Bite and Divide: Malaria and Ethno-Linguistic Diversity*

Matteo Cervellati
University of Bologna
IZA, Bonn

Giorgio Chiovelli
University of Bologna

Elena Esposito
EUI Florence

March 2015

Abstract

We test the hypothesis, emerging from the historical and genetic literature, that a long term exposure to malaria is an important historical determinant of ethno-linguistic diversity since it shaped the patterns of location (and isolation) of communities and induced differential interactions within and between groups across generations. We use cell level geo-localized disaggregate data. The results show that malaria exposure (measured by the stability of malaria in each location) and levels of past malaria exposure (measured by genetic immunities and historical malaria prevalence in the population) is an important and significant determinant of pre-colonial and contemporaneous ethnic diversity in Africa. The exposure to malaria leads to more (and smaller) ethnic groups and higher degrees of ethnic fractionalization and polarization. No significant effect is detected in the Americas where malaria, that affects the population today, was introduced only in the seventeenth century. The patterns are robust to extensive covariates including geo-climatological conditions, geographic distances, and pre-colonial ethnic characteristics. Using individual survey data for Africa (from Afrobarometer and DHS), we find evidence that exposure to malaria increases ethnic identification and ethnic endogamy. The findings, that are robust to extensive controls, provide a first evidence for a role of malaria in shaping ethnic diversity and individual (ethnic) interactions.

JEL Classification: J12, J15, O43, Z13.

Keywords: Ethno-linguistic Diversity, Malaria, Disaggregate Analysis, Ethnic Identity, Endogamy.

*Contacts: Matteo Cervellati m.cervellati@unibo.it, Giorgio Chiovelli giorgio.chiovelli2@unibo.it and Elena Esposito elena.esposito@eui.eu

1 Introduction

The literature has documented the important role of ethno-linguistic diversity for several socio-economic outcomes even at a highly disaggregate level, Alesina et al. (2015) **?**.[1] Michalopoulos (2012) contributes the first investigation of the long term empirical determinants of ethno-linguistic diversity and documents the crucial role of geography at a disaggregate (worldwide) level.[2] On this line of research, this paper provides a first attempt of a systematic investigation of the role of the long-term exposure to malaria for the emergence and persistence of ethnic diversity, group identity and differential interactions within and between groups in terms of differential patterns of endogamy.

The historical and anthropological literature has for a long time emphasized the role of exposure to malaria for the emergence and persistence of ethno-linguistic groups. Specifically, the risk of contagion from malaria has been argued to be an important reason for the isolation of the communities living in malarious areas.[3] That malaria favored the emergence of small isolated groups, increased interaction within groups but reduced interaction across them has been forcefully argued, most importantly, for Africa. For instance, Diamond (2000) makes the point that "Africans were combating malaria with more than just antibodies... by living in relatively small communities, .. [to] limit the level of malaria transmission". These arguments has been recently supported by the genetic literature. Genetic immunities, that emerged in response to high and persistent malaria exposure, are particularly protective of specific malaria strains and are highly location- and ethnic group-specific, Kwiatkowski et. al. (2005).[4] The genetic literature also points out that these immunities have been sustained by high levels of

[1] The early literature has documented the role of ethnic diversity across countries, for development, public policies and conflicts to name some. Easterly and Levine (1997); Alesina and Ferrara (2000), 2005; Fearon and Laitin, (2003),, Esteban et al. (2012)

[2] Ashraf and Galor (2013) show that migratory distance from the cradle of humankind in East Africa correlates with higher contemporary diversity across countries still today.

[3] For instance, Mc Neill (1976) argued that in the Indian sub-continent the heterogeneity in epidemiological endowments between intrusive Aryans and local "forest folks" prevented the local primitive communities from being "digested" into the invaders' civilization, remaining semi-autonomous and separated. See also Brower (2007)

[4] These immunities are based on various types of malaria-protective mendelian blood disorders: like the Duffy antigen, the sickle-cells disease and favism among many others.

endogamy that are necessary to preserve ethnic specific genetic traits.[5]. While the available studies are informative on the emergence and persistence of selected ethnic groups, there is no systematic evidence on the role of malaria for ethnic diversity.

We use grid of cells of 1x1 degree of size, corresponding to about 110 km at the equator as units of observation. This strategy, that follows Michalopoulos (2012) also allows for controlling for relevant country specific factors. The benchmark, location based, measure of long term exposure to malaria is the index of malaria force and stability in a given territory built by Kiszewski et. al. (2004) using biological characteristics of the regionally dominant vector mosquitoes, such as their propensity to feed on humans and their daily survival, and how these features interact with the climatic environment. We begin by documenting a positive large and highly significant role of malaria exposure on the number of ethno-linguistic groups existing today in Africa (using data from The World Language Mapping System database (Lewis et al. 2009)). Given the cross-sectional nature of the data a relevant concern is about cell (location) specific unobserved heterogeneity. A first general strategy involves checking the robustness of results to the inclusion of the extensive covariates adopted in the literature (including geo-climatological cell specific characteristics, distances as well as the inclusion of potentially more endogenous controls like population density and night lights). In line with the historical literature, the results show that a higher exposure to malaria is a relevant and robust predictor of the number of groups in a given location.

The hypothesis put forward by historians and geneticists is about the role of long term and persistent exposure to the pathogen and the emergence of groups specific genetic immunities. To investigate this prediction and isolate the role of malaria beyond location-based measure of malaria exposure we next consider population-based measures.[6] We first consider measures of the historical presence of the plasmodium parasite in the (blood of the) population. The cell level analysis is replicated using data from Lysenko (1968) on malaria endemicity in the african population at the beginning of the 20th century. Compared to location based data

[5]For instance after centuries of exposure to malaria, high isolation and strict endocamic marriage patterns lead the Tharu people of Nepal developed significant genetic resistance to malaria, to the point that they faced an about sevenfold lower malaria incidence than synpatric non-Tharu people, Modiano et. al. (1991)

[6]As discussed above, malaria deeply affects genetic and acquired immunities that can be detected in the population either in terms of parasitization rates or in terms of rates of genetic blood disorders.

on malaria exposure, data on endemicity has the relevant caveat of being heavily based on spatial interpolation but has the advantage of providing information on the actual exposure to malaria in the population at the early stages of African colonialism, rather than the geographical suitability to malaria. To test the role of genetic immunities, we use data on the presence of Duffy negative phenotype from Howes et al. (2011) measured in 2010. These genetic immunities can emerge in given population by a process of genetic selection only in response to a long period of exposure to malaria. These population based measures reflect the long term exposure to malaria of a given population. The results confirm the findings obtained using the malaria stability index and are suggestive, and reassuring, that the effect is indeed related to historical exposure to malaria.

To explore the robustness of the findings we consider the role of relevant pre-colonial characteristics of the different ethnicities using information from the Murduck Ethnographic Atlases that allow to account for heterogeneity in subsistence patterns (like e.g. presence of hunting and gathering, animal husbandry, patterns of agricultural dependence, etc), settlement patterns (and their complexity) and historical institutions (like e.g. property rights, jurisdictional hierarchies, slavery, polygyny, etc). The results show that some historical settlement patterns (that facilitated higher population density) affect the number of ethnic groups but that the effect of malaria (and its magnitude) is not affected by accounting for pre-colonial characteristics.

The analysis is replicated using the GREG data on ethnic groups, rather than ethnolinguistic groups, and confirms the findings. To test the hypothesis that the role of malaria in shaping the patters of groups formation and persistence is related to the long term exposure to the pathogen we next consider pre-colonial historical data on the spatial distribution of ethnic homelands in Africa from Murdock (1959) and in the Americas from Murdock (1951 and 1957) digitized by Chiovelli (2015)). The population in Africa, and more generally in the old world, has been exposed to malaria since at least 10.000 years while in the Americas malaria is a much more recent diseases (being brought to the new world after the discovery of America) and started affecting the population only in the seventeenth century. While malaria is widespread today in the suitable areas of the Americas the exposure of the population to the pathogen has been very limited in time from the perspective of genetic selection since the plasmodium parasite has been introduced in the new world only in the 18th century with the Columbian

exchange. If the effect of malaria is, as postulated by anthropologists and geneticists, related to a process of genetic adaptation in association to behavioural responses in terms of isolation, then we should see no effect of malaria in the Americas.[7] The results consistently show that malaria is a robust and significant predictor of pre-colonial ethnic diversity in Africa but not in the Americas. The results, that confirms the findings, are in line with the view on that malaria affects the patterns of interaction in the population over long periods of time.

To investigate, more directly, the role of malaria for differential interaction patterns within and between ethnic groups, the analysis next exploits contemporary individual level data for Africa. Using individual data from the Afrobarometer for 16 countries we relate local malaria exposure to the strength of ethnic identity. The results, that exploit within country and within ethnic groups variability, suggest that exposure to malaria significantly increases the level of identification. The results hold when controlling for geo-climatological conditions, population and night lights as well as for individual characteristics of the responders. The results suggest that, in line with the arguments put forward by historians, malaria not only affect the number (and size) of ethnic groups but also the intensity of ethnic identification.

High ethnic identification is likely to related to high interactions between members of the same ethnic group. Nonetheless, to test more directly the hypothesis that the role of malaria for ethnic identify is related to differential interactions and to test the hypothesis that genetic immunities are sustained by high endogamy, we check whether exposure to malaria affects intergroups marriage patterns. We use different measures endogamy from the DHS and from the Ethnologue (Lewis et al. 2009). The analysis again exploits within country and within ethnic groups variation and controls for location and individual characteristics. The results show that higher malaria exposure significantly reduces the likelihood of marriages outside own ethnic groups. These patterns are confirmed also with a measure of "esogamic distance" built exploiting information on the distance the language of the wife and husband (using the procedure proposed by Desmet et al. (2012)). The results provide a first shred of evidence on the fact that malaria exposure in Africa is associated to high levels of differential interactions,

[7]The genetic literature relates the presence of groups specific immunities to the exposure to malaria over the last ten thousand years in the old world (in particular in Africa), Kwiatkowski et. al. (2005). A natural check to isolate the specific role of the long term exposure to malaria and further rule out problems of unobserved heterogeneity is to study the americas as a sort of placebo.

both in terms of group identification and levels of endogamy, still today.

The paper is organized as follows. Section 2 introduces the data sources and measurement. Section 3 presents the empirical strategy and results at the cell level. Section 4 presents the result of the individual level analysis.

2 Data

2.1 Cell Level Analysis

Exposure to Malaria. The long term exposure to malaria is measured using information on the malaria suitability of a specific location, on the existence of genetic immunities and on the levels of endemicity of the pathogen in the population.

Malaria Force and Stability. The benchmark measure of long term exposure to malaria is the malaria force and stability constructed byKiszewski, Mellinger, Spielman, Malaney, Sachs, and Sachs (2004), as follows: they associated to each country a dominant vector of *Anopheles* mosquitoes (for countries with different dominant mosquitoes, mosquitoes were association to sub-regions). A monthly index of stability was then computed as a parametric function of the share of blood meal taken by the mosquito, the daily survival rate and the extrinsic incubation period. Once this regional monthly index (constructed for about 260 regions in the world) was created, in order to obtain a finer data resolution, a mimimum lagged threshold of precipitation (10 mm) was exploited as a pre-condition for malaria transmission. The yearly aggregation of such an index is the malaria stability index which we exploit in the analysis.

