DISCUSSION PAPER SERIES

DP13437

BITE AND DIVIDE: MALARIA AND ETHNOLINGUISTIC DIVERSITY

Matteo Cervellati, Giorgio Chiovelli and Elena Esposito

MACROECONOMICS AND GROWTH

ISSN 0265-8003

BITE AND DIVIDE: MALARIA AND ETHNOLINGUISTIC DIVERSITY

Matteo Cervellati, Giorgio Chiovelli and Elena Esposito

Discussion Paper DP13437
Published 09 January 2019
Submitted 30 December 2018

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

BITE AND DIVIDE: MALARIA AND ETHNOLINGUISTIC DIVERSITY

Abstract

We investigate the epidemiological origins of ethnic diversity and its persistence. First, we conceptualize the role of malaria for the incentives to voluntary isolation in a Malthusian environment. The theory predicts that interactions in multiple geographically clustered groups with high sexual endogamy allowed limiting disease prevalence and increasing group fitness in pre-modern populations exposed to malaria. Second, using disaggregate level data, we document the hitherto unexplored and robust role of malaria for pre-colonial, historical and contemporaneous ethnic diversity in Africa. Third, falsification tests based on malaria epidemiology and history further allow us to validate the specific predictions of the model. No effect can be detected for other placebo vector-borne diseases. Malaria is a main driver of pre-colonial ethnic diversity in Africa but not in the Americas, where the pathogen was absent before European colonization. Fourth, the effect of ancestral malaria on endogamic cultures is the main predicted channel for the persistence of African ethnicities. Exploiting within village variation across 18 African countries, we find that ancestral malaria, but not malaria today, still affects the differential persistence of ethnicities through its legacy of active endogamic cultures.

JEL Classification: N10, N30, O10, O40, Z10

Keywords: Malthusian Theory, Ethnic Groups, Cultural and Genetic Selection, Malaria, Endogamy, African Growth

Matteo Cervellati - m.cervellati@unibo.it
University of Bologna and CEPR

Giorgio Chiovelli - gchiovelli@london.edu
London Businnes School

Elena Esposito - elena.esposito.1@unil.ch
University of Lausanne

Acknowledgements
We have greatly benefitted from extensive comments and helpful discussions with several colleagues. In particular, we would like to thank Alberto Alesina, Marcella Alsan, Francesco Amodio, Quamrul Ashraf, Maristella Botticini, Francesco Cinnirella, Marcus Feldman, Christelle Dumas, James Fenske, Oded Galor, Nicola Gennaioli, Boris Gershman, Roland Hodler, Eliana La Ferrara, Sara Lowes, Marco Manacorda, Andreas Madestam, Stelios Michalopoulos, Giovanni Prarolo, Nathan Nunn, Ignacio Ortuno-Ortin, Elias Papaioannou, Luca Pensieroso, Uwe Sunde, Andrea Tesei, Felipe Valencia, Lore Vandewalle, Romain Wacziarg, David Weil and Fabrizio Zilibotti. We are also grateful to the audience at the University of Bologna, Bonn MacroHist, the joint session of the Political Economy and Income Inequality and Macroeconomics NBER Summer Institute, the Brown Workshop on the Deep Roots of Development, the Swiss Economic Development Network, Graduate Institute in Geneva, University of Geneva, Fribourg, King's

College London, and at the workshop on Institutions, Culture, and Long-Run Development in Munich. Cervellati gratefully acknowledges financial support from the Italian Ministry for University and Research (project PRIN 2015T9FYZZ). Esposito gratefully acknowledges the support of the Monique de Meuron programme for Academic Promotion.

Bite and Divide: Malaria and Ethnolinguistic Diversity*

Matteo Cervellati
University of Bologna
CEPR, CESIfo and IZA

Giorgio Chiovelli
London Business School

Elena Esposito
University of Lausanne

December 30, 2018

Abstract

We investigate the epidemiological origins of ethnic diversity and its persistence. First, we conceptualize the role of malaria for the incentives to voluntary isolation in a Malthusian environment. The theory predicts that interactions in multiple geographically clustered groups with high sexual endogamy allowed limiting disease prevalence and increasing group fitness in premodern populations exposed to malaria. Second, using disaggregate level data, we document the hitherto unexplored and robust role of malaria for pre-colonial, historical and contemporaneous ethnic diversity in Africa. Third, falsification tests based on malaria epidemiology and history further allow us to validate the specific predictions of the model. No effect can be detected for other placebo vector-borne diseases. Malaria is a main driver of pre-colonial ethnic diversity in Africa but not in the Americas, where the pathogen was absent before European colonization. Fourth, the effect of ancestral malaria on endogamic cultures is the main predicted channel for the persistence of African ethnicities. Exploiting within village variation across 18 African countries, we find that ancestral malaria, but not malaria today, still affects the differential persistence of ethnicities through its legacy of active endogamic cultures.

JEL Classification: N10, N30, O10, O40, Z10.

Keywords: Malthusian Theory with Endogenous Health, Emergence and Persistence of Ethnic Groups, Heuristics, Cultural and Genetic Selection, Geography of Ethnic Diversity, Long-Term Exposure to Malaria, Ethnic Endogamy, Legacy of Pre-Colonial Past, Disaggregate Data, Individual Data, Determinants of African Development.

*We have greatly benefitted from extensive comments and helpful discussions with several colleagues. In particular, we would like to thank Alberto Alesina, Marcella Alsan, Francesco Amodio, Quamrul Ashraf, Maristella Botticini, Francesco Cinnirella, Marcus Feldman, Christelle Dumas, James Fenske, Oded Galor, Nicola Gennaioli, Boris Gershman, Roland Hodler, Eliana La Ferrara, Sara Lowes, Marco Manacorda, Andreas Madestam, Stelios Michalopoulos, Giovanni Prarolo, Nathan Nunn, Ignacio Ortuño-Ortín, Elias Papaioannou, Luca Pensieroso, Uwe Sunde, Andrea Tesei, Felipe Valencia, Lore Vandewalle, Romain Wacziarg, David Weil and Fabrizio Zilibotti. We are also grateful to the audience at the University of Bologna, Bonn MacroHist, the joint session of the Political Economy and Income Inequality and Macroeconomics NBER Summer Institute, the Brown Workshop on the Deep Roots of Development, the Swiss Economic Development Network, Graduate Institute in Geneva, University of Geneva, Fribourg, King's College London, and at the workshop on Institutions, Culture, and Long-Run Development in Munich. Cervellati gratefully acknowledges financial support from the Italian Ministry for University and Research (project PRIN 2015T9FYZZ). Esposito gratefully acknowledges the support of the Monique de Meuron programme for Academic Promotion. Contacts: Matteo Cervellati (Department of Economics) m.cervellati@unibo.it, Giorgio Chiovelli (London Businnes School) gchiovelli@london.edu, and Elena Esposito (Faculty of Businnes and Economics) elena.esposito.1@unil.ch

1 Introduction

A large scholarly literature documents the role of ethnicities for economic and socio-political outcomes, particularly in Sub Saharan Africa. African ethnicities are deeply rooted in the concept of ancestral homelands and on endogamic cultures.[1] Understanding the emergence of ethnic groups, the peculiarity of the intense African ethnic phenomenon and the mechanisms of its persistence is key to understand its role for contemporary outcomes.

The genesis of ethnic groups has been interpreted, similarly to the process of speciation in ecology, as resulting from a sufficiently long process of differential interactions across subgroups of mankind. Besides geographic isolation and natural barriers, cultural anthropologists emphasize the key role of voluntary, or behavioral, isolation. A large ethnographic literature, discussed below in Section 2, documents cultures that actively, and sometimes fiercely, enforced limited interactions across groups and sexual endogamy.[2] Why would early humans take costly effort to enforce these behaviors, that involve both socio-economic and genetic (health) costs, is a question that has intrigued scholars for a long time.

Locating the drivers of behavioral isolation is a quest that proved conceptually elusive and, so far, is empirically unexplored. In this paper we put forwards, conceptualize and test the hypothesis that the long-term exposure to the intense selective pressure from malaria played a main role in the emergence and persistence of ethnic groups and cultures in Africa. Anthropologists interpreted the emergence of cultures regulating interactions with strangers and mating behavior as instrumental to limit diseases and to increase the survival of the group.[3] The pathogen that most intensively plagued Africans for thousands of years is malaria, the disease caused by plasmodium falciparum parasites that are exclusively transmitted by anopheles mosquitoes.[4]

[1] "*In Africa, ethnic identity is above all other things a territorial identity. Nothing defines the ethnic group better than its "standing place". Thus the term geoethnicity has been used to describe the African ethnic phenomenon.*", Cobbah (1988). "*Ethnic boundaries are created socially by preferential endogamy and physically by territoriality. (..) The prototypical ethny is thus a descent group bounded socially by inbreeding and spatially by territory.*" Van der Berghe (1987). The role of homelands and ancestry is emphasized also by Horowitz (1985).

[2] In contrast to animal species that cannot interbreed, the emergence and persistence of ethnic groups also crucially rest on active limitations of sexual admixing. "*Sheer physical propinquity determines who has sexual access to whom. Geographical barriers (..) isolate animal populations from each other, and create breeding boundaries between them, that can and often do lead to speciation or subspeciation. In humans, however, the story does not stop there. In addition to the purely physical impediments (..) human groups create cultural prescriptions and proscriptions concerning their mating systems. There is not a single known human group that lacks them and that even approximates panmixia* [random mating]", Van der Berghe (1987).

[3] Scholar have interpreted the evolution of human groups as being driven by the relative benefits of different traits and behavior. Cultural change has been modelled in various ways ranging from the selection of groups, traits and heuristics of behavior. A common prediction is that they should be expected to maximize some measure of fitness irrespective of whether the groups know, or understand, that they are beneficial and why. See also Section 2.

[4] In'*Humanity's Burden: A Global History of Malaria*", epidemiological historian Webb (2009) makes the point that "*[Malaria is] a primordial companion of our distant protohuman ancestors and an even earlier companion of the chimpanzees from which we branched off six or seven million years ago*". Malaria (in the vivax type) accompanied the emergence of homo sapiens in Africa more than 100,000 years ago while the most deadly variant, falciparum, plagued

The role of long-term exposure to malaria in pre-modern environments is conceptualized, in Section 2, extending a Malthusian framework to endogenous health by incorporating well-established insights from malaria epidemiology and evolutionary genetics. The theory formalizes the hypothesis that the intense and prolonged exposure to the pathogen favored the emergence of patterns of behavioral isolation in early human settlements. In particular, it predicts that limiting the interactions across multiple geographically clustered groups and enforcing high sexual inbreeding helped containing the prevalence of the disease and sustaining population density.[5] The emergence of endogamic cultures is interpreted as the main channel of persistence of these groups until today.[6] The predictions are in line with a body of scholarly arguments and scattered historical narratives, discussed in Section 2, but are hitherto not empirically investigated.

We empirically explore the main insights on the role of malaria for ethnic groups in Section 3. The baseline results document that exposure to malaria is one of the main and most robust drivers of the number of historical ethnic groups across different locations (grid cells) in Africa. The patterns are robust to a large set of checks. Given the nature of the data, we devote particular attention to assess the role of location-specific unobserved characteristics, the use of alternative measures of malaria exposure and their potential endogeneity, the role of grid cells of different size and the existence of measurement errors in the drawing of ethnic borders.[7]

To explore the specific predictions, and to improve identification, we devise a set of empirical exercises and falsification tests based on the peculiarities of malaria epidemiology and of its global epidemiological history. First, geographic clustering and sexual inbreeding are predicted to be beneficial for malaria but not for other important vector-borne diseases affecting Africans, like trypanosomiasis, dengue and yellow fever (as discussed below). The role of these placebo diseases is systematically explored to validate the specific role of malaria in each of the empirical exercises. Second, malaria was imported in the New World only after European colonization in the context of the so-called Columbus exchange. We assemble a novel database that allows us to document that malaria is a main driver of pre-colonial diversity in Africa while it has no explanatory power for pre-colonial ethnicities in the Americas (as placebo).[8] The analysis finally studies the persistent legacy

Africans for more than 10,000 years.

[5]The process of cultural and genetic evolution of different groups, related to the benefits of behavioral isolation, is interpreted as being shaped by the specific disease ecology. Technically, geographic clustering in multiple groups without visitations limits the incidence of malaria by reducing the size of the human host group in each location. Sexual inbreeding increases homozygosity of genetic traits and favors the spread of group-specific genetic immunities to malaria.

[6]From an epidemiological standpoint, behavioral isolation is not expected to have first-order effects on malaria prevalence in modern African economies characterized by high population density, increased contacts between the population across locations and improved access to medical treatments and prevention. Present-day exposure to malaria should not be expected to have a direct effect on the persistence of ethnic cultures.

[7]The analysis also accounts for, and studies, the pre-colonial ethnographic characteristics of these groups.

[8]Further in line with these findings, robustness checks document that malaria affected the distribution of historical groups in all the Old World.

of malaria on groups admixing and ethnic identities at the village level and on the distribution of ethnolinguistic groups in Africa today.

The predicted channel of persistence is based on the role of ancestral malaria exposure for the emergence and perpetuation of endogamic cultures. Section 4 offers a conceptualization based on ethnolinguistic distances and measures of ethnic endogamy using individual data for marriages from the DHS surveys.[9] Identification exploits within-village variation looking at the marriage choices of respondents not residing in their ancestral ethnic homeland. The results document that higher ancestral malaria, but not current malaria in the location, strongly increases the likelihood of endogamic marriages (or reduces exogamic ethnolinguistic distances).[10] The patterns are robust to several checks and no evidence for any significant role of (placebo) diseases other than malaria is detected. The results, that are insightful also for the process of differential survival of groups, suggest that ethnicities shaped by the strong selective pressure of ancestral malaria still enforce endogamic marriages and, therefore, should be expected to face higher prospects of future persistence.

