speak different languages from different language families (different language, different genus, different family). Further details can be found in (Handley & Mathew, 2020). To calculate geographic distance between groups, we collected GPS coordinate information for the locations in which genetic sampling occurred. If a household fell within one precision of one or more households (within 20 meters), only one GPS measure was recorded. Using the latitude and longitude for each measured household in each population, we calculated the average distance between pairs of populations in kilometers (km). This involved calculating the distance between all households from one population to another population and then averaging these distances. This was computed for all pairs of populations using a custom python script. Pearson correlations were performed in R (R. C. Team & Others, 2013; R. Team & Others, 2015) using package ppcor (Kim, 2015) and visualized using ggplot2 (Wickham, 2011).

Data and code availability

The genotype data generated in this manuscript has been deposited on dbGap (dbGap accession number phs002654.v1.p1) and will be made available upon publication. All original code used in this manuscript can be found on GitHub: https://github.com/SexChrLab/Kenya_Fst.

Results

Intermarriage among ethnolinguistic groups contributes to the clinal pattern of genetic variation

We found a clinal pattern of genetic variation that mirrors the overall geographic distribution of the individuals we sampled (**Figure 1A, 1B**). In the principal components analysis (PCA), the Turkana samples separate from the other three groups along PC1, and along PC2 the Borana samples separate from the Rendille and Samburu (**Figure 1B**). While the Borana samples form a discrete cluster from the other ethnolinguistic groups, there is overlap between Samburu and Rendille and some overlap between Samburu and Turkana (**Figure 1B**). Interestingly, many of the overlapping Samburu and Rendille samples have a family history - a parent and/or grandparent(s) in the Rendille and Samburu, respectively (**Table S2**). In contrast, the Samburu sample that falls near Turkana and the Turkana sample that falls near Samburu have no reported cross-group family history through the grandparent level in these individuals (**Table S2**). We additionally found high rates of intermarriage between some ethnolinguistic groups and nearly non-existent intermarriage between others (**Figure 1C**). For Rendille, 5% of female marriages and 16.4% of male marriages were with a Samburu individual (**Figure 1C**). For Samburu, we observe almost the exact opposite pattern, with 11.4% of female marriages and 4.8% of male

367 marriages with a Rendille individual (**Figure 1C**). The Samburu also have low levels of
368 intermarriage with the Turkana; 3.2% of male Samburu marriages were with a Turkana individual
369 (**Figure 1C**). For the Borana, none of our sampled individuals report marriages with the Turkana,
370 Samburu, or Rendille, but did report varying levels of intermarriage between the Borana and
371 Sakuye, Gabra, Garri, and Somali (**Figure 1C**).

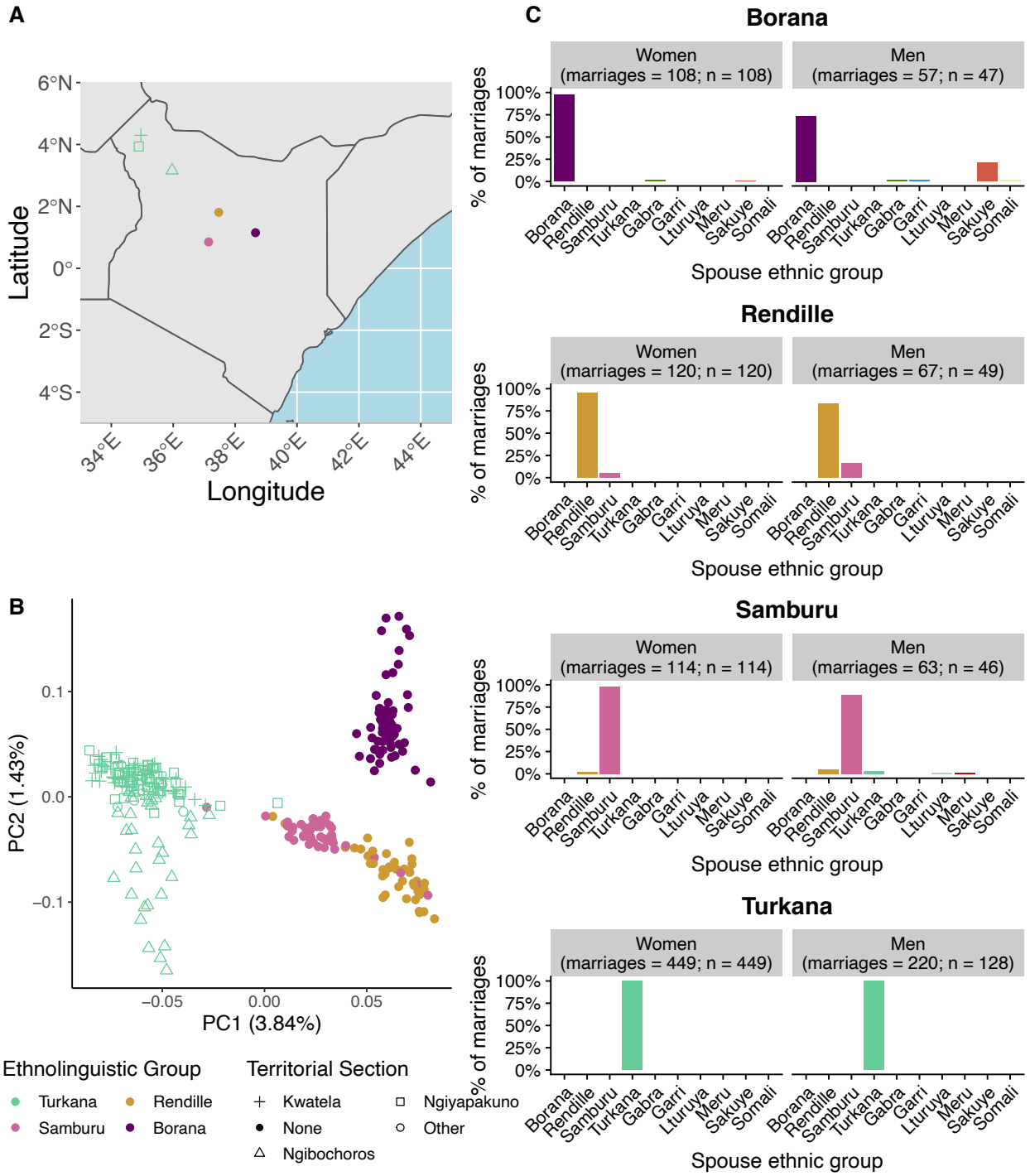


Figure 1. Intermarriage among ethnolinguistic groups contributes to the clinal pattern of genetic variation. Sampling regions, patterns of genetic variation, and rates of intermarriage across northern Kenya human populations. A) We sampled 376 individuals across four ethnolinguistic groups in northern Kenya and for the Turkana only, we additionally sampled across