Malaria Endemicity. A second measure of malaria incidence that we use is a historical one. It was produced by Lysenko and Semashko (1968) and recently digitized by Hay S.I. (2004), it aims at measuring the level of malaria endemicity at the beginning of the 20th century. Endemicity is defined as the parasite rate (PR) in the 2-10 year age cohort [8]. The index takes value 0 wherever malaria is absent, 1 for epidemic areas, 2 where malaria is hypoendemic, 3 for mesoendemic areas, 4 for hyperendemic and 5 for holoendemic areas. As previously mentioned, the level of malaria endemicity could be endogenous to several factors such as agricultural

[8]Hypoendemic with PR lower than 0.1; mesoendemic with PR between 0.11-0.5; hyperendemic for 0.51-0.75 for the holoendemic class (PR higher than 0.75), the PR refers to the 1-year age group

activity and population density, however, insofar as it is measured at the begininning of the 20th it precedes the timing of the massive malaria eradication campaings which took place after the IIWW. It is reasonable to consider than before the era of massive eradication campaing, the incidence of malaria was more tightly related to climatic factors than it is today.

Genetic Immunities to Malaria. Finally we use information on genetic immunities to malaria embodied in the population. Several blood related genetic variants have been shown to be strictly associated with historical malaria incidence. HbS stands for sickle haemoglobin allele frequency in 2010, G6PD for allele frequency for G6PD deficiency in 2010, and Duffy for Duffy negative phenotype again measured in 2010.

Ethno-Linguistic Diversity. We consider various measures of ethnic diversity from several data sources.

WLMS Data. The World Language Mapping System database offers, to our knowledge, the most comprehensive mapping of the world's known living languages[9]. We compute the number of languages spoken in all cells of our dataset, excluding languages spoken by less than 10000 people overall. Figure 1 shows how the number of languages per cell varies across the world. Note that in about half of the cells composing our sample only one language is spoken, while in one fourth of the cell 3 or more languages are currently in use.

GREG Data. While the WLMS data are based on linguistic groups, the GREG dataset has its focus on ethnic groups. The spatial distribution of ethnic groups is derived from the Atlas Narodov Mira. This Atlas maps the spatial distribution of ethnicities across states atound 1960. It has the advantage to consider minority groups and to include only those group with a clear territorial basis. Moreover, the number of overlapping polygons is very little; worldwide, only 17 % of the polygons present more than one ethnic groups. This feature is particularly useful to calculate indexes of ethnic diversity on territorial basis.

Murdock Data. We use the George Peter Murdocks (1951, 1957, 1959) Ethnolinguistic Maps in order to have spatial distribution of ethnicities across Africa and the Americas in the eve of European colonization. As in other applications, the African map includes 843 tribal areas. The

[9]Note that we exploit a definition of ethnicity based on language, this choice follows both theoretical and empirical consideration.

pre-colonial Murdocks Map for the American continent (digitized by Chiovelli (2015)) includes 330 for Latin America and 277 for North America.

Measures of Ethnic Diversity. The first measure we present in the analysis is the logarithm of the number group (as in Michalopoulos (2012)). This measure of numerosity is constructed counting the number of each polygons in each cells. Moreover, in order to employ standard measure of ethnic diversity we create an index of territorial fractionalization and an index of territorial polarization. The last measure we use is the average size of each (polygon) group in the cell.

Covariates. The set of covariates includes geo-climatological and location controls. The Geographic controls include average temperature, average precipitation, average soil suitability, standard deviation of soil suitability, mean elevation, standard deviation of elevation, ruggedness, natural logarithm of water area and tse-tse fly suitability at cell level. The Location controls include distance from the coast, distance from Addis Abeba, distance from closest river, distance from the country border, distance from country capital, absolute latitude of cell centroid, number of countries in the cell, a within-country dummy variable if the cell is contained in a country. The description of the construction of the data and the sources are reported in the appendix. We also pre-colonial characteristics in terms of subsistence patterns, settlement patterns and historical institutions.

2.2 Individual Level Analysis

Ethnic Identity. The exposure to malaria shapes incentives for interaction among individuals. We want to test whether these incentives take place along ethnic lines. In the first part of the analysis, we show that the presence of malaria increases the number of ethnic groups. In this part of the paper we qualify the results of previous section focusing on the salience of the ethnic dimension for sustaining differential interaction among groups. A possible way to test this hypothesis is to look at the identification of individual both on We employ data from the 3rd wave of Afrobarometer for 17 African countries. The potential sample is composed by 21822 respondents. Our sample is reduced to 19809 individuals for which we can retrieve

information at the ethnic level. Specifically, we look at the degree of ethnic identification of each individual in each village. Our variable of interest is the 'Ethnic versus National Identity' (q82). A variable indicating the extent of self-identification with the ethnic group. The variable range from 1 to 5. The respondent was asked: Let us suppose that you had to choose between being a Kenyan and being a [respondents identity group]. Which of these two groups do you feel most strongly attached to? In our recoding, a value of one correspond to the answer I feel only Kenyan while a value of 5 to I Feel Only (respondent group).

Ethnic Endogamy and Linguistic Distances. In order to test this hypothesis we use contemporary geo-referenced data on marriage patterns in Africa provided by the Demographic and Health Survey (DHS). DHS surveys are constructed to provide detailed information on a sub-national representative sample of women of age ranging from 15 to 49. We exploit several waves of the survey conducted in 1815 cluster across 22 African countries. In most of the waves, data on married couples are provided. In this exercise we use those waves for which information on both the ethnic identity of the wife and the husband is available. The exercise of identification of endogamous marriages is complicated by the fact that the definition of ethnicity is not homogenous and varies along waves and countries. More precisely, for the same country we might have a wave where it is specified the ethnicity x of individual W_i, while in the following wave individual W_j is reported to belong to ethnicity $x1$, which actually represents a sub-family of ethnicity x. In fact, in some cases the ethno-lingustic level reported is a dialect, while in other cases we have only official national languages. This is problematic to the extent to which the variability of ethno-linguistic diversity across waves and across country is just artificially generated by the different levels of details and precision that characterized the process of DHS data collection. To mitigate this issue, our strategy is to associate all ethnic groups reported in the DHS to the various branches of the Ethnologue trees Gordon and Grimes (2005). Following this procedure, we are able to assign each ethno-linguistic entry in the DHS to the respective ethno-linguistc family in the Ethnologue. For every spouse in the sample, we are therefore able to see whether he/she formed a couple with an individual belonging to her ethnic group at various level of ethno-linguistic family aggregation.

From the DHS data, we extract individual - both for female and male - socio-economic

(religion, urban residence, education) and democraphic (age) characteristics. Moreover, we are able to proxy for the ethnic group population in the region. Moreover, we exploit the spatial disaggregation of the DHS data in order to construct our geographical controls. Notice that, in order to preserve respondents confindenciality, DHS uses random displacement in the geo-referencing process of different clusters in each country. Urban clusters are displaced with a minimum of 0 and a maximum of 2 kilometers of error. Rural clusters are instead displaced with an error of a minimum of 0 and a maximum of 5 kilometers of positional error with a further 1% of the rural clusters displaced a minimum of 0 and a maximum of 10 kilometers (DHS website) [10]. In order to minimize this displacement issue, we construct in ArcGIS a buffer of 10 km radius around the coordinates of each cluster. In such a way, we are able to use the area inside this circle to construct cluster-level measure of malaria suitability, elevation, suitability of agriculture, temperature and precipitation. Since the area of the buffer contains the "true" location of each cluster, we think this strategy represents a reasonable solution to the random displacement adopted in DHS data.

3 Ethno-Linguistic Diversity: Cell Level Analysis

3.1 Empirical Strategy

Following the seminal contribution by Michalopoulos (2012), our empirical framework exploits information at the "artificial country" level. That is, the unit of observation of the analysis are cells of 1x1 degree of size, corresponding to about 110 km at the equator. This approach has two main advantages: from one hand, the small size of cells allow us to measure more accurately geographical and environmental features, variables such as elevation, average precipitation and temperature are more accurately measured this way than in the aggregated form of a standard cross-country regressions; ii) from another hand, by looking at small artificial cells we can focus on within country variability. In other words, by exploiting "artificial country", we can easily

[10]Notice that there is no issue in the assignment of cluster near to country borders. In fact, "displacement is restricted so that the points stay within the country and within the DHS survey region (DHS website). Morevoer, "In surveys released since 2009 the displacement is restricted to the country's second administrative level where possible." (DHS website)

account for common country effects and get rid the potential confounding factors which act at the country level, such as institutional or cultural features. This approach has also a set of shortcomings, the most compelling one being that many variables for which we would like to control for, such as GDP or average human capital, are not available at such a finer spatial resolution. However, it is also true that all these controls would be endogenous to our variable of interests and could not be included into the regression without fears of biased estimates.

We estimate the relationship between the natural logarithm of the number of linguistic groups in the 1x1 degree cell and the incidence of malaria. Our baseline specification follows:

$$EthoLinguistic_{i,c} = \beta_0 + \beta_1 Malaria_{i,c} + \beta_2 \mathbf{X_{i,c}} + \mu_{\mathbf{c}} + \epsilon_{\mathbf{i,c}}$$

where i indicates the cell, and c the country. We use different measures of ethno-linguistic diversity, $EthoLinguistic_{i,c}$, including modern data on the area occupied by groups speaking different languages (from the World Language Mapping System), different ethnic groups (from GREG) and pre-colonial data on the pre-colonial homelands of ethnic groups in Africa and the Americas (from Murdock). $\mathbf{X_{i,c}}$ includes an extensive set of climatic, geographical and location controls as well as pre-colonial ethnic characteristics μ_c are country fixed effects.

3.2 Results

Table 1 reports baseline results. An increase in malaria incidence is associated with a sizable and robust increase in the number of languages spoken in a cell, see Column 1-4 in Table 1. Our favorite specification, which include country fixed effects and all exogenous covariates, is reported in Column 4. A one point increase in the Malaria Stability Index is associated with a 1.2% increase in the number of spoken languages. In terms of magnitude, this means that going from a cell with the Malaria Stability of the 25th centile (0 malaria) to the Malaria Stability at the 75th centile (14.7), we would observe an increase in the average number of languages spoken in the cell by almost 20%, which would correspond to about half a language additionally spoken cell. By focusing on within country variability, the estimated coefficient almost doubles (Column 2), while including all exogenous controls on top of the country fixed effects brings down the coefficient to the size of the unconditional regression (Column 4).

The positive effect of malaria incidence on ethnic diversity is confirmed when exploiting two

different indexes measuring the actual presence of the parasite in the blood of the population: a modern measure of genetic immunities to malaria and a historical measure of malaria endemicity. Results, reported in Table 2, are in line with baseline findings. The coefficient of both indexes are positive and large. Going from the 25th centile to the 75th centile is associated with a 11% increase in the number of groups for the genetic immunities index, and with a 20% increase in the number of groups when exploiting the coefficient of the historical malaria endemicity index. Table 3 reports regression estimates obtained by controlling for pre-colonial characteristics of the cells. The reference regression is specification 2 of Table 1. The coefficient of the malaria stability index remain almost unaffected by the inclusion of each pre-colonial control.

In Table 4 we shift our focus from linguistic to ethnic diversity. We show that a one point increase in Malaria Stability is associated with a 2% increase in ethnic diversity. The predicted increase in number of ethnic group going from the 25th centile to the 75th centile would be of 28%, effect not far in size from previous estimates. A higher malaria stability is not only associated to higher number of groups (Table 4 Column 1-2), but also to higher territorial fractionalization and polarization (Table 4 Column 3-6).