Literature. The paper contributes to the literature a first conceptualization of the role of malaria for the emergence and persistence of ethnicities, and a systematic empirical investigation of drivers of the spatial distribution of ethnic groups and of ethnic marriages today. The background is the large literature on the role of human diversity across countries.[11] The framework extends Ashraf and Galor (2011) to the consideration of endogenous health, spatial clustering and sexual inbreeding. While leaving the Malthusian dynamics unchanged, the results highlight the role of cultures of isolation and endogamy in shaping the patterns of pre-modern population density. A recent literature focuses on the geographic distribution of ethnic groups using disaggregate data. Alesina, Michalopoulos and Papaioannou (2015) document the role of spatial ethnic inequality for local economic development. Michalopoulos (2012) shows that physical isolation and differential agricultural suitability are important drivers of the global distribution of linguistic groups today. Our analysis contributes to this literature conducting the first systematic investigation of the determinants of the historical and pre-colonial distribution of ethnic groups. The results provide evidence for a main role of long-term

[9]We are not aware of any attempt to measure and systematically explore the drivers of ethnic endogamy in Africa. Measuring endogamy involves conceptual and practical difficulties, discussed below. We locate each individual to her/his ethnolinguistic group at different levels of the ethnolinguistic tree (exploiting a large set of data sources for matching individual ethnicities to ethnic homelands). We consider both dichotomous definitions of endogamy (at different levels of aggregation) and measures of distances between the languages spoken by the spouses.

[10]The empirical strategy allows to isolate the role of ancestral origins of individuals residing in the same location, and thus exposed to the same environment in terms of e.g. diseases, geography, social structure and levels of economic development. The results, that also account for individual characteristics, are not driven by the effect of malaria (either ancestral or current) on selective migration of individuals across different locations.

[11]See Alesina and La Ferrara (2005), Esteban, Mayoral and Ray (2012), Franck and Rainer (2012), and references therein, for the role of ethnic diversity on public good provision, conflicts and ethnic politics, respectively. The literature has studied also the related role of cross country genetic distances, Wacziarg and Spolaore (2013) and genetic diversity, Ashraf and Galor (2013).

exposure to malaria as a specific determinant of the spatial distribution of African, but not American, ethnicities above and beyond the (universal) role of geographic isolation.

The literature has documented several instances of long-term persistence of cultural attitudes. This paper relates in particular to the studies exploiting information on individuals residing outside their ancestral homeland for empirical identification, see Nunn and Wantchekon, (2011) and Michalopoulos, Putterman, and Weil (2018). The question under which condition should specific traits be expected to persist or not has proved more difficult to address, see Nunn (2012) and Alesina and Giuliano (2015) for surveys and Giuliano and Nunn (2017). Endogamic cultures have been widely, but informally, interpreted as a potential main mechanism of differential persistence of groups and attitudes.[12] We conceptualize the drivers of endogamic cultures and offer a first measurement of endogamic marriages in Africa. We build on the concept of ethnolinguistic distance by Desmet, Ortuño-Ortín, and Wacziarg (2012). The results suggest a key role of ancestral ecology in shaping patterns of differential persistence of African ethnicities and cultures.

Finally, the paper contributes to the debate on the drivers of Africa's (under)development and, in particular, on the role of geography and ethnicities. Mounting evidence suggests a key role of ethnic homelands and pre-colonial ethnic institutions, see Gennaioli and Rainer (2007), Michalopoulos and Papaioannou (2013a, 2013b, 2016) and Fenske (2014), among others. The empirical role of long-term exposure malaria, while being often advocated as "Africa's burden", is nonetheless still disputed.[13] Our results also confirm, and put into perspective, the lack of systematic effects of malaria on pre-colonial population density in Africa, see Alsan (2015) and Depetris-Chauvin and Weil (2018). By considering the so far overlooked effect of malaria on ethnic diversity, the findings help to reconcile the unclear evidence. The results suggest that the exceptional selective pressure from malaria might affect the prospects of African development but mostly through its long-lasting imprint on the emergence and persistence of its geo-ethnicities and endogamic ethnic cultures.

2 Conceptual Framework

The framework presented in Section 2.1 builds on, and extends, the Malthusian model by Ashraf and Galor (2011). Section 2.2 presents a simple malaria transmission model that follows the seminal insights by Ross (1909) and the subsequent formalizations by MacDonalds (1956). The role of sexual inbreeding is studied by postulating a simple trade-off between the costs and benefits of increased homozygosity of genetic traits in the presence of malaria. Section 2.3 studies the Malthusian station-

[12] In the specific context of ethnic cultures there is also a vast literature in sociology and anthropology but also economists have made the argument, see e.g. Caselli and Coleman (2013).

[13] Early findings of a harmful effect, see e.g. Sachs (2003), have been put into question, see e.g. Weil (2017) and references therein, among others.

ary state with endogenous health. Section 2.4 characterizes the patterns of behavioral isolation, in terms of geographic clustering and sexual endogamy, that maximize Malthusian population density. Finally, Section 2.5 discusses testable predictions in view of the available historical narratives and scholarly arguments. The derivations and proofs are in the Appendix.

2.1 Set-up

Consider a simple overlapping generation Malthusian set-up along the lines of Ashraf and Galor (2011). Generations are denoted by t.

Production. The unit of observation is a location, e.g. a grid cell, with territorial size normalized to one endowed with a time-invariant fixed factor of production X (e.g. productive land). Denote by $G \geq 1$ as the number of subgroups of individuals existing in a cell.[14] In each generation, t, population size in each location is L_t. Aggregate production is given by,

$$Y_t = (A(G)h_t(G)X)^{\alpha} L_t^{1-\alpha} \tag{1}$$

where *effective* productivity depends on the parameter $A_t(G)$ (that denotes, e.g., the level of technology), and human health, $h(G)$.

This framework nests on the standard Malthusian model as a special case. If $h_t(G) = h_t$, and absent any scale effects in productivity, so that $A(G) = A$ then income per capita in each location does not depend on the number of human groups G.[15] In general, however, the patterns of geographic clustering and interactions may affect per capita income (given population size). Splitting the human population in geographically separated groups may involve economic costs in terms of e.g. reduced scale effects or reduced benefits from interactions across groups (e.g. productive specialization and trade) so that $A'(G) \leq 0$ while, as characterized below, in the presence of malaria geographic separation of human (hosts) in subgroups can be beneficial for health $h'(G) > 0$.

Population Dynamics. Individuals derive utility from consumption, c_t, and surviving children. Fertility, n_t, is chosen to maximize

$$u_i = c_t^{1-\gamma}(\pi_t n_t)^{\gamma} \tag{2}$$

[14] As baseline, we study a pattern of social interactions involving the full separation of the human population in each cell into G geographically clustered subgroups. As discussed below this patterns of social interactions is particularly interesting from the perspective of malaria transmission.

[15] Income per capita in each location is by $y_t = (A(G)h_t(G))^{\alpha}(X/L_t)^{\alpha}$.

under the individual budget constraint

$$\rho \pi_t n_t + c_t \leq y_t \tag{3}$$

where $\pi_t(g) \in (0,1]$ is child survival. The evolution of population across generations is given by,

$$L_{t+1} = \pi_t n_t L_t \tag{4}$$

2.2 Malaria

To illustrate the predicted role of long-term exposure to malaria on the human populations, we study a simple model of malaria transmission and preferential sexual inbreeding.

Malaria Transmission. Consider a simple malaria transmission model with spatial clustering of the human population in subgroups along the lines of the mathematical formalizations that follow MacDonalds (1956).[16] Malaria transmission is characterized by the dynamic system

$$\begin{aligned} \dot{\sigma} &= (1-\sigma) \times \mu M \times s - r\sigma \\ \dot{\mu} &= (1-\mu) \times \sigma (L_t/g) - d\mu \end{aligned} \tag{5}$$

where σ and μ denote the share of infected humans and infected mosquitos, respectively. The intensity of exposure to malaria depends on mosquito density M and the probability of developing the infection upon inoculation of the pathogen, $s \leq 1$, that as discussed below depends on the level of immunities in the population. The parameters r and d denote the recovery rate of humans and the death rate of mosquitos.

The first equation of the system implies that, for given recovery rate r, the share of non-infected humans $(1-\sigma)$ that get infected increases with the size of the infected mosquitos population, given by μM, times the probability of developing the diseases upon being inoculated, s. Similarly the increase in the share of infected mosquitos in the second equation of the system, depends on the probability that a non-infected mosquito $(1-\mu)$ bites an infected human that, under the extreme assumption of no visitations between subgroups, is simply given by the infected human population in each spatial partition: $\sigma (L_t/G)$.[17]

[16]The epidemiological literature on mathematical models of malaria transmission is by now extensive; see Mandal et al. (2011) for a survey.

[17]The case of no visitations is an interesting benchmark because it delivers the largest gains for health (by fully separating the different host populations), it allows to derive a closed form solution of the stationary state of the system (5) and it permits an analytical characterization of the Malthusian equilibrium with endogenous health and to derive testable predictions by means of comparative statics (see Section 2.3 below). The epidemiological literature has studied partial patterns of visitations across multiple, partially interconnected, host populations by means of numerical

Immunities: Genetic Selection and Sexual Inbreeding. Human susceptibility to the *plasmodium falciparum* parasites is reduced by several genetic immunities. The strong selective pressure imposed by the pathogen on humans over the last thousand years materialized in an abnormal spread of several malaria protective blood disorders, which have been linked to the intensity of malaria exposure and to the degree of sexual inbreeding. Higher selective pressure from the pathogen, proxied in the model by M, increases the evolutionary advantage of malaria protective genetic traits and the likelihood that they are transmitted to descendants. Higher levels of sexual inbreeding, denoted by $e \geq 0$, raise homozygosity of genetic traits including, crucially, the malaria protective ones.[18] As a result, malaria is the strongest known selector of the human genome and the most common monogenetic human diseases are malaria protective.[19]

For simplicity, we incorporate these well-established insights in reduced form by assuming that the probability of developing the disease upon inoculation of the pathogen can be represented by a separable function,

$$s = f(M, e) \tag{6}$$

Recalling that higher spread of genetic immunities implies a reduction of disease incidence, we assume that the function (6) is decreasing in both the intensity of malaria exposure, M and the level of endogamy, e, so that: $f_M < 0$ and $f_e < 0$.[20]

2.3 Malthusian Equilibrium with Endogenous Health

Endogenous Health. Higher homozygosity of genetic traits is generally detrimental to human health because of the negative effects of recessive lethal diseases. We denote $p(e)$, with $p_e(e) > 0$, $p_{e,e}(e) < 0$ and $p(0) > 0$, as the probability of developing genetic diseases. The probability of being healthy is given the joint probability of not being infected by malaria (given by $1 - \sigma$) and not being subject to the incidence of genetic diseases $(1 - p)$:

$$h \equiv (1 - \sigma) \times (1 - p) \tag{7}$$

simulations. The higher the frequency of visitations, the higher the prevalence of the disease which is maximal in the limit case in which humans freely move across space (effectively interacting within a unique group, $G = 1$). See, among others, e.g Rodriguez and Torres-Sorando (1997) and Mandal et al. (2011) for a survey.

[18]The variable e is interpreted as the level of "preferential" sexual inbreeding with $e = 0$ denoting random mating. For simplicity we abstract from the explicitly modeling the distinction between sexual inbreeding across extended families and sexual inbreeding within the group (which tend to coincide when the groups are small).

[19]Higher selective pressure increases the spread of non-lethal recessive and co-dominant monogenetic diseases like α-thalassemias, G6pd deficiency and Duffy negative antigen since they involve reproductive advantages to their carriers. Higher sexual inbreeding increase the spread of malaria protective traits for any level of selective pressure. See e.g. Kwiatkowski (2005), Sabeti et al. (2006) and Denic and Nicholls (2007).

[20]To restrict attention to a unique interior maximum, we also assume that genetic immunities attenuate the negative effect of exposure to malaria: $|\varepsilon_{s,M}| < 1$. See also derivation below and in the Appendix.

Conditional on population density and the patterns of social interactions (number of groups and endogamy), the equilibrium level of health is given by,

Lemma 1 (Health)**.** *For any* $\{L_t, G, M, e\}$*, the level of health in an interior stationary state of the malaria transmission model (5) is given by,*

$$h_t^*(L_t, G, M, e) = \left(1 + \frac{d \times G}{L_t}\right)\left(\frac{r}{r + f(e, M) \times M}\right)(1 - p(e)) \tag{8}$$

Proof. Given (7) the equilibrium level of health is obtained by characterizing the interior stationary state of the malaria transmission model (5), by setting $\dot{\sigma} = \dot{\mu} = 0$ as reported in Section A1.1 in the Appendix. □

The level of health decreases with population density, L_t, and in the probability of infection $s \times M$. Limiting interactions within geographically clustered subgroups of humans, G, allows to limit the negative effects of malaria by reducing the size of the parasite host human population for any given population density L_t. Finally, the effect of sexual inbreeding, e, is generally ambiguous and depends on the intensity of exposure to malaria.