three territorial sections. B) Autosomal principal components analysis (PCA). C) Rate of intermarriage across each ethnolinguistic group. Points in A and B represent sampled ethnolinguistic groups and Turkana territorial sections. Colors represent ethnolinguistic group affiliation, and shapes represent Turkana territorial section affiliation. Each point in A represents the geographic location of each sampled group, while the points in B represent individuals.

The Turkana have additional variation and geography-based substructuring

Just as these ethnolinguistic groups are geographically separated and similar to the PCA (**Figure 1B**), we observed clear genetic separation in ADMIXTURE analyses (**Figure 2A**). In the ADMIXTURE analyses, each of the ethnolinguistic groups have their own unique ancestry at K = 5 (**Figure 2A**). Interestingly, at K = 4 we observe possible admixture between Samburu and Rendille (**Figure 2A**).

In the Turkana, we additionally found geography-based genetic substructuring based on territorial region (**Figure 2A; Figure S5**). In the ADMIXTURE analyses, we see substructure within the Turkana before we see all four ethnolinguistic groups being identified separately. For example, at K = 4, the Turkana are characterized by two different ancestries, with one of these ancestries unique to individuals from the Ngibochoros territorial section (**Figure 2A**). Consistent with this, in the PCA, we observe variation within the Turkana along PC2 (**Figure 1B**). The individuals from the Ngibochoros territorial section separate from the other Turkana territorial sections along PC2. We further calculated genetic differentiation, F_{ST}, among the three Turkana territorial sections. We found that F_{ST} between Ngibochoros and either Kwatela or Ngiyapakuno are much higher than F_{ST} between Kwatela and Ngiyapakuno, which are territorial sections that are both adjacent to each other and distant from the Ngibochoros (**Figure 1A; Figure 2B**).

The observed territorial section substructuring in the Turkana may be due to geographic separation among the territorial sections. Alternatively, it is possible that individuals from Ngibochoros - the territorial section that is closer geographically to the Borana, Samburu, and Rendille than the other Turkana territorial sections - may have admixed with the other neighboring groups, resulting in the higher genetic differentiation. To investigate these scenarios further, we calculated genetic F_{ST} between each of the Turkana territorial sections and the other three ethnolinguistic groups and performed permutation tests to investigate whether F_{ST} values were significantly different between each territorial section and ethnolinguistic group. If gene flow was occurring among Ngibochoros and the other three ethnolinguistic groups, we would expect F_{ST} to be lower between Ngibochoros and at least one of the ethnolinguistic groups than between either Kwatela or Ngiyapakuno and the same ethnolinguistic groups. What we find, however, is that F_{ST} is not significantly different for each of the Turkana territorial sections when compared with each of the other ethnolinguistic groups (**Figure 2C**).