If we were to look at ethnic diversity before European colonization of Africa, the same picture would emerge. Table 5 reports results for the analysis on pre-colonial ethnic diversity. It is reassuring to see the same pattern in coefficients, also precisely estimated. Even more reassuringly, note that the estimated coefficient of Malaria Stability on the number of ethic groups with pre-colonial data is very close with the one obtained with modern data (see Table 4 Column 2 and Table 5.

Results seem not to be driven by some environmental features that our explanatory variable, the Malaria Stability Index, is capturing. Evidence of this is provided in Table 5. When running a placebo exercise on pre-colonial America, we do not find the same patterns highlighted for Africa.

4 Ethnic Identity and Ethnic Endogamy

In the previous section we documented that malaria exposure is associated with higher ethnolinguistic diversity. In this section we explore the effect of malaria exposure on patterns of interactions within and between groups. By exploiting individual survey data, we analyze the strength of ethnic identification, i.e. how much an individual feel to belong to his/her own ethnic group versus his/her country of origin. Moreover, as a proxy of group closeness, we explore patterns of endogamy of individuals across a ample set of African countries.

4.1 Empirical Strategy: Individual Level Analysis

For both analysis, the units of observation are individual respondents to surveys (Afrobarometer or DHS) in cluster c in country z for each available wave t. The role of malaria exposure for ethnic identity and endogamy is investigated exploiting within-country across-cluster variability, estimating the following specification:

$$Y_{i,c,z,t} = \beta_0 + \beta_1 Mal_{c,z} + \beta_2 \mathbf{X}_{c,z} + \mathbf{Z}_{i,c,z} + \mu_z + \mu_t + \epsilon_{i,c,z,t}$$

where $Mal_{c,z}$ is a measure of malaria suitability in cluster c in country z. $X_{c,z}$ is a vector of time invariant geographical controls, including elevation, suitability of agriculture, temperature, precipitation, relative humidity and tse-tse fly suitability. $Z_{i,c,z}$ is a vector of individual characteristics (urban resident, and age). μ_z is a country dummy, μ_e is a female ethnic dummy, and μ_t is a wave dummy. We explore the effect of malaria on three variable of interest, $Y_{i,c,z,t}$: i) ethnic saliency, a variable ranging from 1 to 5 measuring the extent of self-identification with the ethnic group; ii) endogamy, a dummy variable taking value one if the individual is married with somebody belonging to the same ethnic group in cluster c in country z in wave t; iii) linguistic distance in marriage pattern, an index ranging from 0 to 1, measuring the linguistic distance of husband and wife.

4.2 Results

Ethnic Identity (Afrobarometer Data). Table 7 reports coefficient estimates on ethnic identification. There is a positive and significant correlation between Malaria Stability on the

strength of ethnic identification. A ten-points increase in Malaria Stability would raise the index by about 2 points. In Column 2 we control for the Murdock ethnic homeland where the village is located, this set of fixed effects should control for all historical circumstances and ethnic features common to the ethnic homeland. In Column 3, we control for respondent ethnic group fixed effect: the positive effect of a malaria location on ethnic identity is present even when comparing people belonging to the same ethnic group, the coefficient shrink but remain positive, large and precisely estimated. The inclusion of geographic controls leave coefficients almost unaffected. In Column 7 we control for a full set of variables that in the previous section we show to be affected by malaria exposure.

Marriage Patterns: Ethnic Endogamy (DHS DATA) Table 8 and Table 9 report estimates for endogamy patterns. The average probability of marrying somebody of the same ethnic group in our sample is about 92%. By going from a village with no malaria transmission to a village of malaria stability of 30, close to the maximum, would increase the share of endogamous marriages by 6 percentage point, which is a sizable increase. As before, in Column 7 we control for a full set of variables that in the previous section we show to be affected by malaria exposure and, again, the magnitude remains unaffected.

5 Concluding Remarks

In this paper we have tested the hypothesis that a long term exposure to malaria is an important historical determinant of ethno-linguistic diversity. The, cell level, results show that malaria exposure (measured by the stability of malaria in each location) and the level of past malaria exposure measured by genetic immunities and historical malaria prevalence in the population is an important and significant determinant of ethnic diversity in Africa. The exposure to malaria materialized in more (and smaller) ethnic groups and a high degree of ethnic fractionalization and polarization. The results are robust to the inclusion of an extensive set of covariates and for controlling for pre-colonial characteristics. The patterns emerge for both historical (pre-colonial) and contemporaneous measures of ethnic diversity but only for Africa (and the old world). No significant patterns can be detected in the Americas where malaria was introduced

(only) two hundred years ago. The findings are consistent with the view that malaria shaped location and isolation patterns and social interactions and over long periods of time.

To further investigate the hypothesis, and the channels, we have used contemporary individual data from Africa from Afrobarometer and the Demographic and Health Survey. The results show that population facing harsher malaria display higher levels of identification in their own ethnic group and are more likely to marry within the group (or display a lower ethno-linguistic distance when marrying outside the group). The findings suggest that the exposure to the pathogen that has been hypothesized to affect the emergence of ethnic groups in the past, is likely to affect social interactions still today. The analysis so far does not allow, however, to discriminate whether social interactions are affected by today's exposure to malaria or are the results of a (cultural) heritage of past exposure to malaria in the past that crystalized in the formation of stronger ethnic identities that still persist today. In this respect, an interesting direction for future research will be to exploit information on individuals that moved outside their ethnic homeland and currently face different disease environments.

References

Alesina, A., and E. La Ferrara (2000): "Participation in heterogeneous communities," *The quarterly journal of economics*, 115(3), 847–904.

——— (2005): "Ethnic Diversity and Economic Performance," *Journal of Economic Literature*, 43(3), 762–800.

Ashraf, Q., and O. Galor (2013): "The out of Africa hypothesis, human genetic diversity, and comparative economic development," *The American Economic Review*, 103(1), 1–46.

Brower, B., and B. R. Johnston (2007): *Disappearing peoples?: indigenous groups and ethnic minorities in South and Central Asia*. Left Coast Press.

Chiovelli, G. (2015): "Pre-Colonial Centralization, Colonial Activities and Development in Latin America," *Universita' di Bologna, mimeo*.

Desmet, K., I. Ortuño-Ortín, and R. Wacziarg (2012): "The political economy of linguistic cleavages," *Journal of development Economics*, 97(2), 322–338.

Diamond, J., and L. Ford (2000): "Guns, germs, and steel: the fates of human societies," *Perspectives in Biology and Medicine*, 43(4), 609.

Easterly, W., and R. Levine (1997): "Africa's growth tragedy: policies and ethnic divisions," *The Quarterly Journal of Economics*, 112(4), 1203.

Esteban, J. M., L. Mayoral, and D. Ray (2012): "Ethnicity and Conflict: An Empirical Investigation," *American Economic Review*, 102, 1310–1342.

Fearon, J., and D. Laitin (2003): "Ethnicity, insurgency, and civil war," *American Political Science Review*, 97(1), 75–90.

Gordon, R. G., and B. F. Grimes (2005): *Ethnologue: Languages of the world*, vol. 15. SIL international Dallas, TX.

Hay S.I., Guerra C.A., T. A. N. A. . S. R. (2004): "The global distribution and population at risk of malaria: past, present and future," *Lancet Infectious Diseases*, 4(6), 327–336.

Howes, R. E., A. P. Patil, F. B. Piel, O. A. Nyangiri, C. W. Kabaria, P. W. Gething, P. A. Zimmerman, C. Barnadas, C. M. Beall, A. Gebremedhin, et al. (2011): "The global distribution of the Duffy blood group," *Nature communications*, 2, 266.

Kiszewski, A., A. Mellinger, A. Spielman, P. Malaney, S. E. Sachs, and J. Sachs (2004): "A global index representing the stability of malaria transmission," *The American journal of tropical medicine and hygiene*, 70(5), 486–498.

Kwiatkowski, D. P. (2005): "How malaria has affected the human genome and what human genetics can teach us about malaria," *The American Journal of Human Genetics*, 77(2), 171–192.

Lewis, M. P., G. F. Simons, and C. D. Fennig (2009): *Ethnologue: Languages of the world*, vol. 9. SIL international Dallas, TX.

Lysenko, A., and I. Semashko (1968): "Geography of malaria. A medico-geographic profile of an ancient disease," *Itogi Nauki: Medicinskaja Geografija*, pp. 25–146.

McNeill, W. (1976): "Plagues and Peoples. 1976," *Garden City, NY: Anchor P.*

Michalopoulos, S. (2012): "The origins of ethnolinguistic diversity," *The American economic review*, 102(4), 1508.

Modiano, G., G. Morpurgo, L. Terrenato, A. Novelletto, A. Di Rienzo, B. Colombo, M. Purpura, M. Mariani, S. Santachiara-Benerecetti, A. Brega, et al. (1991): "Protection against malaria morbidity: near-fixation of the α-thalassemia gene in a Nepalese population," *American journal of human genetics*, 48(2), 390.

Murdock, G. P. (1951): *Outline of South American Cultures*, vol. 2. Human Relations Area Files.

Murdock, G. P., and I. P. Africa (1959): "Their Cultural History," *New York*, p. 73.

Murdock, G. P., and T. J. O'Leary (1975): *Ethnographic Bibliography of North America: General North America*, vol. 1. Human Relations Area Files Press.

Ramankutty, N., J. A. Foley, J. Norman, and K. McSweeney (2002): "The global distribution of cultivable lands: current patterns and sensitivity to possible climate change," *Global Ecology and Biogeography*, 11(5), 377–392.

6 Appendix

6.1 Tables

Table 1: Baseline: Exposure to Malaria and Linguistic Diversity (WLMS) Data

Dependent Variable	Ln Number of Ethnolinguistic Groups			
	(1)	(2)	(3)	(4)
Malaria Ecology	0.013***	0.026***	0.017***	0.012**
	(0.003)	(0.006)	(0.005)	(0.005)
Grographic Controls:				
Average Temperature and Average Precipitation	No	No	Yes	Yes
Soil Suitability (Mean and Standard Deviation)	No	No	Yes	Yes
Elevation (Mean and Standard Deviation)	No	No	Yes	Yes
Ruggedness	No	No	Yes	Yes
Ln(Water Area)	No	No	Yes	Yes
Tse-tse fly Suitability	No	No	Yes	Yes
Location Controls:				
Distances (coast, river, border, capital, Addis Abeba)	No	No	No	Yes
Absolute Latitude	No	No	No	Yes
Number of Countries	No	No	No	Yes
Within Country	No	No	No	Yes
Country FE	No	Yes	Yes	Yes
Observations	1,911	1,911	1,911	1,911
R-squared	0.799	0.419	0.505	0.526

The dependent variable is the natural logarithm of the number of ethno-linguistic groups at cell level; Malaria Stability is the average level of the malaria stability index in the cell; see text for details. All regressions include the natural logarithm of land area. The variable distances include distance from coast, from river, from the country border, from the country capital and from Addis Abeba. The within-country is a dummy variable taking value one if the cell is fully contained in a country. Robust standard errors clustered by country are in parentheses. ***, **, * indicate significance at 1-, 5-, and 10-% level, respectively.