Malthusian Stationary State. Lemma 1 offers a simple formalization of the effect of malaria on health that, as characterized next, represents a main driver of Malthusian population density.[21] The level of health has been so far characterized conditional on population density. We next get back to the characterization of the stationary level of population density in the stationary state with endogenous fertility and health. Recall that population dynamics evolve according to (4). We have the following,

Lemma 2 (Malthusian Stationary State)**.** *There exists a unique level of population in the Malthusian stationary state, denoted as* L*, which is implicitly characterized by*

$$L = \left(\frac{\gamma}{\rho}\right)^{1/\alpha} A(g) X \times h^*(L, G, M, e) \tag{9}$$

where $h^*(L, G, M, e)$ *is equilibrium health as derived in (8).*

Proof. Equation (9) is obtained by substituting optimal net fertility, which is proportional to income per capita, and restricting attention to the stationary state of population ($L_t = L_{t+1} = L$). See also Section A1.2 in the Appendix. □

[21] Malaria infection also seriously affect health, and mortality, of children. Child mortality is not, however, a main predicted driver of the level of Malthusian population density in the stationary state. As well studied in the endogenous fertility literature, and as characterized in the Appendix, the reason is that the first order effect of child mortality is on gross, and not net, fertility.

Notice that the consideration of endogenous health does not change the Malthusian nature of the theoretical framework and, accordingly, its main features are unchanged.[22] In absence of malaria, the level of population density in each location only depends on differences in technology and productive endowments. From (8) health is decreasing in the level of population density. The presence of malaria, therefore, produces a countervailing effect on population in the Malthusian stationary equilibrium (9) that, for any given number of groups and endogamic behavior G and e, tends to decrease with the intensity of malaria.

2.4 Optimal Behavioral Isolation: Spatial Clustering and Endogamy

Researchers in economics have studied the evolution of social behavior in terms of group selection, selection of traits and evolution of heuristics of behavior.[23] Cultural anthropologists have, more specifically, interpreted the emergence of cultures regulating interactions with strangers and mating behavior, as specifically instrumental for the health and the survival of the group. The "minimax" theory posits that "*cultural systems tend to favour practices which minimise the risk of disease and maximise the health and welfare of groups*", Alland (2008).[24] Regardless of the specific approach, a common prediction is that the evolution of social norms and cultural traits should be interpreted as being driven by the relative benefits of different social traits. Importantly, these differential traits should be expected to emerge in the long run even if the reasons why they were beneficial is not known or understood by the population.[25]

Following this perspective, we look at the norms of behavioral isolation that maximize group welfare that, from a Malthusian perspective, is proxied by the stationary level of population density. We have the following,

Proposition 1 (Optimal Number of Groups and Sexual Endogamy)**.** *Consider $A(G)$ with $A'(G) < 0$. There exists a unique interior solution for the number of geographically clustered groups, G and level of preferential sexual endogamy, $\{G^*, e^*\}$, that maximizes stationary state population density as implicitly characterized by (9) with,*

[22]This implies that, as in Ashraf and Galor (2011), heterogeneity in either technology A, factors of production X, and equilibrium health h, all affect the level of population density but not the stationary level of income per capita in each location.

[23]The question of the evolution of social norms is related to the selection of social preferences and group selection (see e.g. Bergstrom (2002)), the evolution preferences (see and Robson and Samuelson (2010), Galor and Michalapolous (2012) and Galor and Ozak (2015)) and heuristics, see Nunn and Giuliano (2017) and Nunn (2012), for a critical survey.

[24]Also in social biology, the emergence of in-group norms of mating has been linked to the level of parasites stress, see e.g. Fincher and Thornhill (2012). A main limitation of this literature, particularly concerning the predictions of the emergence of in-group norms, is that the epidemiological reasons why different social behaviors should be more or less beneficial for health in the face of different pathogens is not spelt out or formalized.

[25]A well known example of heuristic of behaviors is by Webb (2009) that points out how African populations living in malaria-infested regions traditionally relied extensively on tuberous foodstuffs such as yams and cassava, which have been subsequently scientifically discovered to be effective in partially inhibiting the reproduction of the parasite.

1. *the optimal number of spatially clustered groups, G^* is strictly* ***increasing*** *in malaria exposure, $(s^*(M) \times M)$ for any $G^* > 1$;*

2. *the optimal level of preferential endogamy, $e^*(M)$, is strictly* ***increasing*** *in malaria exposure, M for any $e^*(M) > 0$;*

Proof. The effect of malaria on optimal behavioral isolation is characterized by explicitly solving for the equilibrium level of population density in the interior stationary equilibrium and by studying the comparative statics effect of malaria by means of implicit function theorem. The derivation is analytically involved because of the implicit characterization of the optimal number of groups and sexual endogamy. The intuition behind the result is, nonetheless, straightforward. At an interior optimum a marginal increase in M, while not affecting the marginal cost (which is only related to $A'(G)$) increases the marginal health benefits of spatial clustering (which related to $h'(G)$). Similarly, while not affecting the marginal cost on health (that is related only to $p'(e)$), higher malaria implies higher marginal benefits of sexual inbreeding on genetic immunities ($f'(e)$). See the derivation in Section A1.3 in the Appendix. □

2.5 Empirical Predictions and Historical Narratives

The conceptual framework models a simple Malthusian environment with malaria. The theory formalizes the predictions that the intense and prolonged exposure to the pathogen favored the emergence of patterns of behavioral isolation that helped limiting the prevalence of the disease in premodern human settlements. More specifically, areas characterized by a strong selective pressure from malaria should host multiple geographically clustered, stand-alone, human groups enforcing limited interactions across groups and featuring endogamic ethnic cultures sustaining patterns of preferential sexual endogamy.

Scholars produced a body of historical narratives on the role of cultures of isolation and sexual endogamy as adaptive responses to the local malaria ecology. For Diamond "*Their [Africans] entire civilization had evolved to help them avoid infection in the first place*" and, in particular, that "*Africans were combating malaria with more than just antibodies (...) by living in relatively small communities, .. [to] limit the level of malaria transmission*" (Diamond 1997, 2011). Historians interpreted the fierce enforcement of behavioral isolation and limited interactions in malarial areas, notably including explicit inter-groups hostility, as a strategy for limiting the disease. Early observer Lambert (1928, p. 368) suggested that malaria spread was traditionally contained by "*an ancient quarantine of intertribal enmity*". Many narratives document practices that limited contacts even when interactions across groups were profitable, or needed, like in the case of trade. For King (2013), limited contacts and exchanges between groups were common practice as "*Malaria proved a constant*

obstacle to trade [...] [traders] would only trade through a system of silent barter" (see also Ramen, 2002). The predictions on the emergence of endogamic cultures favoring sexual inbreeding is in line with the recent literature on evolutionary genetics. Denic and Nicholls (2007) summarize the existing evidence as providing *"strong support to the hypothesis that the culture of consanguineous marriages and the genetics of protection against malaria have co-evolved by fostering survival against malaria through better retention of protective genes".*[26]

The consideration of a process of relocation of groups across space, while not explicitly modelled, tend to reinforce the predictions on the role of malaria for the emergence of geo-ethnicities. Similar to animal species that are highly specialized to local ecology, the intense cultural and genetic adaptation to malaria in humans should be expected to have favored limited relocation of existing groups across space.[27] In line with this, Webb (2009) argues that adaptation to, and avoidance of, malaria underscored the distinct local cultures leading to the "*the tapestries of ethnicity*" that we observe today. Avoidance and limited admixing in malarial areas have been largely documented, see e.g. McNeill (1976) who argue that malaria also induced well adapted primitive communities from leaving their homelands and from being replaced in their territories remaining separated or semi-autonomous, see also Endalew (2006) and Brower and Johnston (2007).[28] In line with these arguments, Reich (2018) imputes the African "tassellated" patterns of population structure (i.e. the existence of areas of genetic homogeneity demarcated by sharp boundaries), to the limited migration.

The existing wealth of historical narratives and arguments offer background evidence of the potential relevance of the role of long-term exposure to malaria for the process of emergence and persistence of African ethnicities. A systematic exploration of the empirical role of malaria for ethnic diversity and tests of the specific predictions derived in Section 2 is, nonetheless, still missing. Section 3 offers a first attempt to test the predicted impact of long-term exposure to the pathogen for the number and geographic distribution of the ethnic groups. Section 4 studies the predicted mechanisms of persistence through endogamic ethnic marriages.

[26]That the prevalence of sexual inbreeding and the frequency of alleles protective against malaria are both very high in malarial areas is well documented, see e.g. Webb (2006) and Denic et al. (2008). Notice that this perspective is related to, but is different from, the arguments emphasizing the role of culture for genetic selection, along the lines of Cavalli Sforza and Feldman (1981), rather than interpreting the two as mutually affecting each other.

[27]The reason is that, e.g. groups that developed resistance to the pathogen would lose their comparative advantage by leaving their ancestral homelands while groups that are unfit would avoid resettling into malarial areas.

[28]A well documented case study is offered by the Tharu people that after centuries of residence in malaria-infested regions developed a high genetic resistance to malaria facing an about sevenfold lower malaria incidence than non-Tharu people, see Modiano et al. (1991). Strict endogamy practiced by the Tharu confined these traits to this indigenous group and preserved their location-specific advantage allowing them to live undisturbed from neighboring, often more powerful, groups.

3 Long-Term Malaria Exposure and Ethnic Diversity

3.1 Empirical Strategy and Data: Grid-Cell Analysis

We start by exploring the prediction that long-term exposure to malaria increases the number of groups in each location. Following Michalopoulos (2012), the baseline analysis is conducted across equally-sized grid-cells. We use cells of 1x1 degree of size (about 110 km at the equator) as units of observation.[29] As baseline empirical specification we estimate the relationship between the (log of) the number of ethnic groups (in the 1x1 degree cells) according to:

$$LogNumberGroup_{i,c} = \beta_0 + \beta_1 Malaria_{i,c} + \beta_2 \mathbf{X_{i,c}} + \mu_{\mathbf{c}} + \epsilon_{\mathbf{i,c}} \tag{10}$$

where i indicates the cell, and c the country. The dependent variable $LogNumberGroup$ is the natural logarithm of the number of ethnic groups in cell i. The vector, $\mathbf{X_{i,c}}$, indicates the set of covariates. All specifications control for the natural logarithm of the cell land area. In some specifications, particularly, when looking at historical data for the mid-twentieth century, we also include country fixed effects, μ_c.

Data: Geographic Distribution of Ethnicities. The empirical analysis exploits information on the geographic distribution of ethnicities at different points in time, that are used to build measures of the number of ethnic groups in each grid cell.

We start the empirical analysis by studying, in particular, the determinants of the number of historical and the pre-colonial ethnic groups in Africa. Section 3.2 looks at the historical location of the different ethnic groups, exploiting information from the Geo-Referencing of Ethnic Groups (GREG) database. This is the digitized version of the Soviet Atlas Narodov Mira from the 1960s, which depicts the spatial distribution of 226 ethnic groups in Africa in the first half of the twentieth century.[30] In Section 3.3 we next exploit information on the spatial distribution of the homelands of pre-colonial ethnic groups retrieved from the work of George Peter Murdock (1951, 1957 and 1959). The Murdock maps, which are based on anthropological and archaeological data, depicts the borders of the pre-colonial (ancestral) homelands of culturally homogeneous ethnicities, rather than the historical location of their descendants.[31] By looking at pre-colonial ethnicities, the Murdock maps also allow us to explore the drivers of ethnic diversity in each location both in Africa and in the

[29] Looking at squared grids allows to have equally sized randomly assigned units of observation in each location. As discussed below, the robustness of the patterns is explored using several strategies including the replication of the analysis for a full set of alternative sizes of grid cells.

[30] In a 1×1 degree cells framework, the resulting average number of GREG groups at the cell level is 2.12 with a standard deviation of around 1.2.

[31] These homelands refer to the pre-colonial geographic location of groups that are culturally homogeneous, typically at levels seven or eight of the Ethnologue as discussed further below.

Americas that, from the perspective of malaria, represents an interesting placebo.[32] Finally, to dig deeper into the role of malaria for the persistence of the geographic distribution of groups today we also exploit, in Section 3.4, data on the distribution of ethnolinguistic groups today from the World Language Mapping System (Lewis, 2009).

Data on Exposure to Malaria. As a baseline measure of long-term exposure to malaria in the different locations, we use the malaria strength and stability index, henceforth labelled "Malaria Stability", constructed by Kiszewski et al. (2004). The index is a measure of predicted exposure to malaria that is built using information on the geo-climatic conditions and the biologic characteristics of the dominant mosquitoes vector in each location.[33] The index is a long-average of malaria suitability (exploiting climatic information from the beginning of the twentieth century to the early 1990s), which makes the it less prone to climatic changes of the last decades.

To check the validity of the results the analysis is replicated using several alternative measures of malaria exposure. To account for the potential non-randomness in the type of mosquitoes, a specific effort is devoted to exploring the sensitivity of the results to alternative measures of predicted malaria exposure (discussed in further detail below). We use alternative information on geo-climatological suitability to *plasmodium falciparum*, and reconstruct alternative measures of the malaria stability index in each location without exploiting actual information on the distribution of mosquitoes. We also use proxies of the levels of malaria endemicity in the African population around the year 1900 and data on the frequency of genetic immunities to malaria.

Covariates. Given the cross-sectional nature of the empirical exercise, the baseline analysis conditions on a large set of covariates that might be relevant drivers of ethnic diversity. We condition on soil suitability and elevation (mean and standard deviation) to account for geographic isolation and incentives for productive specialization. The *geographic* covariates further account for long-term averages of temperature and precipitation, terrain ruggedness, and caloric suitability before 1500. The *distance* variables include distance from the equator, from the coast, from the river, from the

[32]Murdock's maps for Africa have been digitized and geo-referenced by Nunn (2008) and Chiovelli (2016), respectively. See also the discussion on original data and methodology in the Appendix in Section A2.2.