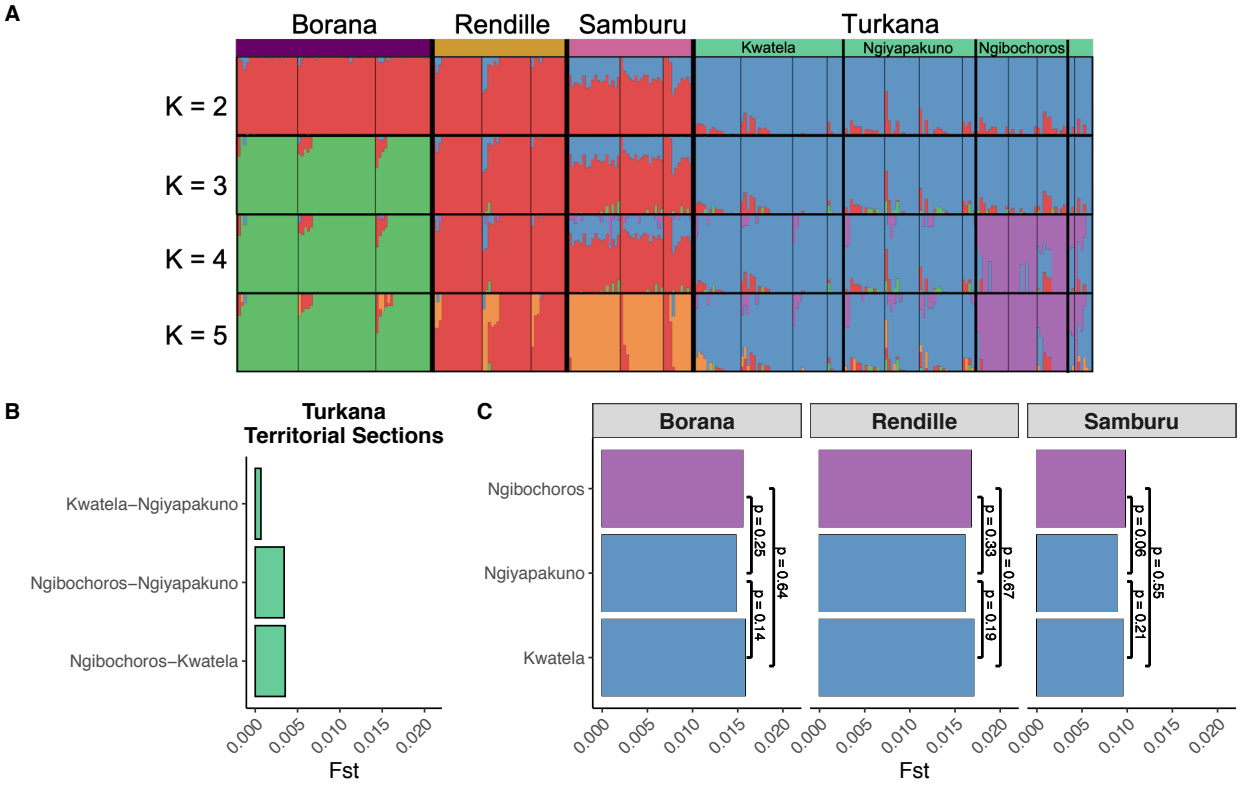
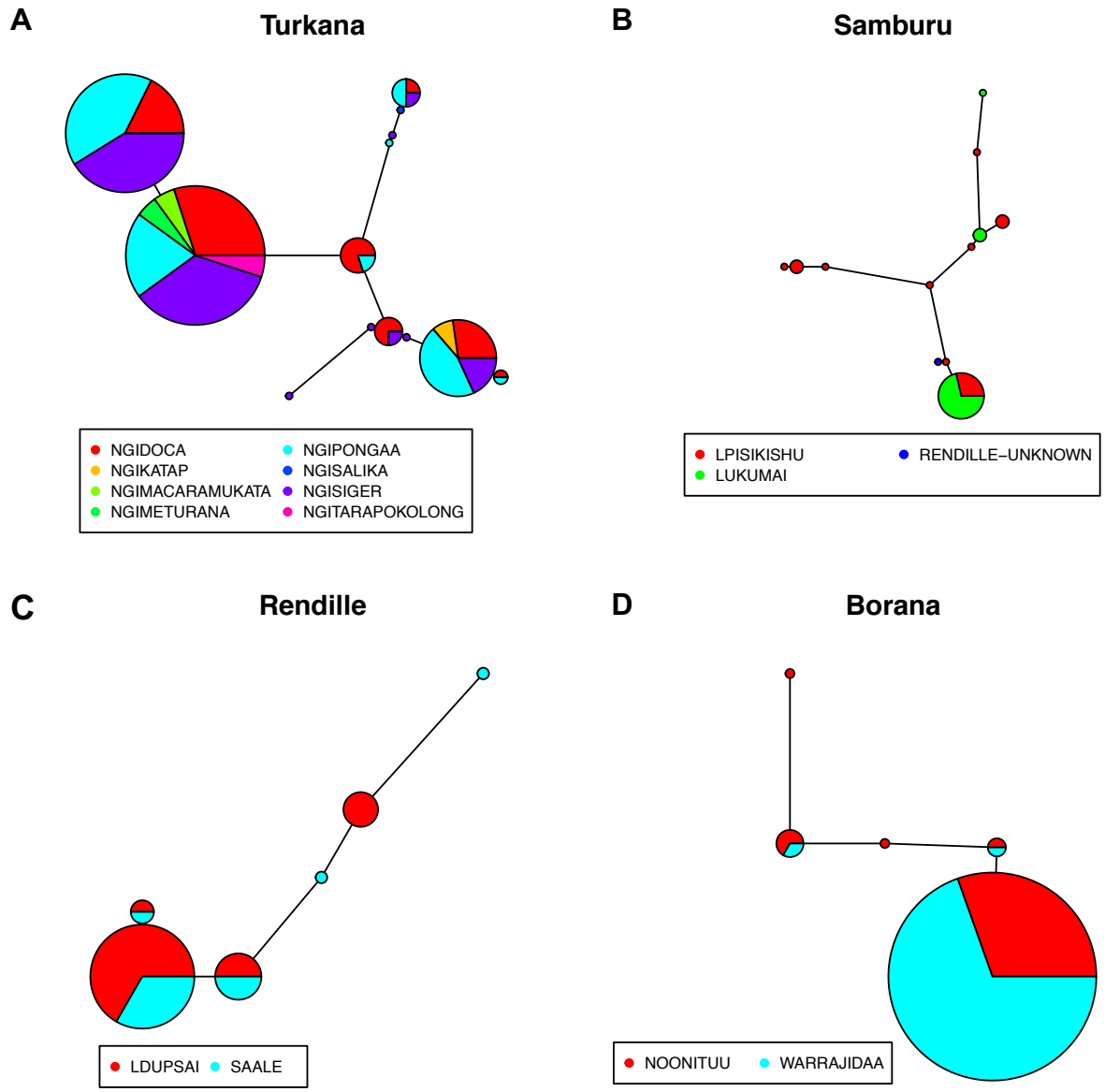


Figure 2. The Turkana have additional variation and geography-based substructuring. A) ADMIXTURE analysis for 10 replicates of K = 2 - 5 for the autosomes. Each vertical bar represents an individual, and the colors represent the proportion of ancestry corresponding to K. Samples are organized by ethnolinguistic groups (separated by thick black vertical bars), then by Turkana territorial sections (separated by medium black vertical bars), and lastly by natal clan affiliation (separated by thin black vertical bars). We observe no substructure based on natal clan affiliation but do observe geographic substructuring in the Turkana based on territorial section (purple and blue clusters at K = 4 and 5). B) Autosomal genetic differentiation (F_{ST}) among Turkana territorial sections. Individuals from Ngibochoros territorial section are more genetically different than individuals from the other sampled territories. C) Autosomal F_{ST} among each Turkana territorial section and the other sampled ethnolinguistic groups. We performed a series of pairwise permutations and found that there is no statistical difference in genetic differentiation among Turkana territorial sections and ethnolinguistic groups. P-values from the permutation tests are annotated on the plot.

Y chromosome haplotypes do not consistently cluster by natal clan affiliation

On the Y chromosome - where we expected Y haplotypes to be more similar for males from the same clan in groups with patrilineal descent than in different clans - we found that haplotypes do not consistently cluster by natal clan affiliation. For Turkana and Borana, there are no haplotypes unique to a clan (**Figure 3A, 3D**). For the Samburu, most of the haplotypes cluster by natal clan affiliation, with the exception of one haplotype that is shared among individuals from both clans we sampled (**Figure 3B**). For the Rendille we observed one haplotype unique to individuals from the Ldupsai clan, however, the rest of the haplotypes were shared among clans (**Figure 3C**). We observe similar characteristics in the mitochondrial DNA haplotype networks

438 (**Figure S6**). Overall, within these patrilineal descent groups, Y chromosome haplotypes generally
439 do not cluster by natal clan affiliation.