Table 2: Population Measures: Genetic Immunities and Malaria Endemicity (WLMS Data)

Dependent Variable	Number of Ethno-Linguistic Groups (Log)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Genetic Immunities	0.480**	1.341***	0.586***	0.398				
	(0.180)	(0.354)	(0.205)	(0.240)				
Malaria Endemicity					0.076***	0.187***	0.098***	0.063**
					(0.020)	(0.039)	(0.025)	(0.031)
Geographic Controls	No	No	Yes	Yes	No	No	Yes	Yes
Location Controls	No	No	Yes	Yes	No	No	Yes	Yes
Country FE	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Observations	1,911	1,911	1,911	1,91	1,911	1,911	1,911	1,9111
R-squared	0.797	0.419	0.500	0.523	0.798	0.424	0.502	0.524

The dependent variable is the natural logarithm of the number of ethnolinguistic groups at cell level; Gentic Immunities measures the average level of Duffy antigen in the cell; Malaria Endemicity is the parasite rate in the cell measured at the beginning of the 1900th century; see text for details. The "Geographic" controls include average temperature, average precipitation, average soil suitability, standard deviation of soil suitability, mean elevation, standard deviation of elevation, ruggedness, natural logarithm of water area and tse-tse fly suitability at cell level. The "Location" controls include distance from the coast, distance from Addis Abeba, distance from closest river, distance from the country border, distance from country capital, absolute latitude of cell centroid, number of countries in the cell, a within-country dummy variable if the cell is contained in a country. All regressions control the natural logarithm of land area. Robust standard errors clustered by country are in parentheses. ***, **, * indicate significance at 1-, 5-, and 10-% level, respectively.

Table 3: Accounting for Pre-Colonial Characteristics (Murduck and WLMS Data)

Dependent Variable	Number of Ethno-Linguistic Groups (Log)				
	Panel A		Panel B		
	Variable	Obs.	Variable	Malaria Stability	Obs.
Subsistence Patterns:					
Gathering	-0.260**	1539	-0.247***	0.026***	1539
	(0.101)		(0.070)	(0.005)	
Hunting	0.011	1539	-0.015	0.026***	1539
	(0.128)		(0.103)	(0.006)	
Fishing	-0.007	1539	-0.164*	0.029***	1539
	(0.141)		(0.087)	(0.006)	
Animal Husbandry	-0.065*	1539	-0.022	0.025***	1539
	(0.033)		(0.007)	(0.11)	
Agricultural Dependence	0.113***	1539	0.079***	0.021***	1539
	(0.021)		(0.023)	(0.007)	
Agricultural Type	0.018	1539	0.074	0.027***	1539
	(0.096)		(0.069)	(0.006)	
Milking	-0.055	1496	0.049	0.026***	1496
	(0.163)		(0.134)	(0.006)	
Settlement Patterns:					
Settlements	0.102***	1496	0.070***	0.021***	1496
	(0.020)		(0.020)	(0.006)	
Complex Settlement	0.220*	1496	0.113	0.025***	1496
	(0.121)		(0.091)	(0.006)	
Local Community	-0.068	1482	-0.070	0.027***	1482
	(0.111)		(0.105)	(0.006)	
Local Community (dummy)	-0.068	1482	-0.070	0.027***	1482
	(0.111)		(0.105)	(0.006)	
Historical Institutions:					
Polygyny	0.159	1516	-0.089	0.027***	1516
	(0.203)		(0.127)	(0.006)	
Clan Communities (dummy)	0.057	1423	0.056	0.025***	1423
	(0.105)		(0.100)	(0.13)	
Slavery	-0.012	1496	-0.106	0.027***	1496
	(0.110)		(0.110)	(0.006)	
Property Rights	0.008	702	-0.009	0.027***	702
	(0.093)		(0.079)	(0.004)	
Jurisdictional Hierarchy	-0.084	1482	-0.076	0.026***	1482
	(0.076)		(0.068)	(0.006)	

Each row of panel A and panel B represents a separate regression. In all regressions, the dependent variable is the natural logarithm of the number of linguistic groups at cell level. Panel A reports coefficients of interest (and number of observations) of regressions that have the average historical feature of the cell as the only explanatory variable. Regression in Panel B repeat the specification of regression in Panel A including Malaria Stability as regressor. Malaria Stability is the average level of the malaria stability index in the cell; see text for details. All regressions controls for country fixed effects. Robust standard errors clustered by country are in parentheses. ***, **, * indicate significance at 1-, 5-, and 10-% level, respectively.

Table 4: Exposure to Malaria and Ethnic Diversity - GREG Data

Dependent Variable	Number of Groups		Fractionalization		Polarization	
	(1)	(2)	(3)	(4)	(5)	(6)
Malaria Stability	0.020***	0.017***	0.007***	0.005**	0.010***	0.007**
	(0.003)	(0.004)	(0.001)	(0.002)	(0.002)	(0.003)
Geographic Controls	No	Yes	No	Yes	No	Yes
Location Controls	No	Yes	No	Yes	No	Yes
Country FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,911	1,911	1,911	1,911	1,911	1,911
R-squared	0.283	0.366	0.198	0.272	0.184	0.244

The dependent variable is the natural logarithm of the number of ethnic groups in Column (1) and (2), the territorial ethnic fractionalization of the cell in Column (3) and (4), the territorial ethnic polarization of the cell in Column (5) and (6). All dependent variables are constructed using GREG data. Malaria Stability is the average level of the malaria stability index in the cell. The "Geographic" controls include average temperature, average precipitation, average soil suitability, standard deviation of soil suitability, mean elevation, standard deviation of elevation, ruggedness, natural logarithm of water area and tse-tse fly suitability at cell level. The "Location" controls are the distance from the coast, distance from Addis Abeba, distance from closest river, distance from the country logarithm of land area. Robust standard errors clustered by country are in parentheses. ***, **, * indicate significance at 1-, 5-, and 10-% level, respectively.

Table 5: Pre-Colonial Data for Africa - Murdock Map

Dependent Variable	Ln # Groups		Fractionalization		Polarization	
	(1)	(2)	(3)	(4)	(5)	(6)
Malaria Stability	0.021***	0.014***	0.008***	0.006***	0.010***	0.007***
	(0.004)	(0.004)	(0.001)	(0.001)	(0.002)	(0.002)
Geographic Controls	No	Yes	No	Yes	No	Yes
Location Controls	No	Yes	No	Yes	No	Yes
Observations	1,911	1,911	1,911	1,911	1,911	1,911
R-squared	0.283	0.366	0.198	0.272	0.184	0.244

The dependent variable is the natural logarithm of the number of ethnic groups in Column (1) and (2), the territorial ethnic fractionalization of the cell in Column (3) and (4), the territorial ethnic polarization of the cell in Column (5) and (6). All dependent variables are constructed using the Murdock map for Africa. Malaria Stability is the average level of the malaria stability index in the cell. The "Geographic" controls include average temperature, average precipitation, average soil suitability, standard deviation of soil suitability, mean elevation, standard deviation of elevation, ruggedness, natural logarithm of water area and tse-tse fly suitability at cell level. The "Location" controls are the distance from the coast, distance from Addis Abeba, distance from closest river, absolute latitude of cell centroid. All regressions include the natural logarithm of land area. Robust standard errors clustered by country are in parentheses. ***, **, * indicate significance at 1-, 5-, and 10-% level, respectively.

Table 6: Pre-Colonial Data for the Americas - Murdock Map

Dependent Variable	Ln # Groups		Fractionalization		Polarization	
	(1)	(2)	(3)	(4)	(5)	(6)
Malaria Stability	0.012	0.023	0.006	0.009	0.012	0.015
	(0.020)	(0.014)	(0.007)	(0.006)	(0.012)	(0.010)
Geographic Controls	No	Yes	No	Yes	No	Yes
Location Controls	No	Yes	No	Yes	No	Yes
Observations	1,700	1,700	1,700	1,700	1,700	1,700
R-squared	0.074	0.168	0.044	0.101	0.043	0.092

The dependent variable is the natural logarithm of the number of ethnic groups in Column (1) and (2), the territorial ethnic fractionalization of the cell in Column (3) and (4), the territorial ethnic polarization of the cell in Column (5) and (6). All dependent variables are constructed using the Murdock map for the Americas. Malaria Stability is the average level of the malaria stability index in the cell. The "Geographic" controls include average temperature, average precipitation, average soil suitability, standard deviation of soil suitability, mean elevation, standard deviation of elevation, ruggedness, natural logarithm of water area at cell level. The "Location" controls are the distance from the coast, distance from Addis Abeba, distance from closest river, absolute latitude of cell centroid. All regressions include the natural logarithm of land area. Robust standard errors clustered by country are in parentheses. ***, **, * indicate significance at 1-, 5-, and 10-% level, respectively.

Table 7: Malaria Stability and Ethnic Identification - Afrobarometer Data

Dependent Variable	Individual Identification in the Ethic Group						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Malaria Stability	0.022***	0.022***	0.012***	0.021***	0.019***	0.012***	0.019***
	(0.005)	(0.005)	(0.004)	(0.005)	(0.005)	(0.005)	(0.005)
Geographic Controls	No	No	No	Yes	Yes	Yes	Yes
Population and Night Lights	No	No	No	No	No	No	Yes
Individual Controls (age, urban dummy)	No	No	No	No	No	No	Yes
Female Group Size in Region	No	No	No	No	No	No	Yes
Number of Groups in Region	No	No	No	No	No	No	Yes
Group Fractionalization in Region	No	No	No	No	No	No	Yes
Female Ethnic Group Share in Region	No	No	No	No	No	No	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Ethnic Homeland FE	No	Yes	No	No	Yes	No	No
Ethnic Group FE	No	No	Yes	No	No	Yes	No
Observations	19,809	19,809	19,809	19,809	19,809	19,809	19,809
R-squared	0.114	0.124	0.157	0.118	0.127	0.158	0.129

The dependent variable is a variable indicating the self-reported degree identification with the ethnic group, it ranges from 1 to 5. A value of 1 corresponds to a full identification with the country, a value of 5 to a complete identification with the ethnic group; Malaria Stability is the average level of the malaria stability index in a radius of 10 km around the centroid of the village of the respondent; see text for details. The "Geographic" controls include average temperature, average precipitation, average soil suitability, mean elevation, ruggedness, relative humidity and tse-tse fly suitability in a radius of 10 km around the centroid of the village of the respondent. The unit of observation is the Afrobarometer respondent. Standard errors are double clustered by ethnic group and district of the respondent, reported in parentheses. ***, **, * indicate significance at 1-, 5-, and 10-% level, respectively.

Table 8: Malaria Stability and Endogamous Marriages - DHS Data

Dependent Variable	Endogamic Marriages (Individual Data)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Malaria Stability	0.004***	0.005***	0.005***	0.004***	0.002***	0.002**	0.002***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Geographic Controls	No	No	No	Yes	Yes	Yes	Yes
Population and Night Lights	No	No	No	No	No	No	Yes
Individual Controls (age, urban dummy)	No	No	No	No	No	No	Yes
Female Group Size in Region	No	No	No	No	No	No	Yes
Number of Groups in Region	No	No	No	No	No	No	Yes
Group Fractionalization in Region	No	No	No	No	No	No	Yes
Female Ethnic Group Share in Region	No	No	No	No	No	No	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	No
Country FE	Yes	Yes	Yes	Yes	Yes	Yes	No
Ethnic Homeland FE	No	Yes	No	No	Yes	No	No
Female Ethnic Group FE	No	No	Yes	No	No	Yes	No
Observations	62,929	62,929	62,929	62,929	62,929	62,929	62,929
R-squared	0.068	0.109	0.141	0.070	0.110	0.144	0.109

The dependent variable is a binary indicator variable taking value 1 if the marriage is between two people from the same linguistic family at level 6 of the Ethnologue Tree, 0 otherwise; Malaria Stability is the average level of the malaria stability index in a radius of 10 km around the centroid of the village of the respondent; see text for details. The "Geographic" controls include average temperature, average precipitation, average soil suitability, mean elevation, ruggedness, relative humidity and tse-tse fly suitability in a radius of 10 km around the centroid of the village of the respondent. The unit of observation is the DHS respondent. Robust standard errors clustered by ethnic group are in parentheses. ***, **, * indicate significance at 1-, 5-, and 10-% level, respectively.