[33]This measure is built upon three main features. The index relates to different types of vectors responsible for malaria transmission: various species of *Anopheles* mosquitoes. Kiszewski et al. (2004) associated with each region a dominant vector of *Anopheles* mosquitoes (for countries with different dominant vectors, mosquitoes were associated with sub-regions). A monthly index of stability was then computed as a parametric function of the share of blood meals taken by the mosquitoes, the daily survival rate, and the extrinsic incubation period which vary with the type of mosquito prevalent in the region. Finally, once this regional index (constructed for about 260 regions in the world) was created, a minimum lagged threshold of precipitation (10 mm) was exploited as a pre-condition for malaria transmission. This procedure allows to obtain a finer data resolution. The original malaria Stability index ranges from 0 (absence of a sustainable environment for malaria transmission) to 34 (high potential for malaria transmission). At cell level, the mean of the Malaria Stability index in our sample is around 11 with a standard deviation around 9. Details on data sources, data construction, and summary statistics are reported in the Appendix in Section A7.

country border and from the country capital. Looking at TseTse suitability, in terms of geographic suitability for Trypanosomiasis transmission, provides information that is useful both as a covariate and as an interesting placebo tropical disease. Finally, we account for the distance from East Africa (as a measure of predicted genetic diversity). For consistency, the same sets of covariates are included in all the empirical exercises.[34]

3.2 Historical Ethnicities

We start by exploring the role of malaria for historical ethnic diversity.

3.2.1 Baseline

Figure 1 depicts the spatial variation of the average number of historical ethnic groups (GREG) and malaria stability. Figure 2 visualizes the relationship by means of binned scatter plots relating long-term malaria exposure and the log number of historical ethnic groups. Panel A reports the unconditional correlation while Panels B and C report associations between malaria and ethnic diversity, conditional on controlling for all the sets of covariates discussed above, without and with country fixed effects, respectively.

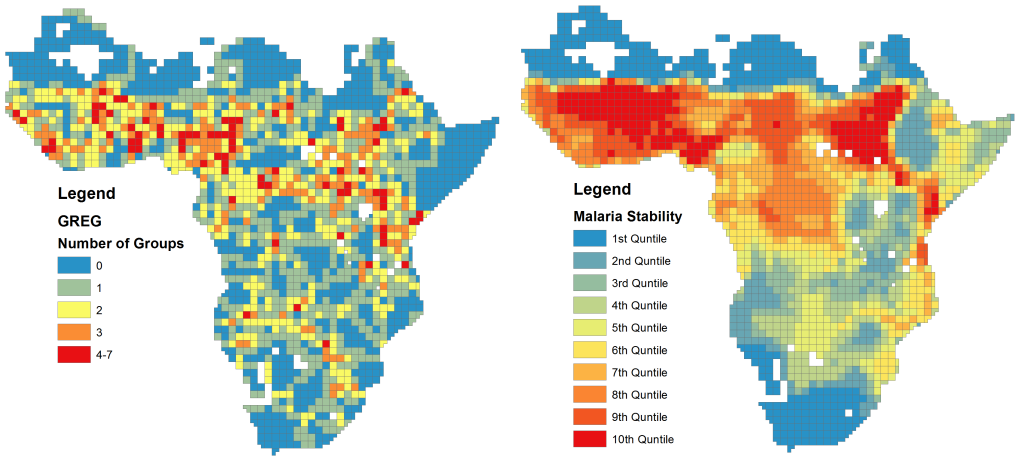
The estimation results of the empirical model (10) are reported in Table 1. Since the data on the historical distribution of the population of the different ethnic groups refers to a period where modern countries where already present, we also account for the number of countries in the cell and include a "within-country" dummy variable indicating whether the cell is fully contained in a country. To account for spatial correlation across nearby units, inference is adjusted reporting both Conley standard errors and standard errors clustered at country level.

Column (1) reports the positive and highly statistically significant unconditional effect of malaria on the log number of ethnic group. The effect is quantitatively sizable. An increase of one (cross-cells) standard deviation of malaria stability increases the log number of ethnic groups by a 0.36 standard deviation, increasing the average number of ethnic group by 0.19 log points.

Column (2) adds country fixed effects to account for the potential role of the process of country formation and country borders. The inclusion of country fixed effects improves the fit of the regression, increasing the R^2 to 0.30, but leaves the magnitude (and the precision of the estimate) of the role of malaria essentially unaffected. Columns (3) and (4) extend the specification to the geographic

[34]The inclusion of soil suitability (mean and standard deviation), elevation and basic geographic covariates follow Michalopoulos (2012). The inclusion of distances allows accounting for possible confounders related to locations. Information on caloric suitability is included, following Galor and Ozak (2016), to account for the Malthusian nature of the hypothesis. Distance from East Africa is included following the arguments on genetic diversity by Ashraf and Galor (2013). The description of the variables of interest, of the main explanatory variables, and of the covariates are reported in the Appendix in Tables E1 and E2, respectively.

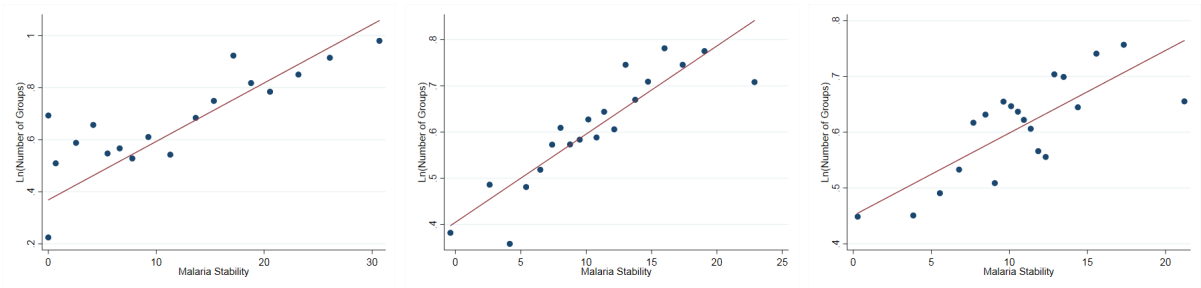
FIGURE 1: NUMBER OF GROUPS AND MALARIA STABILITY



(a) Number of Ethnic Groups (b) Malaria Stability

Notes: The figure on the left plots the total number of groups (GREG) across our 1x1 degree cells. The figure on the right maps the average level of Malaria Stability across the same 1x1 degrees cells.

FIGURE 2: BIN SCATTERS - MALARIA STABILITY AND GREG (LOG) NUMBER OF GROUPS



(a) Unconditional (b) Controls without Country FE (c) Fully Controlled

Notes: The graphs plot the values of Ln Number of Groups and Malaria Stability along the two axes through binned scatterplots (describing the average x-value for each y-value).

and location covariates, respectively. Column (4) includes all the controls that have been proposed and investigated in the literature, as discussed in the previous sections, and extends the specification to the further covariates discussed above. The effect of malaria exposure is generally little affected by the inclusion of other (relevant) covariates. Long-term exposure to malaria accounts for a sizable variation in the size of ethnic groups. The R^2 of the specification of Column (1) is around 0.17 which is about 50 % of the variability explained by the most extensive specification.

Table 1: Long-Term Exposure to Malaria and Historical Ethnicities

	Ln (Number of Groups - GREG)				
	(1)	(2)	(3)	(4)	(5)
Malaria Stability	0.021***	0.020***	0.017***	0.015***	0.015***
Cluster s.e. (Country)	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)
Conley s.e. (500km)	(0.002)	(0.003)	(0.004)	(0.004)	(0.004)
Beta Coefficient	[0.358]	[0.345]	[0.290]	[0.265]	[0.257]
Geographic Controls:					
Ln(Cell Area)	Yes	Yes	Yes	Yes	Yes
Soil Suitability and Elevation (Mean and Std.)	No	No	Yes	Yes	Yes
Avg. Temperature and Precipitation	No	No	Yes	Yes	Yes
Ruggedness and Caloric Suitability Pre 1500	No	No	Yes	Yes	Yes
Location Controls:					
Distances (equator, coast, river, border, capital)	No	No	No	Yes	Yes
Number of Countries and Within Country	No	No	No	Yes	Yes
TseTse suit. & Pr. Genetic Distance	No	No	No	No	Yes
Country FE	No	Yes	Yes	Yes	Yes
Observations	1,976	1,976	1,976	1,976	1,976
R-squared	.17	.299	.351	.376	.379

Notes: The Table reports the OLS specification estimates associating the log number of ethnic groups with the level of Malaria Stability. In all specifications, the dependent variable is the natural logarithm of the number of ethnic groups in the cell; Malaria Stability measures the average malaria suitability in the cell; see text for details. The geographic variables include the natural logarithm of the cell land area, soil suitability and elevation (mean and standard deviation), average temperature and precipitation, terrain ruggedness and caloric suitability before 1500. The distance variables include distance from the equator, from the coast, from the river, from the country border and from the country capital. Number of countries is the number of countries whose land fall within the cell. The within-country is a dummy variable taking value 1 if the cell is fully contained in a country. TseTse suitability measures the geographic suitability for Trypanosomiasis transmission. Predicted Genetic distance is computed as the log of migratory distance from East Africa. Variable description, data sources and summary statistics are reported in Tables E1, E2, S1, and S4, respectively. Beta coefficient in square brackets. Robust standard errors clustered by country and Conley standard errors (500 km cutoff) are reported in round brackets. ***, **, and * indicate significance at the 1-, 5-, and 10-% levels, respectively.

Robustness of Baseline. The baseline results are robust to several checks that are reported in the Appendix. These include the measure of ethnic diversity in terms of territorial fractionalization (Table A1) and sensitivity checks in different samples that restrict attention within malarial areas and within cells with multiple groups (Tables A2 and Table A3), as well as replicating the analysis for the whole Old World (Table A4). Finally, the results are robust to controlling for proxies for current development in terms of lights at night and population density in 2000 (Table A5).

The robustness of the results, the sizable effect of malaria exposure, and the high stability of the point estimates are little affected by conditioning on increasingly large sets of covariates, which builds and extends the ones proposed in the literature. Moreover, standard tests based on observables suggest that the effect of malaria exposure is unlikely to be driven by unobserved (cell specific) characteristics. Further strategies are implemented below to further evaluate the specific predictions and potential threats to empirical identification.

3.2.2 Measures of Long Malaria Exposure

Measuring long-term malaria exposure at local level for the whole of Africa is not straightforward. The malaria stability index used as baseline has the advantage of providing a fine-grained disaggre-

gate measure of predicted malaria exposure. The measure is built exploiting also information on the distribution of the type of mosquitoes across locations which can be affected by the long-term interaction with humans.[35] The presence of humans, particularly in terms of population density and potentially their productive activities, might indirectly influence the construction of the baseline index. While no available argument suggests that the number of spatially clustered groups, rather than population density, should impact the distribution of mosquitoes types, we explore the sensitivity of the results to different measures of malaria exposure.

Alternative Measures of Predicted Malaria. As a first strategy to assess the potential role of mosquitoes types, we employ the alternative index of Plasmodium Falciparum Suitability proposed by Gething et al. (2011) and previously exploited by Depetris-Chauvin and Weil (2018). The index is constructed using temperature-based constraints to malaria transmission rather than actual presence of different mosquitoes' types. As a second strategy, we reconstruct the Malaria Stability index excluding information on the mosquitoes' specific characteristics. Drawing on the epidemiological literature, we predict mosquitoes-specific human biting preferences as a function of climatological conditions only.[36] The results obtained with both intention-to-treat regressions and 2SLS strategies deliver very similar and consistent results reassuring that the baseline results are not driven by the effect of the number of groups on the presence of different types of mosquitoes (see Table A6).

Measures of Historical Malaria Incidence in the Population. Measures of malaria exposure discussed above are informative on predicted, rather than actual, past malaria exposure. This feature is appealing in the context of endogeneity concerns but they are, by construction, not directly informative on the actual selective pressure of ancestral exposure to the pathogen in terms of actual spread of the pathogen in the population.[37] To explore the role of actual exposure to the pathogen, we use the (limited) available information on endemicity of the pathogen in the African population measured around 1900, from Lysenko and Semashko (1968). As alternative measure of the historical ancestral exposure to malaria we also look at the distribution of the frequency of genetic immunities to malaria in terms of the share of individuals with Duffy negative phenotype, from Howes et al. (2011). Conceptually, information on the spatial distribution of these traits is the best available

[35] In fact, biting preference of the locally dominant version of the Anopheles presumably developed over hundreds or thousands of years potentially in co-evolution with humans and other local mammals. See for instance Neafsey et al. (2015) and Constantini et. al. (1999).

[36] These alternative indexes, that are correlated with each other at 92%, display a geographical distribution that is coarser of the baseline malaria measure and is conceptually noisier measures of actual long-term malaria exposure. For the purpose of this specific sensitivity check they are interesting as they have the advantage of not relying on information on the actual types of mosquitoes in each location. Technical details on the construction of these indexes are available in Appendix Section A2.1.