440

441 **Figure 3. Y chromosome haplotypes do not consistently cluster by natal clan affiliation.**
442 Haplotype networks constructed from Y chromosome SNP data from A) Turkana, B) Samburu,
443 C) Rendille and D) Borana male samples. The size of each node (circle) is proportional to the
444 number of samples in the node (larger nodes have more samples and smallest nodes have 1
445 sample). Colors within each node represent natal clan affiliation corresponding to the key in each

446 panel.

447 Genetic differentiation is driven by physical separation, not cultural
448 processes

449 Ethnolinguistic groups that are geographically closer typically had lower genetic F_{ST}
450 (**Figure 4**). The lowest genetic F_{ST} was found between the Samburu and Rendille yet they speak
451 languages from different language families; individuals from these ethnolinguistic groups are
452 closer geographically to each other than to any other ethnolinguistic group. The Turkana sampled
453 here are, on average, furthest geographically from the other ethnolinguistic groups - ranging from
454 286 km to 439 km away. The Turkana also have much higher genetic F_{ST} values with the Rendille,
455 Borana, and Samburu than F_{ST} measured between any two comparisons of the Rendille, Borana,
456 and Samburu (**Figure 4**).

457 For some groups, the pattern of genetic differentiation secondarily paralleled linguistic
458 relationships. Among the Turkana genetic F_{ST} comparisons, genetic F_{ST} between Turkana and
459 Samburu - both Nilo-Saharan speakers - is about two times lower than genetic F_{ST} between the
460 Turkana and Rendille, even though the Rendille individuals sampled here are closer to Turkana
461 by about 63 km (**Figure 4**).

462 We find that cultural differentiation does not predict genetic differentiation among
463 neighboring groups in northern Kenya. Genetic F_{ST} and cultural F_{ST} are not significantly correlated
464 with each other (R = 0.63, p-value = 0.18; **Figure 5; Figure S7**). However, we observe a
465 significant positive correlation between genetic F_{ST} and geographic distance both at the
466 ethnolinguistic group level and also Turkana territorial section level (ethnolinguistic group level: R
467 = 0.82, p-value = 0.048; Turkana territorial section level: R = 0.992, p-value = <0.001; **Figure 5;**
468 **Figure S7**).

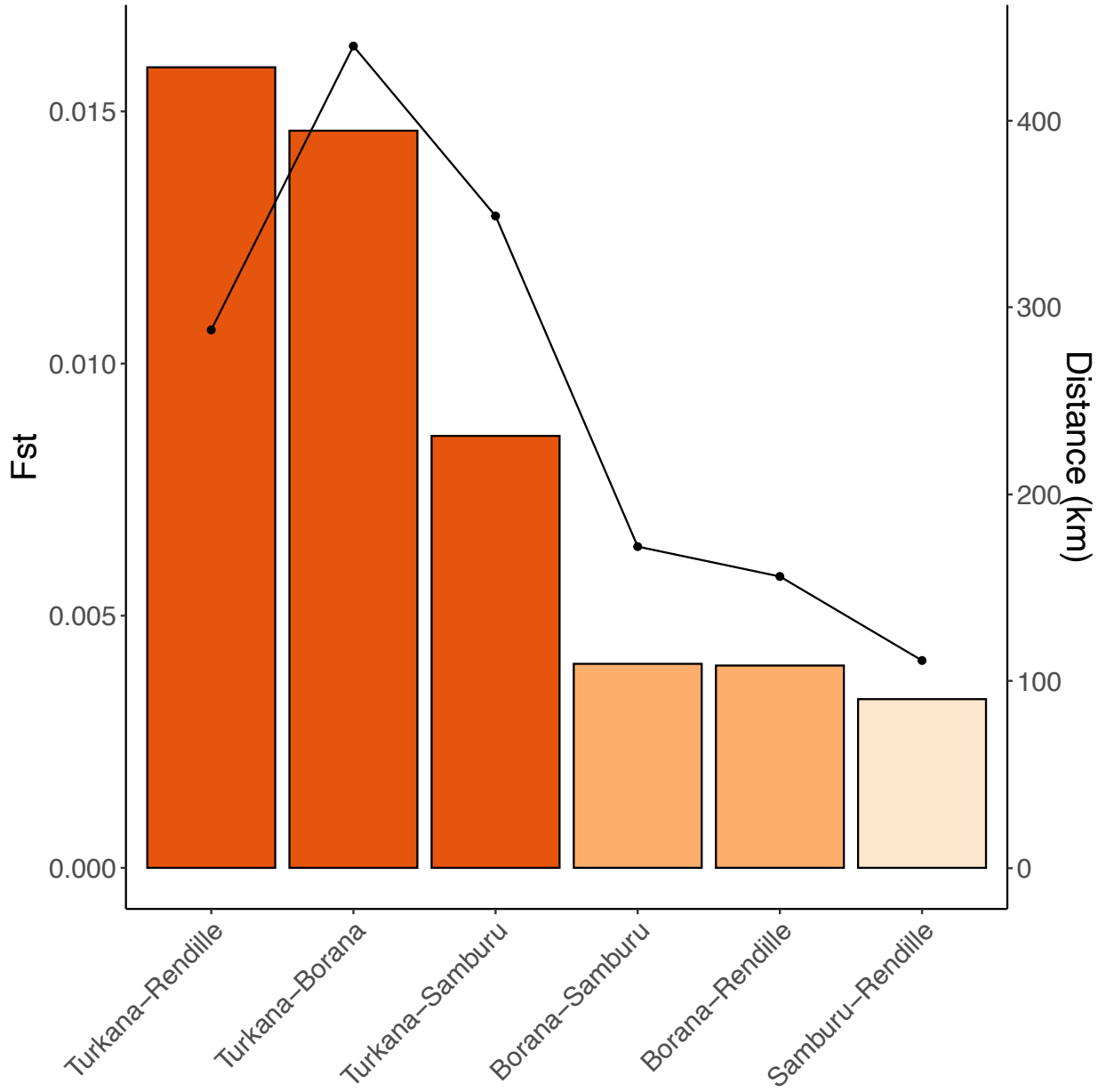


Figure 4. Geography primarily impacts patterns of genetic differentiation among ethnolinguistic groups but for some groups, the pattern of genetic differentiation secondarily parallels linguistic relationships. We calculated autosomal F_{ST} among ethnolinguistic groups. Bars are ordered by F_{ST}. Dark orange (left) are groups furthest geographically, while the lighter orange bars are groups closest geographically. The Samburu and Rendille (pale orange) are two neighboring groups that speak languages from different

language families, yet have the lowest genetic F_{ST} observed in our study. Line graph corresponds to the geographic distance between each pair of ethnolinguistic groups.