Table 9: Malaria Stability and Endogamous Marriages - DHS Data

Dependent Variable	Linguistic Distance between Husband and Wife (Individual Data)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Malaria Stability	-0.009**	-0.003**	-0.004**	-0.014***	-0.005***	-0.004*	-0.014***
	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.001)
Geographic Controls	No	No	No	Yes	Yes	Yes	Yes
Population and Night Lights	No	No	No	No	No	No	Yes
Individual Controls (age, urban dummy)	No	No	No	No	No	No	Yes
Female Group Size in Region	No	No	No	No	No	No	Yes
Number of Groups in Region	No	No	No	No	No	No	Yes
Group Fractionalization in Region	No	No	No	No	No	No	Yes
Female Ethnic Group Share in Region	No	No	No	No	No	No	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Ethnic Homeland FE	No	Yes	No	No	Yes	No	No
Female Ethnic Group FE	No	No	Yes	No	No	Yes	No
Observations	62,914	62,914	62,914	62,914	62,914	62,914	62,914
R-squared	0.025	0.084	0.225	0.033	0.084	0.225	0.144

The dependent variable is a an index of linguistic distance between the ethnic group of the wife and the group of the husband, computed thorugh the Ethnologue Tree from the WLMS; Malaria Stability is the average level of the malaria stability index in a radius of 10 km around the centroid of the village of the respondent; see text for details. The "Geographic" controls include average temperature, average precipitation, average soil suitability, mean elevation, ruggedness, relative humidity and tse-tse fly suitability in a radius of 10 km around the centroid of the village of the respondent. All coefficients are multiplied by 10. The unit of observation is the DHS respondent. Robust standard errors clustered by ethnic group are in parentheses. ***, **, * indicate significance at 1-, 5-, and 10-% level, respectively.

6.2 Data Sources

Data Construction: 1x1 Degree Cells Analysis

Number of Languages, World Language Mapping System
Number of languages in the 1x1 degree cell. Source: constructed with a (ArcGIS) spatial join between a 1x1 degree grid and the World Language Mapping System language shapefile. We keep only languages that are spoken by more than 1000 people overall. We exclude all intersections that measures less than 10 squared km and all cells with an area smaller than 1000 squared km.

Number of Ethnic Groups, GREG
Number of languages in the 1x1 degree cell. Source: constructed with a (ArcGIS) spatial join between a 1x1 degree grid and the GREG shapefile. We exclude all intersections that measures less than 10 squared km and all cells with an area smaller than 1000 squared km.

Number of Ethnic Group in Pre-Colonial Africa, Murdock (1959)
Number of ethnic groups in the 1x1 degree cell. Source: constructed with a (ArcGIS) spatial join between a 1x1 degree grid and the digitalized map of Murdock Africa map, from Nunn (2009). We exclude all intersections that measures less than 10 km squared and all cells with an area smaller than 1000 squared km.

Number of Ethnic Group in Pre-Colonial America, Murdock (1959)
Number of ethnic groups in the 1x1 degree cell. Source: constructed with a (ArcGIS) spatial join between a 1x1 degree grid and the Murdock map for North and South America, digitalized by Chiovelli (2013). We exclude all intersections that measures less than 10 km squared and all cells with an area smaller than 1000 squared km.

Land Fractionalization, World Language Mapping System
Land fractionalization in the 1x1 degree cell. Computed measuring the share of land occupied by each language in the cell (we added up not contingent shares of land of the same language in the same cell). Source: constructed with a (ArcGIS) spatial join between a 1x1 degree grid and the World Language Mapping System language shapefile. We keep only languages that are spoken by more than 1000 people overall. We exclude all intersections that measures less than 10 squared km and all cells with an area smaller than 1000 squared km.

Land Fractionalization, GREG
Land fractionalization in the 1x1 degree cell. Computed measuring the share of land occupied by each group in the cell (we added up not contingent shares of land of the same group in the same cell). Source: constructed with a (ArcGIS) spatial join between a 1x1 degree grid and the GREG shapefile. We exclude all intersections that measures less than 10 squared km and all cells with an area smaller than 1000 squared km.

Land Fractionalization in Pre-Colonial Africa, Murdock (1959)
Land fractionalization in the 1x1 degree cell. Computed measuring the share of land occupied by each group in the cell (we added up not contingent shares of land of the same group in the same cell). Source: constructed with a (ArcGIS) spatial join between a 1x1 degree grid and the digitalized Murdock Africa map, from Nunn (2009). We exclude all intersections that measures less than 10 squared km and all cells with an area smaller than 1000 squared km.

Land Fractionalization in Pre-Colonial America, Murdock (1959)
Land fractionalization in the 1x1 degree cell. Computed measuring the share of land occupied by each group in the cell (we added up not contingent shares of land of the same group in the same cell). Source: constructed with a (ArcGIS) spatial join between a 1x1 degree grid and the Murdock map for North and South America, digitalized by Chiovelli (2013). We exclude all intersections that measures less than 10 squared km and all cells with an area smaller than 1000 squared km.

Average Territorial Extension, World Language Mapping System
Average size of land occupied by each language in the cell (we added up not contingent shares of land of the same language in the same cell). Source: constructed with a (ArcGIS) spatial join between a 1x1 degree grid and the World Language Mapping System language shapefile. We keep only languages that are spoken by more than 1000 people overall. We exclude all intersections that measures less than 10 squared km and all cells with an area smaller than 1000 squared km.

Average Territorial Extension, GREG

Data Construction: 1x1 Degree Cells Analysis

Average size of land occupied by each language in the cell (we added up not contingent shares of land of the same group in the same cell). Source: constructed with a (ArcGIS) spatial join between a 1x1 degree grid and the GREG shapefile. We exclude all intersections that measures less than 10 squared km and all cells with an area smaller than 1000 squared km.

Average Territorial Extension in Pre-Colonial Africa, Murdock (1959)

Average size of land occupied by each language in the cell (we added up not contingent shares of land of the same group in the same cell). Source: constructed with a (ArcGIS) spatial join between a 1x1 degree grid and the digitalized Murdock Africa map, from Nunn (2009). We exclude all intersections that measures less than 10 squared km and all cells with an area smaller than 1000 squared km.

Average Territorial Extension in Pre-Colonial America, Murdock (1959)

Average size of land occupied by each language in the cell (we added up not contingent shares of land of the same group in the same cell). Source: constructed with a (ArcGIS) spatial join between a 1x1 degree grid and the Murdock map for North and South America, digitalized by Chiovelli (2013). We exclude all intersections that measures less than 10 squared km and all cells with an area smaller than 1000 squared km.

Malaria Stability

Average Malaria Stability Index in the 1x1 degree grid cell. Source: average Malaria Stability is constructed as the 1x1 degree cell average of the Malaria Stability index from **?** across grids, computed using ArcGIS with data in EASE GLOBAL GRID projection.

Malaria Endemicity

Average Historical Malaria Endemicity in the 1x1 degree grid cell. Source: average Historical Malaria Endemicity is constructed as the 1x1 degree cell average of the Malaria Endemicity level, devised by **?** and digitalized by **?**, computed using ArcGIS with data in EASE GLOBAL GRID projection.

Duffy Antigen

Contemporary share of Duffy negative phenotype in the 1x1 degree grid cell. Source: Howes et al. (2011), computed using ArcGIS with data in EASE GLOBAL GRID projection.

Average Temperature

Mean annual 1x1 degree cell temperature (baseline period 1961-1990). Source: average temperature is constructed as the 1x1 degree cell of the mean annual temperature across raster grids, computed using ArcGIS with data in EASE GLOBAL GRID projection, from FAO/IIASA, 2011-2012. Global Agro-ecological Zones (GAEZ v3.0). FAO Rome, Italy and IIASA, Laxenburg, Austria.

Average Precipitation

Average 1x1 degree cell monthly precipitation mm/month (baseline period 1961-1990). Source: average monthly precipitation is constructed as the 1x1 degree cell average of the mean monthly precipitation across 10 minute grids, computed using ArcGIS with data in EASE GLOBAL GRID projection, with CRU CL 2.0 data from **?**.

Land Suitability

Average land suitability in the 1x1 degree cell . Source: average Land Suitability is constructed as the 1x1 degree cell average of the land suitability index from Ramankutty, Foley, Norman, and McSweeney (2002) across 0.5 degree raster grids, computed using ArcGIS with data in EASE GLOBAL GRID projection.

Variation in Land Suitability

Standard deviation of land suitability in the 1x1 degree cell. Source: standard deviation for Land Suitability is constructed as the 1x1 degree cell standard deviation of the land suitability index from across 0.5 degree raster grids, computed using ArcGIS with data in EASE GLOBAL GRID projection.

Mean Elevation

Average 1x1 degree cell elevation. Source: mean elevation is constructed as the 1x1 degree cell average of elevation across grids, computed using ArcGIS with data in Africa EASE GLOBAL GRID projection, with data from National Oceanic and Atmospheric Administration (NOAA) and U.S. National Geophysical Data Center, TerrainBase, release 1.0 (CD-ROM), Boulder, Colo.

Variation in Elevation

Data Construction: 1x1 Degree Cells Analysis

Standard deviation of elevation in the 1x1 degree cell. Source: constructed as the 1x1 degree cell standard deviation of elevation across grids, computed using ArcGIS with data in Africa EASE GLOBAL GRID projection, with data from National Oceanic and Atmospheric Administration (NOAA) and U.S. National Geophysical Data Center, TerrainBase, release 1.0 (CD-ROM), Boulder, Colo.

Ruggedness

Average 1x1 degree cell ruggedness (Terrain Ruggedness Index, 100 m). Source: mean ruggedness is constructed as the 1x1 degree cell average of elevation across grids, computed using ArcGIS with data in EASE GLOBAL GRID projection, with data from Terrain Ruggedness Index originally devised by Riley, DeGloria, and Elliot (1999), obtained through http://diegopuga.org.

Total Water Area

Total area occupied by water within the 1x1 degree cell. Source: constructed with ArcGIS by intersecting the 1x1 degree cell grid and the Digital Chart of the World inwater shapefile, by intersecting the 1x1 degree cell grid and the Digital Chart of the World oceans and sea shapefile. We sum up total in-cell water area and the areas of the cell occupied by seas and oceans, areas computed with data in EASE GLOBAL GRID projection.

Total Area

Total area of the 1x1 degree cell. Source: constructed with ArcGIS by intersecting the 1x1 degree cell grid and the World Language Mapping System shapefile from the Digital Chart of the World. We exclude cell parts not covered by World Language Mapping System data (and by the Africa Murdock Map, and the Murdock map for North and South America), areas computed with data in EASE GLOBAL GRID projection.

Number of Countries

Total number of countries in the 1x1 degree cell. Source: constructed with ArcGIS by intersecting the 1x1 degree cell grid and the country boundaries shapefile from the Digital Chart of the World, areas computed with data in EASE GLOBAL GRID projection.