[37] This concern is attenuated in the case of the baseline malaria stability index, that accounts for actual mosquitoes types and has been argued to deliver a good measure of actual long-term exposure to the pathogen.

proxy of past exposure to malaria for the whole of Africa.[38] A limitation of these population-based measures is that they display a limited spatial variability which limits their informative content. Replicating the analysis nonetheless confirm the baseline patterns delivering further evidence on the role of the high selective pressure imposed by malaria on the ancestral African population (see Table A7).[39]

3.2.3 Multi-Host Vector Transmitted Pathogens: Placebo Diseases

The African populations are deeply affected by other serious diseases that are transmitted by flying vectors with important implications also for long-term development, see e.g. Alsan (2015) and Lowes and Montero (2018). Compared to these diseases, that include trypanosomiasis, dengue and yellow fever, malaria has distinctive features that are important for the purposes of our investigation. First, unlike other pathogens transmitted by flying vectors with also non-human hosts reservoirs, plasmodium falciparum parasites are hosted almost exclusively in humans.[40] A relevant implication is that geographic separation of humans across non interacting subgroups, which reduces the host group of malaria (see Section 2.2), did not represent an effective strategy to limit incidence of multi-host diseases like trypanosomiasis, dengue and yellow fever unless it is effectively applied also to the domestic and wild animal reservoir hosts.[41] Second, for these diseases the evidence on the role of genetic immunities is scant and no arguments have to our knowledge been proposed on their role for sexual inbreeding. In contrast, malaria is considered the strongest selector of the human genome as discussed above. Finally, among tropical diseases, malaria is by far the one that imposed the highest

[38]This is the case for several reasons. First, the Duffy negative genotype offers effective protection against plasmodium vivax (to the point of leading to almost complete elimination of this variant in Africa) and limits the risk of severe infections from falciparum. Second, its mild health consequences favored a sharp response to ancestral malaria and imply a large persistence of these traits still today. Finally, since it is not incompatible with other malaria protective genetic mutations, it offers a comparable measure of past exposure across all locations of Africa. This is different from other mutations, for example, those of the HBB gene, for which different populations have developed different evolutionary responses at the local level, see e.g. Kwiatkowski (2005). For instance, the HbS variant (sickle cell) is more concentrated in central Africa while the HbC variant (haemoglobin C) is more concentrated in West Africa. Looking at the spatial spread of each HB variant alone is therefore not necessarily informative on the overall past exposure to the pathogen across all locations in Africa.

[39]A map of the measures aggregated at 1×1 degree cell level is reported on the right panel of Appendix Figure D3. Both measures display little spatial variability since the data are highly spatially interpolated. This also makes inference more noisy and limits the residual variability across locations particularly after conditioning on the geographic covariates that are, on the contrary, measured at a finer spatial resolution.

[40]Plasmodium falciparum protozoa have been sometimes detected in other apes like gorillas (that host different variant of plasmodium) but humans are considered the main host. In contrast, a wide set of mammals, including domestic and wild animals, have been documented as important reservoir hosts for trypanosomiasis (transmitted by *TseTse flies*) including the variant that most intensively affect humans. Similarly, yellow and dengue viruses (mostly transmitted by *Aedes mosquitoes*), exploits wild animals as reservoirs. See e.g. the Global Infectious Disease and Epidemiology Network (GIDEON) Database or the WHO for summaries of the specific features of vector-borne diseases affecting Africans.

[41]In fact, the presence of multi-hosts including wild and sylvatic species is considered a relevant impediment also for the ability to control the spread of these diseases still today and limits the future prospects of their eradication. See, e.g., the WHO website.

health and mortality burden on pre-modern Africans.

According to the framework presented in Section 2.2, the exposure to these multi-host vector-borne diseases should not be expected to be a prime driver of the number of clustered groups in each location and of the emergence of cultures enforcing sexual endogamy, since no sizable benefits should be expected from these costly adaptation strategies. Comparing the exposure to malaria to these other vector-borne diseases, therefore, represents a valuable empirical test that allows to further explore the channel and offers a further strategy to assess identification concerns by mean of an interesting placebo. The results confirm that the role of malaria exposure is unaffected by accounting for the predicted exposure to trypanosomiasis, dengue and yellow fever which, in turn, do not systematically and significantly affect the number of ethnic groups, see Tables A8, A9 and A10.

3.2.4 Cell Size and Drawing of Ethnicity Borders

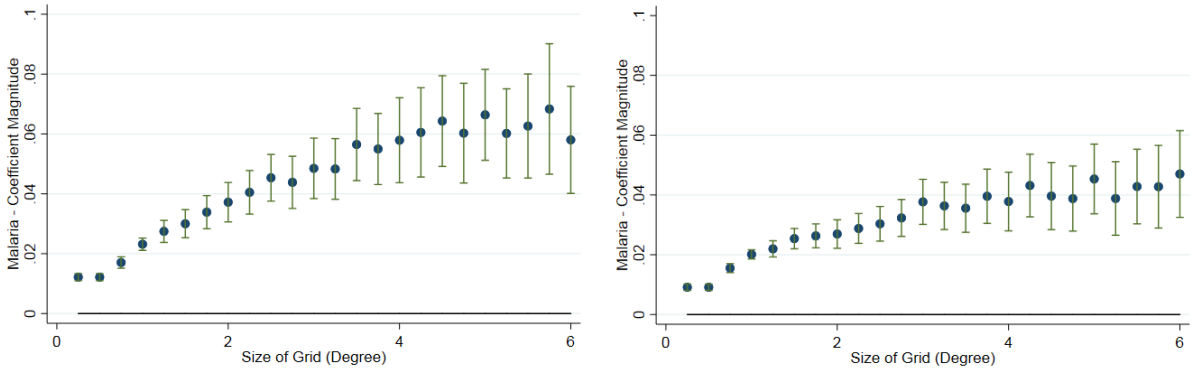
Cells are drawn randomly. The distribution of the number of groups in the sample depends, however, on how cells randomly cut the distribution of ethnic groups. This feature raises two main questions. First, a larger cell size mechanically (weakly) increases the number of groups in each location.[42] Second, it is not clear to which degree the estimated number of groups in a cell is affected by possible errors in the original drawing of the borders of each group. A further issue is about the conceptual difficulty, that is common to all disaggregated analysis with grid cells, to establish a theoretically congruent level of aggregation for the problem at hand.

We explore these issues by systematically replicating the analysis varying grid cell sizes. We reconstruct the database for alternative units of observation: from 0.25 degrees (about 28km x 28km) to 6 degrees (about 666km x 666km) (in steps of 0.25 degrees). The results, which are comparable to Column 1 of Table 1 and reported in Figure 3 (a), confirm the baseline patterns and suggest that the baseline effects obtained with 1-degree cells are, actually, conservative. The point estimate of the effect of malaria increases in absolute magnitude with cell size and tends to increase up to cells of 2.5-3 degrees, and then stabilizes.

To account for potential (non-random) errors in the drawing of borders of historical ethnicities, the analysis is also replicated, by applying Thiessen-polygons transformations of the original borders of each ethnic group and by re-computing the number of groups in each cell accordingly, using full set of different cell sizes. The results, depicted in Figure 3 (b), deliver practically identical results for the baseline grid cells size and slightly lower point estimates for larger cell sizes.

[42]This is related to the modifiable areal unit problem, MAUP.

FIGURE 3: GRID-CELLS SIZE AND THIESSEN POLYGON TRANSFORMATIONS



(a) GREG Polygons - Africa

(b) Thiessen Polygons GREG - Africa

Notes: The graph plots coefficient estimates of the regression of the effect of malaria on the log number of groups (GREG) by replicating the baseline regression using the different cells of different size (x-axis). In panel (a) the number of groups is computed using the borders as mapped by the GREG database. In panel (b), we reconstruct group borders of the GREG dataset through a Thiessen polygon transformation of the original data.

3.2.5 Characteristics and Population Density of Pre-colonial Ethnic Groups

The hypothesis under investigation is based on the view of a long-term adaptation of mankind to the local ecology. The conceptual framework in Section 2 predicts that malaria should be expected to increase the number of spatially clustered and sexually endogamic groups. The theory does not deliver specific insights and predictions on the role of malaria for the organization of these ethnic groups (besides a limitation of contacts across groups which could reduce trade and incentives for productive specialization). Still, it is interesting to check the robustness of the baseline patterns explicitly accounting for, potentially omitted, pre-colonial characteristics of these ethnic groups and also exploring the potential role of malaria in shaping the organization of pre-colonial groups.

Pre-colonial Characteristics. We look at pre-colonial ethnographic characteristics from the Ethnographic Atlas (1967), an anthropological database contains detailed information for 534 ethnic groups in Africa, and create a set of variables measuring the average pre-colonial subsistence, settlement, institutional and cultural patterns in the cell.[43] The results show that the effect of malaria on the number of ethnic groups is not driven by specific pre-colonial characteristics (Table A11). When considered as dependent variables, the analysis delivers no evidence of systematic effects of malaria

[43]Subsistence variables include: Gathering, Hunting, Fishing, Animal Husbandry, Agricultural Dependence, Agricultural Types, and Milking. Settlement patterns variables include: Settlement complexity, Political Complexity at local level. Institutional and cultural characteristics include: Polygyny, Clan Communities, Slavery, Property right, and Political Complexity beyond the local level. See Section A3.5.1 in the Appendix for a description of data construction and presentation of the results.

on the pre-colonial organization of groups, besides an increased likelihood of relying on fishing as a main subsistence activity (Table A12).

Pre-Colonial Population Density. The conceptual role of malaria for pre-colonial population density requires a specific discussion in view of the predictions of the overall role of malaria in a Malthusian stationary state (see Lemma 2 in Section 2.3). For given patterns of geographic clustering and genetic immunities, a higher level of malaria worsen health and reduces the level of Malthusian population density. The framework predicts a countervailing effect coming from spatial clustering and endogamy that reduce disease incidence and, from Proposition 1, are predicted to increase with malaria. In other words, human groups that were intensively exposed to malaria are predicted to experience a process of cultural and genetic adaptation that allows them to sustain higher levels of population density for any level of exposure to the pathogen. The overall predicted effect of malaria on the level of population density in the Malthusian stationary state is therefore ambiguous. Concerning the other expected drivers of pre-colonial population density, the Malthusian framework confirms the predictions that all factors that increase total factor productivity and human health should increase population density.

The analysis delivers several interesting insights. First, the effect of malaria on the number of ethnic groups is essentially unchanged by controlling for the level of pre-colonial population density (Table A11). Second, the results show no consistent effects of malaria on pre-colonial population density (see Figure D8 and Table A13) and, confirming the results by Depetris-Chauvin and Weil (2018), generally not statistically significant when including geographic covariates.[44] Finally, the (unreported) main drivers of pre-colonial population density are level of precipitations and crop caloric suitability (with a positive effect as in Ashraf and Galor, 2011) and TseTse suitability (with a negative effect as in Alsan, 2015).

3.3 Pre-Colonial Ethnicities

We next focus attention on the role of ancestral exposure to malaria for the predicted spatial distribution of pre-colonial, rather than historical, ethnic groups.

3.3.1 The Role of Malaria in Africa *vs* the Americas (Placebo)

Malaria is a global burden today affecting populations in Africa and the Americas. The epidemiological history of the disease is very different on the two sides of the Atlantic ocean, however. As

[44]Specifically, the effect tends to be positive for the baseline measure of malaria exposure (with a strongly declining point estimate when including geographic covariates) but it is negative (and generally insignificant) when malaria is measured using alternative indexes that do not rely on information on the type of mosquito vectors and conditioning on covariates.

mentioned in Section 1, *plasmodium falciparum* has plagued Africans for thousands of years but was "exported" to the Americas only after the European colonization of the continent, presumably around the seventeen century, in the context of the so-called Columbus Exchange.[45]

The lack of exposure to the disease before colonization implies that the postulated mechanism linking malaria to the process of emergence of pre-colonial groups should not be at work in the New World. While African scholars have considered the burden of malaria as something at the root of African civilizations, as discussed in Section 2.5, we are not aware of accounts or narratives on the role of malaria in shaping pre-colonial civilizations in the Americas. This (lack of) historical narratives and scholarly arguments suggest that, if the hypothesis under test has empirical validity, we should expect no systematic effects of malaria on the patterns of geographic distribution of pre-colonial ethnicities in the New World.[46]

The geographic distribution of pre-colonial homelands is available for both sides of the Atlantic ocean. Information on the spatial distribution of ethnic groups at the eve of colonization is retrieved from the works of Murdock (1951, 1957, 1959) that apply the same methodology of data collection and treatment for both Africa and the Americas. The Murdock's maps provide the best available representation of the distribution of ethnic groups before colonization.[47] Figure 4 depicts the spatial distribution of the pre-colonial ethnic homelands in Africa and the Americas.

Exploring the role of malaria for the spatial distribution of pre-colonial ethnicities in Africa and the Americas allows us to move one step forward in testing the specific predictions and in the identification of the role of malaria for the emergence of ethnicities as resulting from a process of long-term adaptation to the local ecology in Africa by using the Americas as a placebo.