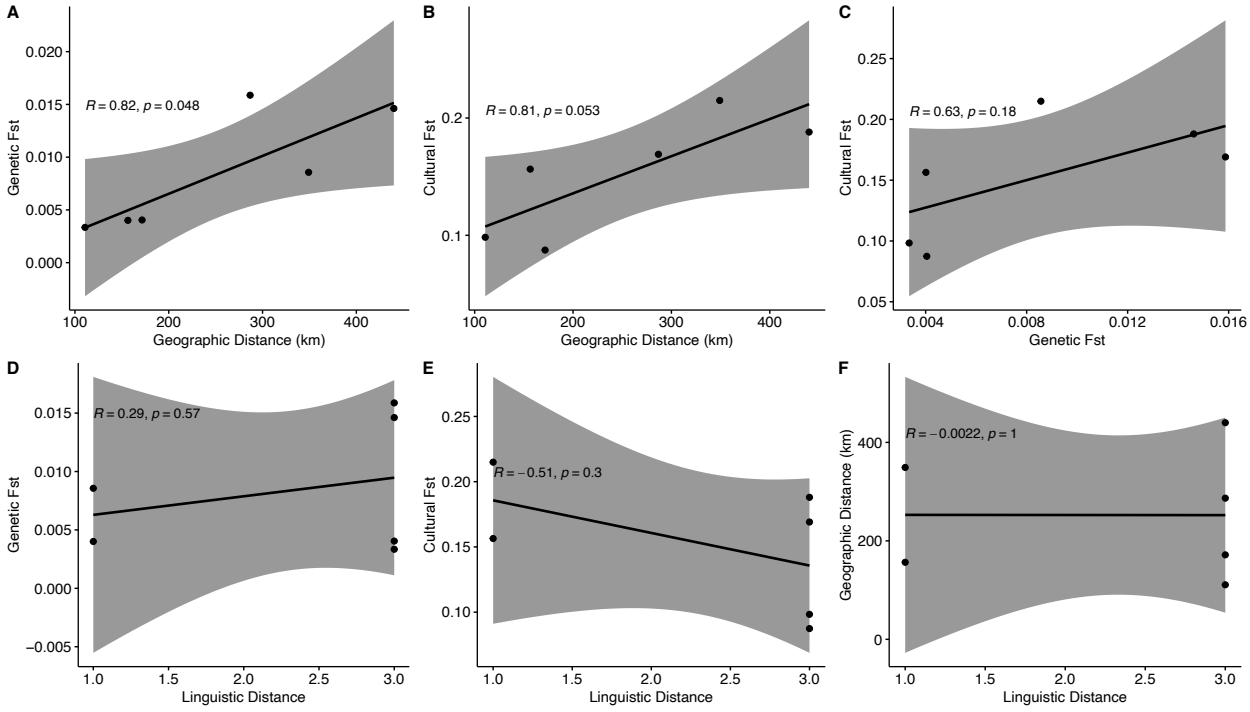


Figure 5. Cultural differentiation does not predict genetic differentiation among human ethnolinguistic groups in northern Kenya. We performed a series of Pearson's correlations to explore whether cultural differentiation may impact genetic F_{ST}. Pearson correlations for A) genetic F_{ST} and geographic distance, B) cultural F_{ST} and geographic distance, C) genetic F_{ST} and cultural F_{ST}, D) genetic F_{ST} and linguistic distance, E) cultural F_{ST} and linguistic distance, and F) geographic distance and linguistic distance. R corresponds to the correlation coefficient; p corresponds to the p-value.

Discussion

In this study, we generated genome-wide SNP genotype data and investigated the extent to which geographic and cultural processes shape genetic variation within and among four

490 pastoral populations in northern Kenya. We sampled across multiple layers of social organization
491 - ethnolinguistic groups, clans, and territorial sections - finding that geography, rather than cultural
492 processes, predominantly shape patterns of genetic variation in northern Kenya.

493 Among ethnolinguistic groups, we observed a clinal pattern of variation with a lack of
494 discrete clustering, particularly between the Rendille and Samburu, and comparatively high levels
495 of intermarriage between them. These results suggest ongoing gene flow between the Rendille
496 and Samburu, the two most closely geographically located groups, but not the most culturally
497 similar. Previous literature has noted intermarriage between the Rendille and Samburu (Spencer,
498 2012) and relatively higher levels of cooperation (Handley & Mathew, 2020). Genetic clustering
499 of Cushitic and Nilo-Saharan speaking groups (of which the Rendille and Samburu are a part of,
500 respectively) has previously been observed, supporting evidence of gene flow between these
501 larger linguistic groups (Tishkoff et al., 2009). Our findings confirm previous genetic and cultural
502 observations and provide an example in humans where genes are shared between different
503 ethnolinguistic groups at a local geographic scale.

504 Though we found a clinal pattern of variation, the Waso Borana and Turkana formed fairly
505 discrete clusters of unique genetic variation. Strikingly, in the individuals we sampled, no
506 intermarriage was reported in the Turkana at all, and for Waso Borana, there was no intermarriage
507 with the other three ethnolinguistic groups. These results suggest isolation in Waso Borana and
508 Turkana from the ethnolinguistic groups sampled in this study. However, it is possible our
509 sampling locations may be driving part of these observations. The regions that the Waso Borana
510 inhabit border the other ethnolinguistic groups; however, we sampled individuals from the Merti
511 region, which is an interior region of the Waso Borana territory. Likewise, the Turkana individuals
512 we sampled were from the north and west regions of Turkana that do not directly border the other
513 groups. We speculate that although these are nomadic groups that can traverse large distances,
514 intermixing may occur in boundary regions rather than in interior regions. Future studies including
515 individuals from both interior and border regions may shed light on this.