Within Country

Dummy variable taking value one if the 1x1 degree cell belong to one single country, 0 otherwise. Source: constructed with ArcGIS by intersecting the 1x1 degree cell grid and the country boundaries shapefile from the Digital Chart of the World.

Ln Migratory Distance

Migratory distance, on a land path, from Adis Ababa. Source: computed following **?**. The distance of the centroid of 1x1 degree cell from from Adis Abeba is computed using the Haversine formula. In order to replicate the most likely migration pattern followed by early men, we calculated the distance from Adis Ababa of the path that connect several obligatory intermediate points, and namely: Cairo, Istambul, Phnom Phen, Anadyr and Prince Rupert.

Ln Distance Coast

Distance to closest coast. Source: constructed with ArcGIS using the digital Chart of the World coastline shapefile.

Ln Distance Border

Distance to closest border. Source: constructed with ArcGIS using the digital Chart of the World boundaries shapefile.

Ln Distance Capital

Distance to the capital of the country where lies the centroid of the 1x1 degree cell. Source: constructed with ArcGIS using the digital World Capital shapefile.

Ln Distance River

Distance to closest river. Source: constructed with ArcGIS using Major Rivers World selected p3w shapefile, retrived from www.naturalearth.com.

Absolute Latitude

Absolute latitudinal distance from the equator in decimal degrees of the 1x1 degree cell. Source: computed using ArcGIS with data in WSG1984.

Night Lights

Data Construction: 1x1 Degree Cells Analysis
Average population in 1x1 degree cell. Source: constructed as the 1x1 degree cell average of population across raster grids, computed using ArcGIS with data in EASE GLOBAL GRID projection, with data from the Center for International Earth Science Information Network - CIESIN - Columbia University, United Nations Food and Agriculture Programme - FAO, and Centro Internacional de Agricultura Tropical - CIAT. 2005. Gridded Population of the World, Version 3 (GPWv3): Population Count Grid. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). http://sedac.ciesin.columbia.edu/data/set/gpw-v3-population-count.
Ln Population Density
Average night lights intesity in 1x1 degree cell. Source: constructed as the 1x1 degree cell average of night lights intesity grids, computed using ArcGIS with data in EASE GLOBAL GRID projection, with data from NOOA National Geophysical Data Centre for the year 2000.

Table 11: Summary statistics: Old World Sample

Variable	Mean	Std. Dev.	Min.	Max.	N
Ln(N. of Language WLMS)	0.948	0.737	0	3.932	4186
Land Fractionalization, WLMS	0.329	0.279	0	0.943	4186
Avg. Land Extension, WLMS	4.707	3.382	0.208	12.308	4186
Ln(N. of Ethnic Groups GREG)	0.648	0.562	0	2.303	4162
Land Fractionalization, GREG	6.060	3.513	0.363	12.308	4162
Avg. Land Extension, GREG	6059.686	3513.199	363.465	12308.464	4162
Malaria Stability	5.552	8.222	0	34.728	4186
Duffy Antigen	0.378	0.431	0	1	4186
Malaria Endemicity	2.636	1.49	0	5	4121
Average Temperature	19.765	8.694	-7.813	30.386	4186
Average Precipitation	80.381	66.388	0.403	462.413	4186
Mean Soil Suitability	0.378	0.295	0	0.992	4186
Variation Soil Suitability	0.046	0.048	0	0.409	4186
Mean Elevation	863.811	949.699	-33.9	5725.512	4186
Variation Elevation	189.392	193.727	0	1868.89	4186
Ruggedness	1.231	1.479	0	10.168	4186
Total Water Area	0.504	1.014	0	8.233	4186
Total Area	9.797	2.703	1.005	12.308	4186
Number of Country	1.29	0.531	1	4	4186
Within Country	0.746	0.435	0	1	4186
Ln(Migratory Distance)	8.43	0.799	4.358	9.602	4186
Ln(Distance Coast)	5.553	1.635	-4.104	7.636	4186
Ln(Distance Border)	4.221	1.517	-4.104	6.907	4186
Ln(Distance River)	5.519	1.355	-2.026	8.146	4186
Ln(Distance Capital)	6.385	0.924	2.743	9.143	4186
Absolute Latitude	21.123	13.292	0.25	53.5	4186
Ln(Night Light)	1.059	0.353	0.698	3.455	4186
Ln(Population)	3.162	1.794	-2.197	8.497	4186

Table 12: Summary statistics: Pre-Colonial Africa

Variable	Mean	Std. Dev.	Min.	Max.	N
Ln(N. of Language Murdock)	0.891	0.54	0	2.565	1742
Land Fractionalization, Murdock	0.34	0.247	0	0.855	1742
Avg. Land Extension, Murdock	5.302	3.094	0.585	12.308	1742
Malaria Stability	12.019	9.038	0	34.728	1742
Duffy Antigen	0.851	0.215	0.089	1	1742
Malaria Endemicity	3.398	1.335	0	5	1734
Average Temperature	24.38	3.455	12.253	30.386	1742
Average Precipitation	79.301	48.458	0.403	270.494	1742
Mean Soil Suitability	0.347	0.251	0.001	0.992	1742
Variation Soil Suitability	0.05	0.045	0	0.282	1742
Mean Elevation	656.774	453.579	-33.9	2517.88	1742
Variation Elevation	136.199	130.142	2.963	795.081	1742
Ruggedness	0.656	0.831	0.004	6.905	1742
Total Water Area	0.377	0.852	0	7.206	1742
Total Area	10.723	2.487	1.249	12.308	1742
Number of Country	1.351	0.571	1	4	1742
Within Country	0.695	0.46	0	1	1742
Ln(Migratory Distance)	7.729	0.700	4.358	8.693	1742
Ln(Distance Coast)	5.891	1.331	-1.452	7.445	1742
Ln(Distance Border)	4.224	1.348	-2.711	6.391	1742
Ln(Distance River)	5.503	1.188	-0.786	7.274	1742
Ln(Distance Capital)	6.091	0.789	2.743	7.576	1742
Absolute Latitude	11.803	7.453	0.417	34.417	1742
Ln(Night Light)	0.957	0.235	0.698	2.801	1742
Ln(Population)	2.367	1.458	-2.197	6.901	1742

Table 13: Summary statistics: Pre-Colonial America

Variable	Mean	Std. Dev.	Min.	Max.	N
Ln(N. of Language Murdock)	0.598	0.469	0	1.792	1461
Land Fractionalization, Murdock	0.233	0.226	0	0.747	1461
Avg. Land Extension, Murdock	6.402	3.109	0.715	12.308	1461
Malaria Stability	2.247	2.243	0	6.949	1461
Duffy Antigen	0.039	0.079	0	0.612	1461
Malaria Endemicity	1.984	1.297	0	4	1438
Average Temperature	21.738	5.441	3.043	28.372	1461
Average Precipitation	115.756	70.233	0.115	640.889	1461
Mean Soil Suitability	0.453	0.271	0	0.997	1461
Variation Soil Suitability	0.058	0.051	0	0.325	1461
Mean Elevation	661.324	855.958	-7.5	4627.625	1461
Variation Elevation	192.496	243.253	0	1403.824	1461
Ruggedness	0.745	0.985	0.001	5.451	1461
Total Water Area	0.467	0.904	0	7.374	1461
Total Area	10.787	2.313	1.08	12.308	1461
Number of Country	1.181	0.414	1	3	1461
Within Country	0.831	0.375	0	1	1461
Ln(Migratory Distance)	10.076	0.112	9.762	10.243	1461
Ln(Distance Coast)	5.385	1.472	-1.37	7.34	1461
Ln(Distance Border)	4.606	1.479	-2.494	6.947	1461
Ln(Distance River)	5.203	1.331	-0.661	7.646	1461
Ln(Distance Capital)	6.539	0.835	2.982	8.212	1461
Absolute Latitude	17.076	11.355	0.5	54.583	1461
Ln(Night Light)	1.063	0.354	0.698	3.293	1461
Ln(Population)	1.868	1.728	-2.197	7.641	1461

Data Sources: Malaria and Endogamy

Individual Level Variables

Endogamous Marriages Tree Level 6

Dummy variable equal to one for a marriage between two individuals of an ethnic group belonging to the same linguistic family at level 6 of the Ethnologue, zero otherwise. Since the number of branches varies among linguistic families and subfamilies, we follow the baseline approach from **?** and extending the last available branch to the bottom. As a robustness, we exclude observations that do not have the same information on linguistic branches. Source: Demographic and Health Survey (DHS). Variable created based on ethhnic variables ($v131$ for female and $mv131$ for male for most of the waves) in the DHS. We matched DHS ethnic group to Ethnologue linguistic group. Through this matching, for each DHS declared ethnicity we have information on the linguistic family at each level of the linguistic tree.

Marriage Partners' Distance

Measure of language diversity of the married couple, computed following **?** . Given that the maximum level of branches was 12, the measure was computed as such $Dis = 1-(L/12)^0.005$, where L is the number of branches in common. Since the number of branches varies among linguistic families and subfamilies, we follow the baseline approach from **?** and extending the last available branch to the bottom. Source: Demographic and Health Survey (DHS). Variable created based on ethhnic variables ($v131$ for female and $mv131$ for male for most of the waves) in the DHS. We matched DHS ethnic group to Ethnologue linguistic group. Through this matching, for each DHS declared ethnicity we have information on the linguistic family at each level of the linguistic tree.

Year

Survey year. Source: Demographic and Health Survey (DHS) (variable $v007$).

Country

Country Survey. Source: Demographic and Health Survey (DHS) (variable $v000$).

Female Age

Age of the female respondent. Source: Demographic and Health Survey (DHS) ($vv012$ for female and $mv012$ for male).

Urban Residence Female

Dummy variable taking value one the female respondent is living in a urban area, zero otherwise. Source: Demographic and Health Survey (DHS) ($v025$ for female and $mv025$ for male).

Woman Ethnic Group

Variable indicating the ethnic group identity of the female respondent. Source: Demographic and Health Survey (DHS) ($v131$)

Female Group Size in the Region

Variable measuring the number of individuals surveyed in the region belonging to the same group of the female respondent. Source: Demographic and Health Survey (DHS) ($v131$)

Female Group Share in the Region

Variable measuring the share of individuals surveyed in the region belonging to the same group of the female respondent. Source: Demographic and Health Survey (DHS) ($v131$)

Cluster Level Variables

Malaria Stability

Average Malaria Stability Index in the 10 km radius around the coordinates of each cluster. Source: Malaria Stability index from **?**.

Average Temperature

Mean annual temperature in the 10 km radius around the coordinates of each cluster (baseline period 1961-1990). Source: FAO/IIASA, 2011-2012. Global Agro-ecological Zones (GAEZ v3.0). FAO Rome, Italy and IIASA, Laxenburg, Austria.

Average Precipitation

Average monthly precipitation mm/month in the 10km radius around the coordinates of each cluster (baseline period 1961-1990). Source: CRU CL 2.0 data from **?**.

Land Suitability

Average land suitability in the 10 km radius around the coordinates of each cluster. Source: Ramankutty, Foley, Norman, and McSweeney (2002).

Data Sources: Malaria and Endogamy

Mean Elevation

Average elevation in the 10 km radius around the coordinates of each cluster. Source: National Oceanic and Atmospheric Administration (NOAA) and U.S. National Geophysical Data Center, TerrainBase, release 1.0 (CD-ROM), Boulder, Colo.

Average Humidity

Average relative humidity (%): how much water vapor is in the air in the 10 km radius around the coordinates of each cluster. Source: data from **?** accessed through Atlas of Biosphere.