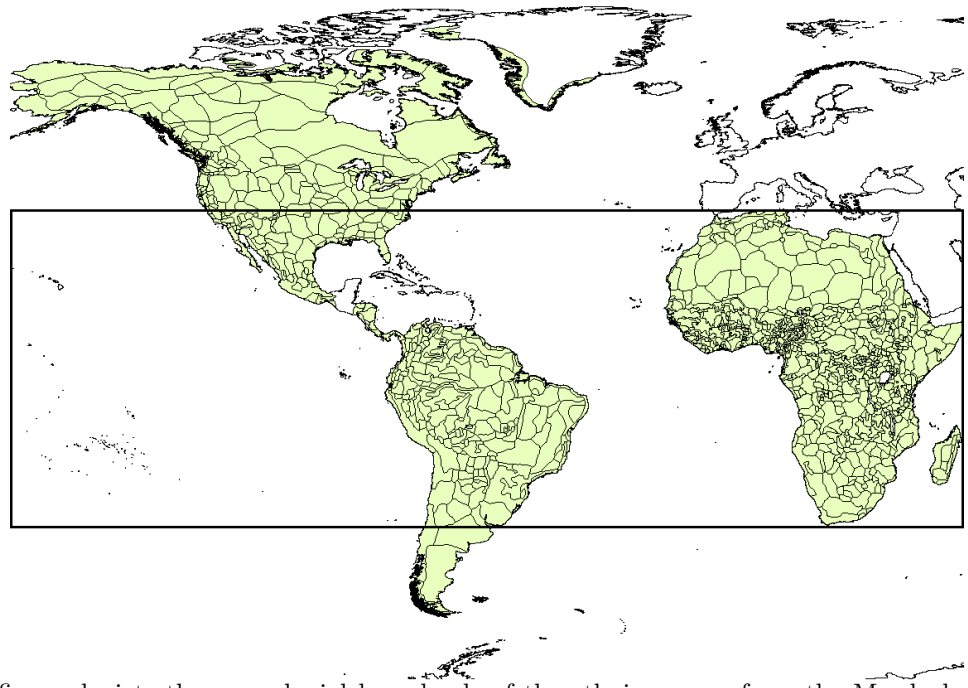
Table 2 replicates the baseline specifications of Table 1 for Africa and the Americas. The results in Columns (1) to (3) document a positive and highly statistically significant effect of malaria on

[45]Milder variants of plasmodium, the *plasmodium simium*, that primarily affected monkeys have been detected in the Americas prior to European colonization. The Columbus Exchange brought to the New World the plasmodium vivax already in the early phases of European colonization and, most importantly, *plasmodium falciparum*, the more lethal variant, responsible for the strongest selective pressure from malaria in Africa.

[46]This should be expected concerning the distribution of pre-colonial ethnicities and not necessarily the distribution of historical or contemporaneous ethnic groups that in the Americas (differently from Africa) has been deeply reshaped by the process of colonization and intense settlements of colonizers. Malaria might have affected the differential survival and the spatial distribution and relocation of the surviving native populations after colonization, the intense settlement patterns of Europeans and the relocation in specific areas of the Americas of malaria-resistant Africans in the context of the colonial slave trade (on this see Esposito, 2018). This implies, in particular, that information from the GREG database, used in the analysis so far and that reports the historical locations of the population belonging to the different ethnicities in the first half of the twentieth century (up to 1960), cannot be used to explore the validity of our hypothesis for the Americas.

[47]Importantly for purposes of this paper, they do not include information on colonial groups (and European descendants) and ethnic groups that were displaced after colonization, such as, for instance, the descendants of African slaves in the Americas. In this respect, the information allows us to test more directly the hypothesis on the long-term (pre-colonial) effect of exposure to malaria. The data also implicitly allow us to isolate the effect of malaria on African ethnicities from the settlement patterns of colonizers and the (possibly related) most recent migration of members of different African ethnicities during the process of colonization.

Figure 4: Murdock's Maps - Africa and Americas



Note: The figure depicts the pre-colonial homelands of the ethnic groups from the Murdock (1951, 1957, 1959) maps of Africa and the Americas. For comparability, the baseline estimation restricts attention to the locations in America that lie within the latitudes of the African continent (depicted by the rectangle).

the number of pre-colonial ethnic groups in Africa. The effect is quantitatively important and, in fact, comparable in size to the effect on historical diversity documented above. An increase of one cross-cells standard deviation of the Malaria Stability index increases the number of ethnic groups by a third of a standard deviation, adding 0.20 log points to the average number of ethnic groups. This implies an increase of the average number of ethnic groups by about 0.5 groups, around one-fifth of the sample mean. The pattern is confirmed when extending the specification to account for the geographic and location covariates. As in Table 1 malaria is a main predictor of the number of groups with the R^2 of column (1) being about 50 percent of one of the most extensive specification in column (3).

The results in Columns (4) to (6) show no consistent and statistically significant pattern of exposure to malaria on the number of pre-colonial ethnic groups in the Americas. Already in the baseline specification in Column (4) the coefficient is insignificant and malaria tends to explain a negligible share of the variation in the number of ethnic groups in the data. Interestingly, and reassuringly, the (unreported) effect of the other main determinants of ethnic diversity, in particular, the standard deviations of elevation which proxies for geographic isolation and the suitability for agriculture, which proxy for the incentives for productive specialization, are similar for Africa and the Americas. The increase in the R^2 due to the inclusion of the covariates other than malaria is

Table 2: Pre-Colonial Ethnic Diversity - Africa *vs* Americas (Placebo)

	Ln (Number of Groups - Murdock's Data)					
	Africa			Americas		
	(1)	(2)	(3)	(4)	(5)	(6)
Malaria Stability	0.021***	0.014***	0.013***	0.004	0.003	0.001
Cluster s.e. (Country)	(0.001)	(0.002)	(0.002)	(0.005)	(0.009)	(0.010)
Conley s.e. (500km)	(0.003)	(0.003)	(0.003)	(0.008)	(0.013)	(0.013)
Beta Coefficient	[0.356]	[0.235]	[0.216]	[0.018]	[0.016]	[0.004]
Geographic Controls:						
Ln(Cell Area)	Yes	Yes	Yes	Yes	Yes	Yes
Soil Suitability and Elevation (Mean and Std.)	No	Yes	Yes	No	Yes	Yes
Avg. Temperature and Precipitation	No	Yes	Yes	No	Yes	Yes
Ruggedness and Caloric Suitability Pre 1500	No	Yes	Yes	No	Yes	Yes
Location Controls:						
Distances (equator, coast, river)	No	No	Yes	No	No	Yes
Observations	1,973	1,973	1,973	1,503	1,503	1,503
R-squared	.154	.286	.292	.0636	.135	.137

Notes: The Table reports the OLS specification estimates associating the number of pre-colonial ethnic groups with the level of Malaria Stability in Africa and Americas. The dependent variable is the natural logarithm of the number of pre-colonial groups (Murdock 1951 and 1959) in the cell. All dependent variables are constructed using the Murdock maps for Africa and Americas (see Appendix A2.2 for further details). Malaria Stability is the average level of the malaria suitability in the cell. See caption of Table 1 and the text for details and for the list of covariates. The geographic variables include the natural logarithm of the cell land area, soil suitability and elevation (mean and standard deviation), average temperature and precipitation, terrain ruggedness and caloric suitability before 1500. The distance variables include distance from the equator, from the coast, from the river, from the country border and from the country capital. Variable description, data sources and summary statistics are reported in Tables E1, E2, E3, and S1, respectively. Beta coefficient in square bracket. Robust standard errors clustered by country and Conley standard errors (500 km cutoff). ***, **, and * indicate significance at the 1-, 5-, and 10-% levels, respectively.

also similar for both samples. This should be expected since, differently from the long-term role of malaria in behavioral isolation, geographic isolation and productive specialization should be universal mechanisms at work also in the Americas.[48]

3.3.2 Robustness

The patterns are robust to several checks reported in the Appendix. These include, in particular replicating the analysis for:

Alternative Malaria Measures that do not rely on the mosquitoes types. The results consistently emerge also when using the alternative measures of predicted malaria exposure (Table B5).

Cell Sizes and Thiessen Polygons. We have replicated the analysis for the full set of alternative grid cell sizes. Figures D9 and D10 report the results for Africa and the Americas, respectively, both for the original homelands and the ethnic diversity recomputed using Thiessen polygon transformations that correct for errors in the drawing of homelands' borders. The results confirm the patterns

[48]For comparability across continents, Table 2 does not condition on the information on TseTse suitability which is not publicly available for the American continent. In line with the previous discussion on the role of other diseases, however, the results for Africa are unaffected when including TseTse suitability (see Table B1).

for Africa (suggesting again that the effect of malaria in grid cells of one degree represents a conservative estimate), while for the Americas no systematic patterns emerge. For the Americas, the sign of the coefficient also varies depending on the size of the cell and the effects (positive or negative) are never statistically significant.

Multi-host vector: Placebo Diseases. The results are confirmed when accounting for the predicted distribution of Trypanosomiasis, Dengue and Yellow fever replicating the placebo exercises also for pre-colonial African ethnicities, see Tables B2, B3 and B4.

3.4 The Legacy of Malaria for Ethnic Diversity Today

The findings in Sections 3.2 and 3.3 provide the first existing attempt to systematically investigate the drivers of the geographic location of historical and pre-colonial ethnic groups. The analysis explores only some of the predictions from the conceptual framework of Section 2 and leaves open the question about the persistence of the effect of malaria on the territorial distribution of ethnic groups today. Before turning, in Section 4, to study the predicted channels of persistence of ethnicities in terms of endogamic cultures, we briefly discuss the role of malaria for ethnic diversity today in term of spatial clustering, ethnic identities and the distribution of groups in Africa.

Spatial Clustering and Ethnic Admixing (Village Level Data). A specific prediction of the framework in Section 2.2 rests on the epidemiological evidence that, for malaria, reducing the size of human host population requires separation of mankind into geographic clustered, or stand alone, groups. As discussed in Section 2.5 the effect of malaria on the (cultural and genetic) adaptation to the local ecology should be expected to materialize also in reduced incentives for the relocation of groups and people across locations. This prediction is in line with arguments of African scholars and historical narratives discussed above. The maps on the geographic distribution of historical and pre-colonial homelands allow us to the test the prediction on the number of groups but not to explore the prediction of limited admixing, since no data on the actual past distribution of the population at the local level is available. To explore the prediction of low ethnic admixing, we look at the actual distribution of individuals in Africa today. Using DHS data, we compute the number of ethnic groups (and the level of ethnic fractionalization) at the village level. The results, that on top of geographic covariates and proxies for the level of development (in terms of population density and lights at night) and condition on the number of ethnic groups in the region, document that villages with higher malaria display significantly lower degree of ethnic admixing, see Tables C1 and C2).[49]

[49]We exploit information on around 14,000 villages within countries (Benin, Burkina Faso, Cameroon, Central African Republic, Ethiopia, Gabon, Ghana, Guinea, Ivory Coast, Kenya, Liberia, Malawi, Mali, Mozambique, Namibia, Niger, Nigeria, Senegal, Sierra Leone, Togo, Uganda, and Zambia).

Identification with own Ethnic Group (Village Level Data). Historical narratives and scholarly arguments uncover a peculiarity of the African ethnic phenomenon (sometimes called geo-ethnicities) in the territorial element of ethnic identities. The conceptual framework provides a rationale for this peculiarity since ethnic identities in malaria areas should be shaped in the context of the emergence of (and are instrumental to) the enforcement of cultures of geographic and behavioral isolation in each location. The prediction on the (persistent) role of malaria for ethnic identity is explored using Afrobarometer data, which allows us to measure the intensity of ethnic identification with own ethnic group in different locations across Africa. Besides including all geographic and location covariates, we also control for individual characteristics and macro characteristics on the distribution of ethnic groups in the region. While interpreting self-reported data on identity is not straightforward, the results document a strong and statistically significant effect of malaria on ethnic identification with own ethnic group (Table C3).[50]

Ethnolinguistic Diversity Today (Grid Cell Level). As discussed in Section 2, the predicted effect of malaria relates to the emergence and persistence of spatially clustered ethnic groups rather than to a direct impact on the distribution of the population today. To explore the legacy of malaria for the distribution of the African population today we replicate the baseline analysis of Table 1 using data from the World Language Mapping System, which portrays the distribution of ethnolinguistic groups today. The results show that malaria exposure increases the number of ethnolinguistic groups also today but further documents that this effect vanishes once controlling for historical and pre-colonial diversity, see Table C4. Together with the results in Tables 1 and 2, the findings suggest that the legacy of malaria on the distribution of ethnic groups today is related to its imprint on borders of pre-colonial and historical ethnic homelands.

4 Channel of Persistence of Ethnicities: Endogamic Cultures

According to the conceptual framework presented in Section 2.4, the emergence of cultures of behavioral isolation involving, in particular, endogamic sexual inbreeding allowed ancestral populations to limit disease incidence in the face of strong pressure from malaria. As mentioned in Section 2.5, the persistent of endogamic ethnic marriages is considered crucial for the survival of ethnic groups.

[50]Throughout specifications, the coefficient remains stable, precisely estimated and sizable in terms of magnitude. Individual controls include living conditions, education, religion, occupation, rural or urban residence. The results consistently emerge also accounting for socio-economic development in terms of population and night lights and when controlling for the presence of different ethnic groups in terms of the respondents' group size (and share) in the region, the total number of ethnic groups, and the index of ethnic fractionalization in the region.

Historical narratives on sexual inbreeding and the high frequency of group-specific genetic immunities in malarial areas provide background evidence on the potential role of malaria for this predicted channel of persistence.

We are not aware, however, of any attempt to measure and systematically explore the determinants of ethnolinguistic endogamy in Africa today. Accordingly, no investigation of the potential role ancestral malaria in shaping endogamic cultures and the differential patterns of persistence of ethnicities is available. In this Section, we study the determinants of ethnolinguistic endogamy today using DHS individual survey data for Africa. The first step is the conceptualization and measurement of individual ethnic endogamy. The identification strategy of the role of ancestral malaria, acting through the predicted channel of persistent endogamic cultures, requires isolating the role of ancestral characteristics from the individual and location-specific drivers of ethnolinguistic marriages. By looking at respondents not residing in their ancestral ethnic homeland, we exploit within-village variation to isolate the role of the ancestral origins of individuals residing in the same location and therefore facing the same local environment (in terms of e.g. current malaria exposure, geography, social structure, and economic development).