Perhaps one of the most intriguing results in this study was the observed genetic substructuring within Turkana based on territorial section affiliation. Individuals from the Ngibochoros territorial section are further geographically from the individuals sampled from the other territorial sections, and our results suggest that this geographic separation results in high genetic differentiation between individuals from Ngibochoros with individuals from Kwatela and Ngiyapakuno. Because there is no cultural barrier to Turkana individuals marrying individuals from different territorial sections (i.e., clan level exogamy) and because there is extensive migration in dry season across territorial section boundaries in these nomadic groups, we did not expect to observe genetic substructuring within the Turkana. Rather, we were expecting the Turkana to be largely homogenous, similar to what we observed within the other three ethnolinguistic groups we sampled in this study. The Turkana are however the most populous of the four ethnic groups, numbering approximately 1 million individuals, and having the largest geographic span. It is possible that although there is a shared cultural identity over this larger area, interpersonal interactions and co-mingling between distant Turkana territorial sections are limited. Genetic substructuring has been observed across humans on larger geographic scales (i.e., (Alsmadi et al., 2013; Bryc et al., 2010; Jakkula et al., 2008; Salmela et al., 2011; Tian et al., 2008; Tishkoff et al., 2009; Xu et al., 2009)); however, due to a lack of dense sampling within individual populations across Africa, genetic substructuring within a single ethnolinguistic group has not been widely observed within the continent, nor indeed within as small of a region as we investigate here. Because underlying genetic structure can have implications for case-control studies (Price, Zaitlen, Reich, & Patterson, 2010), our results highlight the importance of accounting for fine-scale substructuring in future genomic studies, particularly with the Turkana, and emphasize the continued importance of characterizing genetic structure across globally diverse human populations.

Within ethnolinguistic groups, we found that Y chromosome haplotypes do not consistently cluster by natal clan affiliation, suggesting that patrilineality may not have a strong impact on

patterns of male-specific genetic variation in northern Kenya pastoral populations. Previous research has found that, in groups with patrilineal descent, like pastoralists in Central Asia (Chaix et al., 2007) and the Bimoba in Ghana (Sanchez-Faddeev et al., 2013), males from the same clan have identical or similar Y haplotypes. However, this is not always the case, as seen in tribal Yemen, where Y haplotypes do not clearly cluster by clan (Raaum, Al-Meeri, & Mulligan, 2013). A possible explanation for our finding is that cultural conception of fatherhood, and therefore clan affiliation, does not always correspond with who one's biological father is. For example, in these groups, offspring from unofficial marriages - unions in which the bride price has not been paid - take on their mother's clan. This would result in a mismatch in clan assignment for these offspring. Adoption can also result in a mismatch in clan affiliation. Adoption is known to occur in Turkana and an adopted child takes on the clan of their adopted father. Overall, these results highlight that patrilineal descent groups do not always correspond with genetic patriline.

We found that genetic differentiation was highest between ethnolinguistic groups separated by the largest geographic distances, suggesting that geography primarily impacts patterns of genetic differentiation among northern Kenyan populations. Previous studies of genetic structure among human populations in Africa have found correspondence between genetic structure and linguistic affiliation and/or geography, with some studies reporting correspondence of genetic structure predominantly with linguistic affiliations (Bryc et al., 2010; Tishkoff et al., 2009), while others found that patterns of genetic structure predominantly mirror geography and ecological barriers (Babiker, Schlebusch, Hassan, & Jakobsson, 2011; Uren et al., 2016). For northern Kenyan human populations, our results suggest that geography primarily shapes the observed patterns of genetic differentiation.

Though genetic differentiation primarily paralleled geographic distances, for some groups in our study, the pattern of genetic differentiation secondarily paralleled linguistic relationships. Specifically, we found that Turkana and Samburu had lower F_{ST} than the Turkana and Rendille, despite the former being geographically more distant from one another than Turkana and

Rendille. The close genetic relationship between Turkana and Samburu compared to Turkana and Rendille could be due to shared Nilo-Saharan ancestry between the Turkana and Samburu but could also be the result of sampling from areas not directly bordering each group. Although not commonplace, the Turkana are known to intermarry with both the Rendille and Samburu, and likely occurs in regions bordering each group. Given our sampling strategy, we were unable to assess the extent of gene flow in regions directly bordering each group and if this differs from interior regions. Therefore, though our results suggest that patterns of genetic differentiation may be secondarily influenced by local linguistic affiliations for populations with similar geographic distances, additional sampling across the entire region of each ethnolinguistic group may be needed to validate this finding.

Lastly, we found that cultural differentiation does not predict genetic differentiation among neighboring populations in northern Kenya. Previous studies have used and compared linguistic or cultural distance to genetic distance to understand the extent to which genes and culture/language travel among human populations, with examples in human history where genes and culture/language have been shown to travel together (i.e., (Filippo, de Filippo, Bostoen, Stoneking, & Pakendorf, 2012; Hewlett, De Silvestri, & Guglielmino, 2002; Hunley et al., 2008; Karafet et al., 2016; Lansing et al., 2007)) and others where spoken language has been shown to have no effect on genetic structure (Veeramah et al., 2010). Here, we used cultural F_{ST} to test whether genes and culture travel together on a local geographic scale and find that cultural F_{ST} and genetic F_{ST} are not correlated with each other among northern Kenya pastoralists. We caution against the overinterpretation of this result, however, due to the limited number of groups sampled here. We anticipate this metric of cultural similarity will be of interest for future studies aimed at assessing questions of the movement of genes and culture in humans on both larger and local geographic scales. Taken together with the other results in this study, our findings suggest that geographic proximity, not cultural similarity, may provide a better explanation for the observed patterns of genetic variation among these groups.

Acknowledgments

This work was funded by the John Templeton Foundation (grant no. 48952) to SM, the National Institute of General Medical Sciences of the National Institutes of Health R35GM124827 to MAW, and The Graduate College at Arizona State University, Achievement Rewards for College Scientists (ARCS) Foundation Phoenix Chapter as a Pierson Scholar, and Arizona State University Chapter Sigma Xi to AMTO. The authors acknowledge Research Computing at Arizona State University for providing high-performance computing resources that have contributed to the research results. The National Museums of Kenya provided institutional support to conduct the research in Kenya. We thank our field research assistants for translating the questionnaires and aiding with data collection: Ekiru Carlystus, Amuria Lotiira, Chegem Muya, Gilbert Topos, Dismas Lomelu, Mohamed Noor Guyo, Abdi Wario, Paul Leramato, Damaris Lekilaui, Julius Longonyek, Sinyati Lesowapir, Rafael Letele, Simon Harugura, Benson Morsa, Ejere Ballo, and Lebo Parkeri. We also thank our participants and host communities for their hospitality and for their continued support in this project.