Ruggedness

Average ruggedness (Terrain Ruggedness Index, 100 m) in the 10 km radius. Source: data from Terrain Ruggedness Index originally devised by Riley, DeGloria, and Elliot (1999), accessed through http://diegopuga.org.

Tse Tse Suitability

Average predicted suitability for tsetse flies in the 10 km radius. Source: predicted suitability for tsetse flies is constructed as the sum of predicted suitability (0 to 1) for the presence of Tsetse groups (Fusca, Morsitans and Palpalis). Data produced for FAO - Animal Health and Production Division and DFID - Animal Health Programme by Environmental Research Group Oxford (ERGO Ltd) in collaboration with the Trypanosomosis and Land Use in Africa (TALA) research group at the Department of Zoology, University of Oxford.

Ethnic Homeland

Variable indicating the ethnic homeland where the DHS cluster falls, based on the map Murdock Africa map (1959), digitized by Nunn (2009).

Region Level Variables

Number of Groups in the Region

Number of different ethnic groups in the region, retrieved through the sample of all female respondents in the region. Source: Demographic and Health Survey (DHS) ($v131$)

Group Fractionalization in the Region

Ethnic fractionalization in the region, retrieved through the sample of all female respondents in the region. Source: Demographic and Health Survey (DHS) ($v131$)

Table 15: Summary statistics: Endogamy and Malaria Stability

Variable	Mean	Std. Dev.	Min.	Max.	N
Endogamy, Tree Level 1	0.99	0.099	0	1	62915
Endogamy, Tree Level 2	0.961	0.192	0	1	62915
Endogamy, Tree Level 3	0.956	0.205	0	1	62915
Endogamy, Tree Level 4	0.947	0.224	0	1	62915
Endogamy, Tree Level 5	0.938	0.241	0	1	62915
Endogamy, Tree Level 6	0.925	0.263	0	1	62915
Endogamy, Tree Level 7	0.913	0.282	0	1	62915
Endogamy, Tree Level 8	0.911	0.285	0	1	62915
Endogamy, Tree Level 9	0.869	0.337	0	1	62915
Endogamy, Tree Level 10	0.837	0.37	0	1	62915
Marriage Partners' Distance	0.017	0.1	0	1	62915
Malaria Stability	17.725	10.598	0	36.957	62915
Average Humidity	61.66	11.842	21	86	62915
Terrain Ruggedness	58986.754	120980.076	0	1931305	62915
Elevation	536.249	554.594	0	3126	62915
TseTse Suitability	0.341	0.308	0	1	62915
Precipitation	98.921	48.081	0	289.548	62915
Mean Suitability	0.463	0.236	0.001	0.999	62915
Average Temperature	25.162	3.667	0	30.165	62915
Night Lights	5.342	9.663	2	63	62915
Population Density	330.723	1062.63	0.133	11133.601	62915
Female Respondent Urban	0.258	0.437	0	1	62915
Female Age	30.41	8.285	15	49	62915
Female Group Size in the Region	263.673	389.124	1	1998	62915
Number of Groups in the Region	7.878	4.176	1	26	62915
Group Fractionalization in the Region	0.523	0.234	0	0.901	62915
Ethnic Group Share in the Region	0.445	0.296	0.001	1	62915

Data Construction: Malaria and Ethnic Salience

Individual Level Variables

Ethnic Salience

A variable indicating the extent of self-identification with the ethnic group. The variable range from 1 to 5. The respondent was asked: 'Let us suppose that you had to choose between being a Kenyan and being a [respondents identity group]. Which of these two groups do you feel most strongly attached to?' In our recoding, a value of one correspond to the answer 'I feel only Kenyan', a value of 2 with the answer 'I Feel More Kenyan than (respondent' group)', a value of 3 to the answer 'I Feel Equally Kenyan and (respondent' group)', a value of 4 to 'I Feel More (respondent' group) than Kenyan', and 5 to 'I Feel Only (respondent' group)'.

Country

Country Survey. Source: Afrobarometer (variable *country*).

Respondent Age

Age of the respondent. Source: Afrobarometer (variable $q111$).

Urban Residence

Dummy variable taking value one the female respondent is living in a urban area, 2 if in a rural. Source: Afrobarometer (variable *urbrur*).

Ethnic Group

Variable indicating the ethnic group indicated by the respondent. Source: Afrobarometer (variable $q3$).

Group Size in the Region

Variable measuring the number of individuals surveyed in the region belonging to the same group of the respondent. Source: Afrobarometer (variable $q3$).

Group Share in the Region

Variable measuring the share of individuals surveyed in the region belonging to the same group of the respondent. Source: Afrobarometer (variable $q3$).

Cluster Level Variables

Malaria Stability

Average Malaria Stability Index in the 10 km radius around the coordinates of each cluster. Source: Malaria Stability index from **?**.

Average Temperature

Mean annual temperature in the 10 km radius around the coordinates of each cluster (baseline period 1961-1990). Source: FAO/IIASA, 2011-2012. Global Agro-ecological Zones (GAEZ v3.0). FAO Rome, Italy and IIASA, Laxenburg, Austria.

Average Precipitation

Average monthly precipitation mm/month in the 10km radius around the coordinates of each cluster (baseline period 1961-1990). Source: CRU CL 2.0 data from **?**.

Land Suitability

Average land suitability in the 10 km radius around the coordinates of each cluster. Source: Ramankutty, Foley, Norman, and McSweeney (2002).

Mean Elevation

Average elevation in the 10 km radius around the coordinates of each cluster. Source: National Oceanic and Atmospheric Administration (NOAA) and U.S. National Geophysical Data Center, TerrainBase, release 1.0 (CD-ROM), Boulder, Colo.

Average Humidity

Average relative humidity (%): how much water vapor is in the air in the 10 km radius around the coordinates of each cluster. Source: data from **?** accessed through Atlas of Biosphere.

Ruggedness

Average ruggedness (Terrain Ruggedness Index, 100 m) in the 10 km radius. Source: data from Terrain Ruggedness Index originally devised by Riley, DeGloria, and Elliot (1999), accessed through http://diegopuga.org

Tse Tse Suitability

Data Construction: Malaria and Ethnic Salience

Average predicted suitability for tsetse flies in the 10 km radius. Source: predicted suitability for tsetse flies is constructed as the sum of predicted suitability (0 to 1) for the presence of Tsetse groups (Fusca, Morsitans and Palpalis). Data produced for FAO - Animal Health and Production Division and DFID - Animal Health Programme by Environmental Research Group Oxford (ERGO Ltd) in collaboration with the Trypanosomosis and Land Use in Africa (TALA) research group at the Department of Zoology, University of Oxford.

Ethnic Homeland

Variable indicating the ethnic homeland where the DHS cluster falls, based on the map Murdock Africa map (1959), digitized by Nunn (2009).

Region Level Variables

Number of Groups in the Region

Number of different ethnic groups in the region, retrieved through the sample of all respondents in the region. Source: Afrobarometer (variable $q3$).

Group Fractionalization in the Region

Ethnic fractionalization in the region, retrieved through the sample of all respondents in the region. Source: Afrobarometer (variable $q3$).

Table 17: Summary statistics: Ethnic Identification and Malaria Stability

Variable	Mean	Std. Dev.	Min.	Max.	N
Ethnic Salience	2.551	1.174	1	5	19809
Malaria Stability	10.963	10.12	0	36.04	19809
Average Humidity	64.381	11.397	23	90	19809
Terrain Ruggedness	92673.797	155789.959	0	1460313.5	19809
Elevation	789.494	634.625	0	2768	19809
TseTse Suitability	0.21	0.271	0	1	19809
Precipitation	85.181	37.34	1.058	249.726	19809
Mean Suitability	0.461	0.229	0	0.988	19809
Average Temperature	22.635	4.395	0	30.104	19809
Night Lights	7.075	12.599	2	63	19809
Population Density	475.81	1388.337	0.004	10124.006	19809
Respondent Urban	1.634	0.482	1	2	19809
Female Age	27.928	6.103	18	68	19809
Respondent Group Size in the Region	108.277	99.31	1	425	19809
Number of Groups in the Region	8.087	5.894	1	27	19809
Group Fractionalization in the Region	0.431	0.26	0	0.899	19809
Ethnic Group Share in the Region	0.573	0.334	0.002	1	19809