4.1 Data and Empirical Strategy

4.1.1 Measurement and Data

Ethnic Endogamy. Measures of ethnic endogamy are built using information on the ethnicities of wives and husbands from DHS data. Conceptually, ethnolinguistic endogamy refers to the fact that the two spouses belong to the same group. The measurement of endogamy along ethnolinguistic lines is, nonetheless, empirically not straightforward. Our aim is to measure endogamic marriages in a way that: (i) is comparable throughout all DHS waves and countries, and that (ii) maps endogamy at various level of the Ethnologue language tree. To this end, we match the ethnicities reported in the DHS survey to their respective Ethnologue language trees by employing several sources: Ethnologue (Lewis, 2009), Murdock (1959), and the Joshua Project (and some minor others).[51]

We propose two conceptually different types of measures of ethnolinguistic endogamy. Endogamy indicators, that take value 1 if husband and wife belong to the same ethnic group, and 0 otherwise. These measures code endogamy depending on whether the spouses belong to the same group, at the different levels of the language tree ranging from macro families to local dialects. A relevant caveat of the endogamy indicator is that there is no metric on distances between spouses and the dichotomous coding of endogamy may be overly sensitive to the chosen level of the language tree. In other words, couples who are coded as exogamous at the more disaggregated branches of the

[51] A more detailed discussion of the conceptual measurement of endogamy across different trees is reported in the Appendix Section A6.1.

Ethnologue trees can be defined as endogamic at the upper branches.[52] To address this issue we also build a measure that exploits information on the ethnolinguistic *distance* between spouses along the Ethnologue tree, following Desmet et al. (2012). The index, that ranges from 0 to 1, increases with the distance between the two ethnic groups along the linguistic tree. The resulting index provides a measure of "exogamic" distances, rather than the existence of endogamy at any given level of the language tree.

Current Location and Ethnic Homelands: Movers. Empirical identification eventually rests on the sample of respondents that do not reside in their ancestral ethnic homeland, along the lines of Nunn and Wantchekon (2011) and Michalopoulos, Putterman, and Weil (2018). The ancestral ethnic homeland of each respondent is tracked by matching the DHS ethnicity of responders to Murdock's ethnic homelands (1959) and Murdock's Ethnographic Atlas (1967).[53] Using the geo-location of DHS respondents, we finally identify the respondents living outside their ancestral ethnic homeland.[54]

Covariates. The analysis conditions on ancestral characteristics (at the level of ancestral ethnic homelands), individual characteristics as well as information on the size of the group of the responder at the region level (to proxy for the size of local ethnic marriage market).[55] The level of malaria in the location is computed averaging in a 10 km radius around respondent's location. Ancestral characteristics are the average in the polygon of the ethnic homeland of the ethnic group of the respondent.

4.2 Empirical Strategy

The drivers of endogamic marriages are explored estimating specifications like,

[52]More specifically, couples not belonging to the same group at a given level are coded as non-endogamous irrespective of whether they would be endogamous using a slightly broader definition of groups along the tree (e.g. if they speak different dialects or variants of the language of a common ancestral ethnic group) or whether they would be coded as endogamous only at very low levels of the tree (e.g., they speak completely different languages and have completely different ancestral homelands). Figure D11 in the Appendix illustrates the example of an Ethnologue tree for the Niger-Congo linguistic family.

[53]The majority of observations, 58% is matched directly. When the name of the ethnicity differs across sources we use the existing alternative names of ethnic groups from the Ethnologue (Lewis, 2009) as a secondary source to establish a match (26% of the observations). For the remaining cases, we complete the match using information from (i) Murdock (1959) (7% of the sample) and following Nunn and Wantchekon (2011) on the Afrobarometer ethnic groups (something less than 5%). The remaining groups are matched using the Joshua Project (one ethnic group) and Wikipedia (two ethnicities).

[54]Given that DHS coordinates are perturbed by 5 to 10 km to ensure confidentiality, we restrict the definition of movers only to those female respondents living at least 10 km away from their historical homeland. For robustness, we check the sensitivity of the results to various definitions of movers and distances (see below). Figure D13 in the Appendix depicts the distribution of movers based on the distance from their ancestral ethnic homeland.

[55]For consistency, the analysis controls for the geographic covariates included in the previous part of the analysis at the ancestral home levels. We also control for individual characteristics in terms of a urban dummy, years of education, age, religion fixed effects, and dummies of relative wealth (poorest, poorer, middle, richer, richest) for each respondent. See the Appendix for further details.

$$Endog_{i,z,g,t,c} = \beta_0 + \beta_1 MalLoc_{z,c} + \beta_2 AncMal_g + \beta_4 \mathbf{Y}_g + \beta_5 \mathbf{I}_{i,z,g,t,c} + \mu_{z,t,c} + \epsilon_{i,z,g,t,c}$$

where $Endog_{i,z,g,c}$ is a measure of endogamy of the female respondent i, belonging to ethnic group g, to whom the survey was administered in year t, living in survey village z in country c. Endogamy is measured using either the indicator variables (at various levels of the language tree) or measures of exogamic distances, as discussed above. The explanatory variables of main interest are (depending on the specification) : (i) malaria in the location of the respondent $MalLoc_{z,c}$, (ii) ancestral malaria in the homeland of the respondent's ethnic group $MalAnc_g$.

The vector $\mu_{\mathbf{z,c,t}}$ indicates the inclusion of different types of fixed effects at the level of location z, country c and/or DHS wave t. All specifications include wave and country fixed effects, so that $\mu_{z,c,t} = \mu_t + \mu_c$. Some specifications include instead location fixed effect, so that $\mu_{z,c,t} = \mu_t + \mu_z$.[56] The inclusion of location fixed effects automatically drops malaria location $MalLoc_{z,c}$, which cannot be estimated as all location-specific characteristics will be absorbed by the fixed effects. Econometric identification effectively exploits variability across individuals with different ancestral homelands living in the same location. The vector Y includes the covariates related to the respondent's ancestral homeland. Finally, the analysis also conditions on proxies for the potential ethnic marriage markets, in terms, e.g. of the size of the ethnic group of the respondent in the region, and individual characteristics, $I_{i,z,g,c}$.

4.3 Results

4.3.1 Baseline

The estimation sample involves information on 19,414 couples over 3,514 locations belonging to 179 ethnicities, spanning 18 African countries.[57] About half of respondents live outside their ancestral ethnic homeland. Ethnic endogamy is decreasing at lower (more disaggregated) branches of the Ethnologue tree. At level 6 of the Ethnologue tree, which roughly corresponds to the level of the Murdock's ethnic homelands maps, the average share of endogamous couples is 0.91 with a standard deviation of 0.28. For movers, the percentage is around 0.88 with a standard deviation around 0.3.[58]

Table 3 reports in Columns (1) to (4) estimates of the effect of malaria in the location of the respondent and the ancestral malaria on the likelihood ethnic endogamy at level 6 of the Ethnologue. Columns (5) to (8) replicate the analysis using the "exogamy index" which provides information on the linguistic distance between spouses.

[56]The inclusion of location fixed effects automatically drops country fixed effects.

[57]From the DHS waves from 1992-2012, we exploit surveys that: 1) contain the coordinates of the location of the female respondent; 2) information on the ethnicity of the spouses; 3) have information on individual characteristics.

[58]The summary statistics of the full sample and the sample of movers, including endogamy at all levels of the Ethnologue tree, are reported in Tables S7 and S8.

Table 3: Ancestral Exposure to Malaria and Ethnic Endogamy

	Endogamy (Dummy)				Exogamy (ethnolinguistic Distance)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Location Malaria	0.002*	0.003	0.002		-0.000	-0.000	0.000	
	(0.001)	(0.002)	(0.002)		(0.000)	(0.000)	(0.000)	
	[0.087]	[0.085]	[0.073]		[-0.030]	[-0.041]	[0.012]	
Ancestral Malaria			0.019***	0.029***			-0.007***	-0.010***
			(0.003)	(0.005)			(0.001)	(0.003)
			[0.531]	[0.821]			[-0.576]	[-0.745]
Sample	Full	Movers	Movers	Movers	Full	Movers	Movers	Movers
Ancestral Controls	No	No	Yes	Yes	No	No	Yes	Yes
Individual Controls	No	No	No	Yes	No	No	No	Yes
Size Group	No	No	No	Yes	No	No	No	Yes
Country FE	Yes	Yes	Yes	No	Yes	Yes	No	No
Wave FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Village FE	No	No	No	Yes	No	No	Yes	Yes
Observations	19416	9400	9400	9400	19416	9400	9400	9400
R-squared	0.12	0.17	0.20	0.53	0.02	0.02	0.06	0.43

Notes: The table reports the OLS estimates associating the probability of being endogamous (or the exogamy index) with the location and ancestral level of Malaria Stability. The dependent variable in Columns (1)-(4) 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. The dependent variable in Columns (5)-(8) index of exogamy measuring the linguistic distance between the spouses, constructed following Desmet et. al (2011). Location Malaria is the average level of the Malaria Stability in the respondent's location, and Ancestral Malaria is the average level of the Malaria Stability in the Murdock ethnic homeland of the respondent's ethnicity; see text for details. Ancestral controls include soil suitability and elevation (mean and standard deviation), average temperature and precipitation, terrain ruggedness, caloric suitability before 1500, distance from the equator, from the coast, from the river, from the country border and from the country capital, TseTse suitability and Predicted Genetic distance. The DHS Individual controls include urban dummy, years of education, age, religion fixed effects, and dummies of relative wealth (poorest, poorer, middle, richer, richest) for each respondent. Size of Group is the number of individuals belonging to the ethnic group of the respondent in the location region. The unit of observation is the female DHS respondent. Variable description, data sources, and summary statistics are reported in Tables E6, E7, and S7, and S8 respectively. Beta coefficient in square brackets. Robust standard errors clustered by ethnic group (DHS) are in parentheses. ***, **, and * indicate significance at the 1-, 5-, and 10-% levels, respectively.

Considered alone, high levels of malaria exposure in a location tend to increase the probability of being in an endogamous marriage (and to decrease, but not significantly, the ethnolinguistic distance of spouses) in the full sample, columns (1) and (5). Malaria of the location does not significantly affect ethnic endogamy for the movers, columns (2) and (6). Ancestral malaria is, in turn, a strong and statistically significant driver of both endogamy and exogamic distances, columns (3) and (7). Finally, columns (4) and (8) identify the effect of ancestral malaria by exclusively exploiting variation across individuals within the same location by including village fixed effects, accounting for all location-specific characteristics that affect the level of endogamy. The magnitude of the effect of ancestral malaria is sizable. For instance, depending on the specification, a 1 standard deviation increase in ancestral malaria makes endogamic marriages 13%-18% more likely.

4.3.2 Further Exploration

The baseline results are robust to several exploration and checks.

Endogamy at different levels of the Ethnologue Tree. As discussed above, the exogamic distance measure allows to address relevant caveats in the interpretation of the endogamy indicators. Nonetheless, it is interesting to explore the role of ancestral malaria for endogamy at different levels of the Ethnologue tree. Figure D14 depicts the estimated coefficients obtained replicating the baseline results of Table D1 for endogamy measured at various levels of the Ethnologue tree. Ancestral malaria increases endogamy throughout.

Movers: Drawing of Ethnicity Borders and Distances. Malaria, in either the location or in the ancestral homeland, is not a systematic relevant predictor of the probability of being a mover which is mostly explained by locations fixed effects (see Table D2). The sensitivity to possible errors in the identification of movers is explored by using Thiessen polygons transformation of homeland borders that leave the results unaffected (Table D3). We explore the existence of heterogeneous effects depending on the distance from historical homelands.[59] The results show that the effect of ancestral malaria on endogamic marriages is positive and can be detected regardless of the actual locations of movers (see Table D4).

Measures of Malaria Exposure. The patterns are confirmed also when exploiting measures of malaria exposure that do not rely on information on mosquitoes types (both in intention-to-treat and IV specifications although estimated with lower precision when including location fixed effects, see Tables D5 and D6) and the findings are confirmed, although again estimated with lower precision, also when using the information on genetic immunities to malaria as an alternative proxy of ancestral malaria exposure (see Table D7).

Placebo Diseases. As discussed above, multi-host diseases should not be expected to have had a prime impact on the incentives for spatial clustering and they do not affect the distribution of historical and pre-colonial ethnic groups. Also, differently from exposure to malaria, that has been documented to be a main selector of the human genome and should be expected to shape endogamic cultures, no systematic accounts exist suggesting a similar effect of other multi-host pathogens on sexual inbreeding. We replicate the placebo exercise by extending the analysis to the consideration of the ancestral role of trypanosomiasis, dengue and yellow fever. The results confirm the role of ancestral malaria while no evidence of significant patterns is detected for the other diseases, see Table D8.

[59] We check, in particular, whether the effect changes with the distance of the homeland from current location looking at individuals that moved less than 50 km, between 50 and 100 km, between 100 and 300 km or above.

5 Conclusion

This research provides the first attempt to systematically address the important, but so far largely unexplored, question of the drivers of the emergence of pre-colonial ethnic groups and of the mechanisms of the differential persistence of ethnic groups and cultures.

We offer a conceptual framework that allows to predict and rationalize the emergence of differential patterns of geographic segregation and behavioral isolation including preferential sexual endogamy. The theory extends an otherwise standard Malthusian set-up to the consideration of geographic clustering of human population and endogenous health in the presence of malaria. By incorporating well-established insights from malaria epidemiology and evolutionary genetics, it delivers a set of predictions on the specific patterns of voluntary isolation that allow sustaining population size in the face of malaria. The predictions, that are specific to malaria and not to other important vector-borne diseases, and the peculiarity of the global history of the disease also allowed us to devise specific falsification tests.