Declaration of Interests

The authors declare no competing interests.

Authors Contributions

AMTO, CH, ACS, SM, and MAW conceived the study. AMTO and MAW designed the methodology. SM and CH collected the samples and demographic data. AMTO and EKH extracted DNA and prepared samples for genotyping. SM, CH, and AMTO processed and cleaned the demographic data. AMTO performed all formal analyses and led the writing of the manuscript.

All authors contributed critically to writing and editing the drafts and gave final approval for publication.

References

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., … Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74.

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664.

Alsmadi, O., Thareja, G., Alkayal, F., Rajagopalan, R., John, S. E., Hebbar, P., … Thanaraj, T. A. (2013). Genetic substructure of Kuwaiti population reveals migration history. *PloS One*, *8*(9), e74913.

Babiker, H. M., Schlebusch, C. M., Hassan, H. Y., & Jakobsson, M. (2011). Genetic variation and population structure of Sudanese populations as indicated by 15 Identifiler sequence-tagged repeat (STR) loci. *Investigative Genetics*, *2*(1), 12.

Behr, A. A., Liu, K. Z., Liu-Fang, G., Nakka, P., & Ramachandran, S. (2016). pong: fast analysis and visualization of latent clusters in population genetic data. *Bioinformatics* , *32*(18), 2817–2823.

Bell, A. V., Richerson, P. J., & McElreath, R. (2009). Culture rather than genes provides greater scope for the evolution of large-scale human prosociality. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(42), 17671–17674.

Bhatia, G., Patterson, N., Sankararaman, S., & Price, A. L. (2013). Estimating and interpreting FST: the impact of rare variants. *Genome Research*, *23*(9), 1514–1521.

Bose, A., Platt, D. E., Parida, L., Drineas, P., & Paschou, P. (2021). Integrating Linguistics, Social Structure, and Geography to Model Genetic Diversity within India. *Molecular Biology and*

Evolution. doi:10.1093/molbev/msaa321

Bryc, K., Auton, A., Nelson, M. R., Oksenberg, J. R., Hauser, S. L., Williams, S., … Bustamante, C. D. (2010). Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(2), 786–791.

Cavalli-Sforza, L. L., Piazza, A., Menozzi, P., & Mountain, J. (1988). Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. *Proceedings of the National Academy of Sciences*, Vol. 85, pp. 6002–6006. doi:10.1073/pnas.85.16.6002

Chaix, R., Quintana-Murci, L., Hegay, T., Hammer, M. F., Mobasher, Z., Austerlitz, F., & Heyer, E. (2007). From social to genetic structures in central Asia. *Current Biology: CB*, *17*(1), 43–48.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, *4*, 7.

Choudhury, A., TrypanoGEN Research Group, Aron, S., Botigué, L. R., Sengupta, D., Botha, G., … H3Africa Consortium. (2020). High-depth African genomes inform human migration and health. *Nature*, Vol. 586, pp. 741–748. doi:10.1038/s41586-020-2859-7

Filippo, C. de, de Filippo, C., Bostoen, K., Stoneking, M., & Pakendorf, B. (2012). Bringing together linguistic and genetic evidence to test the Bantu expansion. *Proceedings of the Royal Society B: Biological Sciences*, Vol. 279, pp. 3256–3263. doi:10.1098/rspb.2012.0318

Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., … Sandhu, M. S. (2015). The African Genome Variation Project shapes medical genetics in Africa. *Nature*, *517*(7534), 327–332.

Handley, C., & Mathew, S. (2020). Human large-scale cooperation as a product of competition between cultural groups. *Nature Communications*, Vol. 11. doi:10.1038/s41467-020-14416-8

665 Hewlett, B. S., De Silvestri, A., & Guglielmino, C. R. (2002). Semes and Genes in Africa. *Current*
666 *Anthropology*, *43*(2), 313–321.
667 Heyer, E., Balaresque, P., Jobling, M. A., Quintana-Murci, L., Chaix, R., Segurel, L., … Hegay, T.
668 (2009). Genetic diversity and the emergence of ethnic groups in Central Asia. *BMC*
669 *Genetics*, *10*, 49.
670 Hudson, R. R., Slatkin, M., & Maddison, W. P. (1992). Estimation of levels of gene flow from DNA
671 sequence data. *Genetics*, *132*(2), 583–589.
672 Hunley, K., Dunn, M., Lindström, E., Reesink, G., Terrill, A., Healy, M. E., … Friedlaender, J. S.
673 (2008). Genetic and linguistic coevolution in Northern Island Melanesia. *PLoS Genetics*,
674 *4*(10), e1000239.
675 Jakkula, E., Rehnström, K., Varilo, T., Pietiläinen, O. P. H., Paunio, T., Pedersen, N. L., …
676 Peltonen, L. (2008). The genome-wide patterns of variation expose significant
677 substructure in a founder population. *American Journal of Human Genetics*, *83*(6), 787–
678 794.
679 Karafet, T. M., Bulayeva, K. B., Nichols, J., Bulayev, O. A., Gurgenova, F., Omarova, J., …
680 Hammer, M. F. (2016). Coevolution of genes and languages and high levels of population
681 structure among the highland populations of Daghestan. *Journal of Human Genetics*,
682 *61*(3), 181–191.
683 Kebathi, J. N. (2008). Measuring Literacy: The Kenya National Adult Literacy Survey. *Adult*
684 *Education and Development*, *71*.
685 Kim, S. (2015). ppcor: An R Package for a Fast Calculation to Semi-partial Correlation
686 Coefficients. *Communications for Statistical Applications and Methods*, *22*(6), 665–674.
687 Lansing, J. S., Cox, M. P., Downey, S. S., Gabler, B. M., Hallmark, B., Karafet, T. M., … Hammer,
688 M. F. (2007). Coevolution of languages and genes on the island of Sumba, eastern
689 Indonesia. *Proceedings of the National Academy of Sciences of the United States of*
690 *America*, *104*(41), 16022–16026.