Figure 1: Spatial Distribution of Ethnologue polygons

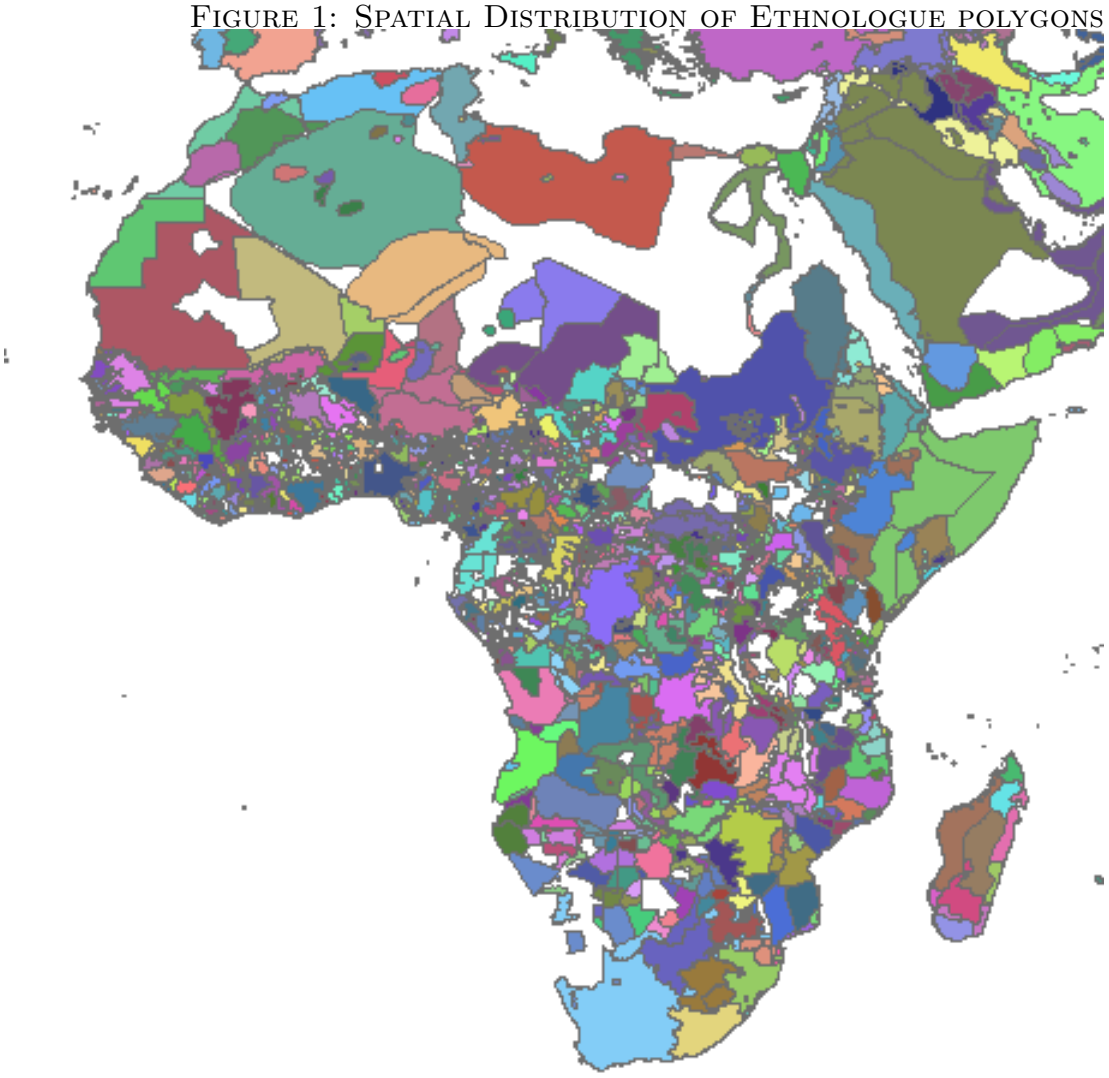


Figure 2: Malaria Stability Index

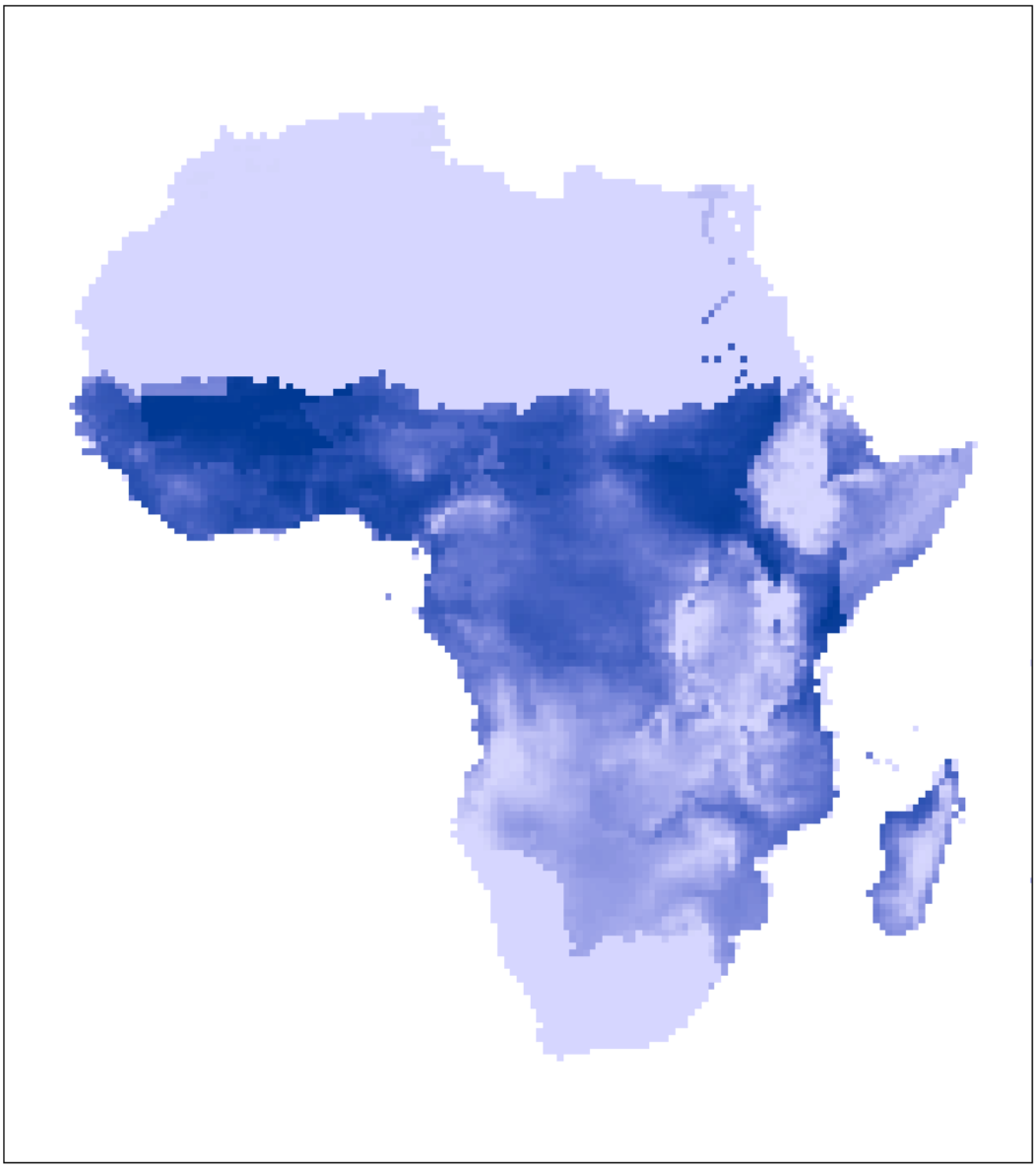


Figure 3: Murdock's Maps - Africa and America

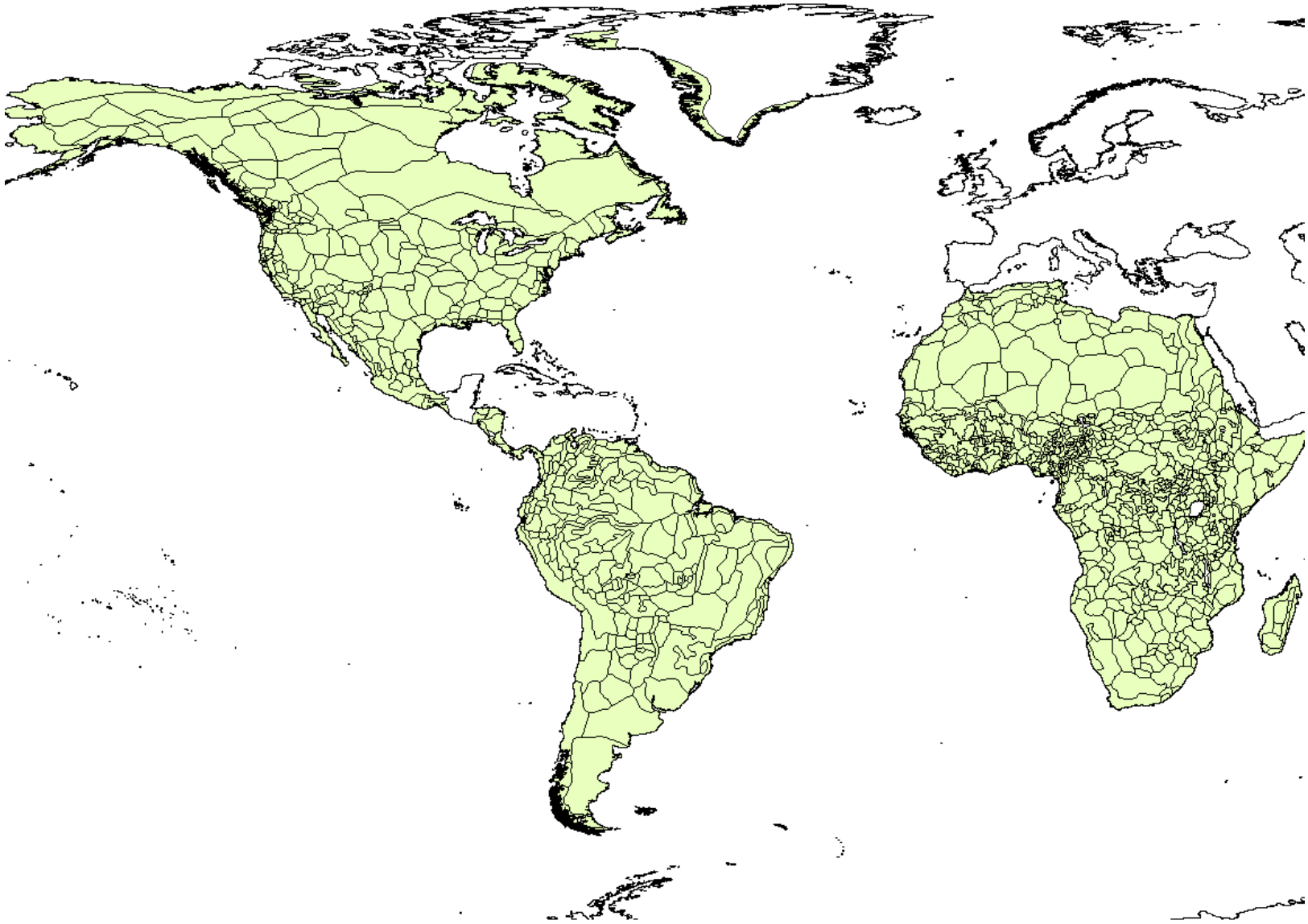
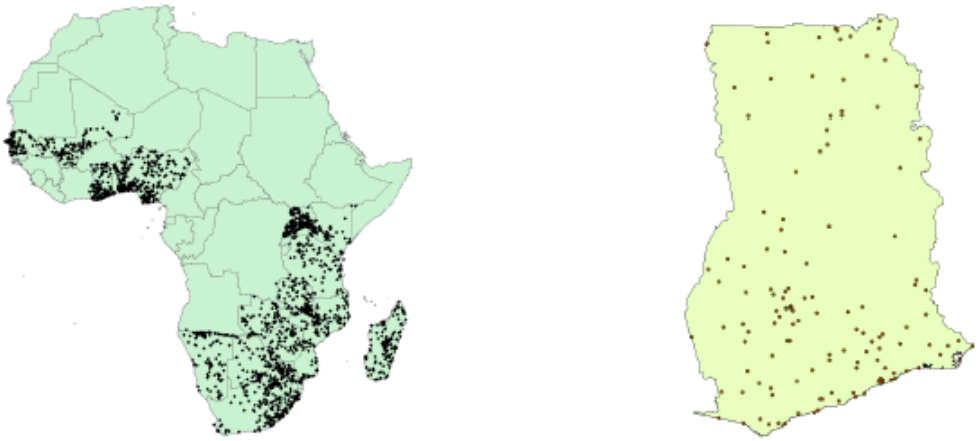
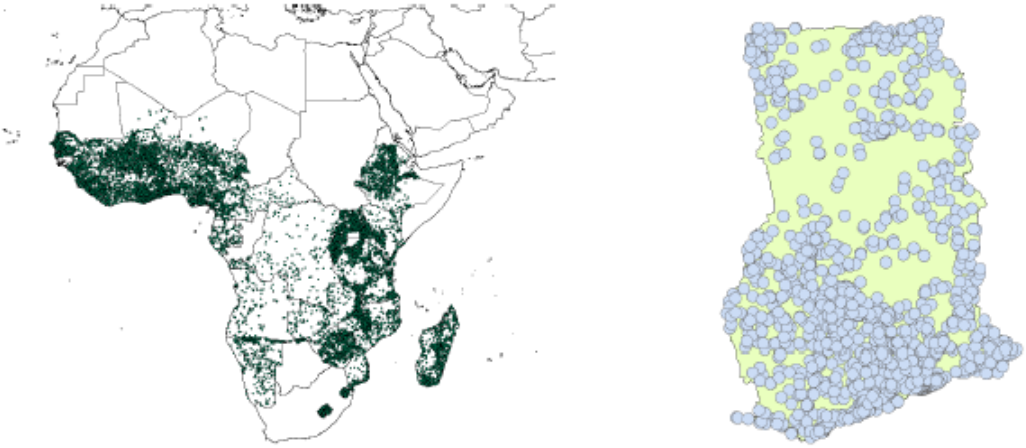


Figure 4: Afrobarometer - Africa and Ghana



(a) Across Africa (b) Within Ghana

Figure 5: Demographic and Health Survey - Africa and Ghana



(a) Across Africa (b) Within Ghana