691 Manica, A., Prugnolle, F., & Balloux, F. (2005). Geography is a better determinant of human
692 genetic differentiation than ethnicity. *Human Genetics*, *118*(3–4), 366–371.
693 Marchi, N., Hegay, T., Mennecier, P., Georges, M., Laurent, R., Whitten, M., … Heyer, E. (2017).
694 Sex-specific genetic diversity is shaped by cultural factors in Inner Asian human
695 populations. *American Journal of Physical Anthropology*, *162*(4), 627–640.
696 Mulindwa, J., Noyes, H., Ilboudo, H., Pagani, L., Nyangiri, O., Kimuda, M. P., … TrypanoGEN
697 Research Group of the H3Africa Consortium. (2020). High Levels of Genetic Diversity
698 within Nilo-Saharan Populations: Implications for Human Adaptation. *American Journal of*
699 *Human Genetics*, *107*(3), 473–486.
700 Nettle, D., & Harriss, L. (2003). Genetic and linguistic affinities between human populations in
701 Eurasia and West Africa. *Human Biology*, *75*(3), 331–344.
702 Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., … Bustamante, C. D.
703 (2008). Genes mirror geography within Europe. *Nature*, *456*(7218), 98–101.
704 Pagani, L., Kivisild, T., Tarekegn, A., Ekong, R., Plaster, C., Gallego Romero, I., … Tyler-Smith,
705 C. (2012). Ethiopian genetic diversity reveals linguistic stratification and complex
706 influences on the Ethiopian gene pool. *American Journal of Human Genetics*, *91*(1), 83–
707 96.
708 Paradis, E. (2010). pegas: an R package for population genetics with an integrated–modular
709 approach. *Bioinformatics* , *26*(3), 419–420.
710 Paradis, E., & Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and
711 evolutionary analyses in R. *Bioinformatics* , *35*(3), 526–528.
712 Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006).
713 Principal components analysis corrects for stratification in genome-wide association
714 studies. *Nature Genetics*, *38*(8), 904–909.
715 Price, A. L., Zaitlen, N. A., Reich, D., & Patterson, N. (2010). New approaches to population
716 stratification in genome-wide association studies. *Nature Reviews. Genetics*, *11*(7), 459–

463.

Raaum, R. L., Al-Meeri, A., & Mulligan, C. J. (2013). Culture modifies expectations of kinship and sex-biased dispersal patterns: a case study of patrilineality and patrilocality in tribal Yemen. *American Journal of Physical Anthropology*, *150*(4), 526–538.

Ramachandran, S., Deshpande, O., Roseman, C. C., Rosenberg, N. A., Feldman, M. W., & Cavalli-Sforza, L. L. (2005). Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(44), 15942–15947.

Salmela, E., Lappalainen, T., Liu, J., Sistonen, P., Andersen, P. M., Schreiber, S., … Kere, J. (2011). Swedish population substructure revealed by genome-wide single nucleotide polymorphism data. *PloS One*, *6*(2), e16747.

Sanchez-Faddeev, H., Pijpe, J., van der Hulle, T., Meij, H. J., van der Gaag, K. J., Slagboom, P. E., … de Knijff, P. (2013). The influence of clan structure on the genetic variation in a single Ghanaian village. *European Journal of Human Genetics: EJHG*, *21*(10), 1134–1139.

Spencer, P. (2012). *Nomads in alliance: symbiosis and growth among the Rendille and Samburu of Kenya*. Retrieved from http://eprints.soas.ac.uk/20803/1/NOMADS%20IN%20ALLIANCE%202012.pdf

Sun, H., Zhou, C., Huang, X., Liu, S., Lin, K., Yu, L., … Yang, Z. (2013). Correlation between the linguistic affinity and genetic diversity of Chinese ethnic groups. *Journal of Human Genetics*, *58*(10), 686–693.

Team, R. C., & Others. (2013). *R: A language and environment for statistical computing*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.470.5851&rep=rep1&type=pdf

Team, R., & Others. (2015). RStudio: integrated development for R. *RStudio, Inc. , Boston, MA*

URL Http://Www. Rstudio. Com, *42*, 14.

Tian, C., Kosoy, R., Lee, A., Ransom, M., Belmont, J. W., Gregersen, P. K., & Seldin, M. F. (2008). Analysis of East Asia genetic substructure using genome-wide SNP arrays. *PloS One*, *3*(12), e3862.

Tishkoff, S. A., Reed, F. A., Friedlaender, F. R., Ehret, C., Ranciaro, A., Froment, A., … Williams, S. M. (2009). The genetic structure and history of Africans and African Americans. *Science*, *324*(5930), 1035–1044.

Uren, C., Kim, M., Martin, A. R., Bobo, D., Gignoux, C. R., van Helden, P. D., … Henn, B. M. (2016). Fine-Scale Human Population Structure in Southern Africa Reflects Ecogeographic Boundaries. *Genetics*, *204*(1), 303–314.

Veeramah, K. R., Connell, B. A., Ansari Pour, N., Powell, A., Plaster, C. A., Zeitlyn, D., … Thomas, M. G. (2010). Little genetic differentiation as assessed by uniparental markers in the presence of substantial language variation in peoples of the Cross River region of Nigeria. *BMC Evolutionary Biology*, *10*, 92.

WALS Online - Home. (n.d.). Retrieved September 23, 2021, from https://wals.info/

Wickham, H. (2011). ggplot2. *Wiley Interdisciplinary Reviews: Computational Statistics*, *3*(2), 180–185.

Xu, S., Kangwanpong, D., Seielstad, M., Srikummool, M., Kampuansai, J., Jin, L., & HUGO Pan-Asian SNP Consortium. (2010). Genetic evidence supports linguistic affinity of Mlabri--a hunter-gatherer group in Thailand. *BMC Genetics*, *11*, 18.

Xu, S., Yin, X., Li, S., Jin, W., Lou, H., Yang, L., … Jin, L. (2009). Genomic dissection of population substructure of Han Chinese and its implication in association studies. *American Journal of Human Genetics*, *85*(6), 762–774.