To explore the empirical validity of the hypothesis and identify the effects, we, therefore, build and exploit several novel databases. These allow, in particular, to study the drivers of the geographic distribution of pre-colonial ethnicities in both Africa and the Americas and the effect of malaria on ethnic diversity and ethnic identities today. We also propose a measurement of the ethnolinguistic distance between spouses and build a database and an empirical strategy that allows us to measure and study the determinants of ethnolinguistic endogamy in Africa today. A very extensive set of robustness checks confirm the baseline patterns, the specific predictions and the evidence on the mechanisms of persistence of African ethnicities.

The results suggest that one of the main legacies of the intense selective pressure of malaria in Africa can be traced in the emergence of, and persistence of, multiple ethnic groups with strong identities that are deeply rooted on the borders of ancestral homelands and on ancestral endogamic cultures. The results on the effect of ancestral malaria on ethnic marriages are also insightful on the prospects of differential persistence ethnic groups. The findings suggest that ethnicities that have been more intensively shaped by the cultural and genetic adaptation to the malaria ecology are more likely to face higher prospects of persistence also in the future.

Our findings deliver some interesting insights for future research. The predictions tested in this paper represent only a subset of testable predictions. A main untested prediction, that is in line with historical narratives but has not been empirically explored for lack of data, is about the role of malaria in shaping interactions with strangers above and beyond sexual reproduction. In particular, norms that limit visitation patterns, exchange and trade, shape trust, and inter-groups enmities (and possibly conflict). Finally, a related, but different, interesting question is about the existence

of a relationship between ethnolinguistic and epidemiological and possibly genetic distances between groups. Answering these questions would require further investment in developing a specific testable hypothesis and a dedicated effort for the construction of a different database (based on, e.g., dyads of groups as units of observation).

References

Alesina, A., and P. Giuliano (2015): "Culture and Institutions," *Journal of Economic Literature*, 53(4), 898–944.

Alesina, A., and E. La Ferrara (2005): "Ethnic Diversity and Economic Performance," *Journal of Economic Literature*, 43(3), 762–800.

Alesina, A., S. Michalopoulos, and E. Papaioannou (2016): "Ethnic Inequality," *Journal of Political Economy*, 124(2), 428–488.

Alland, A. (2008): *Evolution and Human Behaviour: An Introduction to Darwinian Anthropology*, vol. 1. Routledge.

Alsan, M. (2015): "The Effect of the TseTse Fly on African Development," *American Economic Review*, 105(1), 382–410.

Ashraf, Q., and O. Galor (2011): "Dynamics and Stagnation in the Malthusian Epoch," *American Economic Review*, 101(5), 2003–2041.

Ashraf, Q., and O. Galor (2013): "The Out of Africa Hypothesis, Human Genetic Diversity, and Comparative Economic Development," *American Economic Review*, 103(1), 1–46.

Bergstrom, T. C. (2002): "Evolution of Social Behavior: Individual and Group Selection," *Journal of Economic Perspectives*, 16(2), 67–88.

Brower, B., and B. R. Johnston (2007): *Disappearing Peoples?: Indigenous Groups and Ethnic Minorities in South and Central Asia*. Left Coast Press.

Caselli, F., and W. J. Coleman (2013): "On the Theory of Ethnic Conflicts," *Journal of the European Economic Association*, 11, 161–192.

Cavalli-Sforza, L., and M. W. Feldman (1981): *Cultural Transmission and Evolution*, vol. 1. Princeton University Press.

Chiovelli, G. (2016): "Pre-Colonial Centralization, Colonial Activities and Development in Latin America," *Mimeo, London Business School*.

Cobbah, J. A., and A. D. Smith (2014): "Toward a Geography of Peace in Africa: Redefining Sub-State Self-Determination Rights," *Nationalism, Self-Determination and Political Geography (Routledge Library Editions: Political Geography)*, 6, 70.

Costantini, C., N. Sagnon, A. Torre, and M. Coluzzi (1999): "Mosquito Behavioural Aspects of Vector-Human Interactions in the Anopheles Gambiae Complex," *Parassitologia*, 41(1/3), 209–220.

Denic, S., N. Nagelkerke, and M. M. Agarwal (2008): "Consanguineous Marriages and Endemic Malaria: Can Inbreeding Increase Population Fitness," *Malaria Journal*, 7(1), 150.

Denic, S., and M. G. Nicholls (2007): "Genetic Benefits of Consanguinity through Selection of Genotypes Protective against Malaria.," *Human Biology*, 7(1), 150.

Depetris-Chauvin, E., and D. N. Weil (2018): "Malaria and Early African Development: Evidence from the Sickle Cell Trait," *Economic Journal*, 128(610), 1207–1234.

Desmet, K., I. Ortuño Ortín, and R. Wacziarg (2012): "The Political Economy of Ethnolinguistic Cleavages," *Journal of Development Economics*, 97(2), 322–338.

Diamond, J. (1997): "Guns, Germs, and Steel: the Fates of Human Societies," *New York: W. W. Norton*.

Diamond, J. (2011): "Guns, Germs, and Steel - Transcript Episode 1: Out of Eden," .

Endalew, T. (2006): *Inter-Ethnic Relations on a Frontier: Mätakkäl (Ethiopia), 1898-1991*, vol. 69. Otto Harrassowitz Verlag.

Esposito, E. (2018): "Side Effects of Immunity: The Rise of African Slavery in the US South," Discussion paper, Université de Lausanne, Faculté des HEC, DEEP.

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

Fenske, J. (2014): "Ecology, Trade and States in Pre-Colonial Africa," *Journal of the European Economic Association*, (12), 612–640.

Fincher, C. L., and R. Thornhill (2012): "Parasite-Stress Promotes In-Group Assortative Sociality: The Cases of Strong Family Ties and Heightened Religiosity," *Behavioral and Brain Sciences*, 35(02), 61–79.

Franck, R., and I. Rainer (2012): "Does the Leader's Ethnicity Matter? Ethnic Favoritism, Education, and Health in Sub-Saharan Africa," *American Political Science Review*, 106, 294–325.

Galor, O., and S. Michalopoulos (2012): "Evolution and the Growth Process: Natural Selection of Entrepreneurial Traits," *Journal of Economic Theory*, 147(2), 759–780.

Galor, O., and Ö. Özak (2016): "The Agricultural Origins of Time Preference," *American Economic Review*, 106(10), 3064–3103.

Gennaioli, N., and I. Rainer (2007): "The Modern Impact of Precolonial Centralization in Africa," *Journal of Economic Growth*, 12(3), 185–234.

Gething, P. W., T. P. Van Boeckel, D. L. Smith, C. A. Guerra, A. P. Patil, R. W. Snow, and S. I. Hay (2011): "Modelling the Global Constraints of Temperature on Transmission of Plasmodium Falciparum and P. vivax," *Parasites & Vectors*, 4(1), 92.

Giuliano, P., and N. Nunn (2017): "Understanding Cultural Persistence and Change," Discussion paper, National Bureau of Economic Research.

Horowitz, D. L. (1985): *Ethnic Groups in Conflict*. University of California Press.

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.

King, V. T. (2013): *Environmental Challenges in South-East Asia*. Routledge.

Kiszewski, A., A. Mellinger, A. Spielman, P. Malaney, S. E. Sachs, and J. Sachs (2004): "A Global Index Representing the Stability of Malaria Transmission," *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.

Lowes, S. R., and E. Montero (2018): "The Legacy of Colonial Medicine in Central Africa," Discussion paper, CEPR Discussion Papers.

Lysenko, A. J., and N. I. Semashko (1968): "Geography of Malaria. A Medico-Geographic Profile of an Ancient Disease," *Itogi Nauki: Medicinskaja Geografija*, pp. 25–146.

Macdonald, G. (1956): "Epidemiological Basis of Malaria Control," *Bulletin of the World Health Organization*, 15(3-5), 613.

Mandal, S., R. Sarkar, and S. Sinha (2011): "Mathematical Models of Malaria - A Review," *Malaria Journal*, 10(1), 202.

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

Michalopoulos, S. (2012): "The Origins of Ethnolinguistic Diversity," *American Economic Review*, 102(4), 1508–1539.

Michalopoulos, S., and E. Papaioannou (2013a): "National Institutions and Subnational Development in Africa," *Quarterly Journal of Economics*, 129(1), 151–213.

——— (2013b): "Pre-Colonial Ethnic Institutions and Contemporary African Development," *Econometrica*, 81(1), 113–152.

——— (2016): "The Long-Run Effects of the Scramble for Africa," *American Economic Review*, 106(7), 1802–1848.

Michalopoulos, S., L. Putterman, and D. N. Weil (2018): "The Influence of Ancestral Lifeways on Individual Economic Outcomes in Sub-Saharan Africa," *Journal of the European Economic Association, forthcoming.*

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.

Neafsey, D. E., R. M. Waterhouse, M. R. Abai, S. S. Aganezov, M. A. Alekseyev, J. E. Allen, J. Amon, B. Arcà, P. Arensburger, G. Artemov, et al. (2015): "Highly evolvable malaria vectors: the genomes of 16 Anopheles mosquitoes," *Science*, 347(6217), 1258522.

Nunn, N. (2008): "The Long-term Effects of Africa's Slave Trades," *The Quarterly Journal of Economics*, 123(1), 139–176.

——— (2012): "Culture and the Historical Process," *Economic History of Developing Regions*, 27(s1), 108–126.

Nunn, N., and L. Wantchekon (2011): "The Slave Trade and the Origins of Mistrust in Africa," *American Economic Review*, 101(7), 3221–52.

Ramen, F. (2002): *Sleeping Sickness and Other Parasitic Tropical Diseases*. The Rosen Publishing Group.

Reich, D. (2018): *Who We Are and How We Got Here: Ancient DNA and the new science of the human past*. Oxford University Press.

Robson, A. J., and L. Samuelson (2010): "The Evolutionary Foundations of Preferences," *Handbook of Social Economics*, 1, 221–310.

Ross, R. (1909): "Report on the Prevention of Malaria in Mauritius," *Churchill, London.*

Sabeti, P. C., S. F. Schaffner, B. Fry, J. Lohmueller, P. Varilly, O. Shamovsky, A. Palma, T. Mikkelsen, D. Altshuler, and E. Lander (2006): "Positive Natural Selection in the Human Lineage," *Science*, 312(5780), 1614–1620.

Sachs, J. D. (2003): "Institutions Don't Rule: Direct Effects of Geography on Per Capita Income," *NBER Working Paper*, 9490.

Torres-Sorando, L., and D. J. Rodriguez (1997): "Models of Spatio-Temporal Dynamics in Malaria," *Ecological Modelling*, 104(2), 231–240.

Van den Berghe, P. L. (1987): *The Ethnic Phenomenon*. ABC-CLIO.

Wacziarg, R., and E. Spolaore (2013): "How Deep are the Roots of Economic Development," *Journal of Economic Literature*, 51(2), 1–45.

Webb, J. L. (2006): "Ecology and Culture in West Africa," *Themes in West Africa's History, ed. Emmanuel Akyeampong (Athens: Ohio University Press, 2006)*.

Webb, J. L. A. J. (2009): *Humanity's Burden: a Global History of Malaria*, vol. 69. Cambridge University Press.

Weil, D. N. (2017): "Gyrations in African Mortality and Their Effects on Economic Growth," *Journal of Demographic Economics*, 83(1), 103–110.

Bite and Divide: Malaria and Ethnolinguistic Diversity

Supplementary Appendix

Matteo Cervellati
University of Bologna
CEPR, CESIfo and IZA

Giorgio Chiovelli
London Business School

Elena Esposito
University of Lausanne

December 30, 2018

Contents

A1 Analytical Derivations and Proofs

A1.1 Lemma 1: Health

Proof. The stationary state of the malaria transmission model (5) is characterized by setting $\dot{\sigma} = \dot{\mu} = 0$ and solving the system,[1]

$$(1-\sigma) \times \mu M \times s = r\sigma \tag{A1}$$

$$(1-\mu) \times \sigma \left(L_t/G\right) = d\mu \tag{A2}$$

Isolating μ from (A2) one gets,

$$\mu = \sigma \left[\frac{1 + \frac{r}{Ms}}{1 + \frac{d \times G}{L}} \right] \tag{A3}$$

and substituting back into (A1) and solving for $(1-\sigma)$ we have,

$$(1-\sigma) = \left(1 + \frac{d \times G}{L}\right) \left(\frac{r}{r + s \times M}\right) \tag{A4}$$

The level of equilibrium health for given population density (8) is finally obtained using (A4) and (6) directly into (7). □

A1.2 Lemma 2: Malthusian Stationary State

Proof. For a given level of health (8) characterized in Lemma 1, the characterization of the level of population density in the stationary state of the Malthusian model follows closely the derivation in Ashraf and Galor (2011). In particular, given the set-up in Section 2.1, optimal net fertility is proportional to per capita income with a proportionality factor γ/ρ. Substituting net fertility into the law of motion of population dynamics in equation (4), solving for the stationary state $L_{t+1} = L_t = L$ and rearranging gives the level of Malthusian population conditional on health of equation (9). □

A1.3 Proposition 1: Optimal Number of Groups and Sexual Endogamy

Proof. From (8) and (9), the level of population density in the Malthusian stationary state with endogenous health, denoted as L to simplify notation, can be expressed as,

$$L = \left(\frac{\gamma}{\rho}\right)^{\frac{1}{\alpha}} X A(g) \left(1 + \frac{d \times G}{L}\right) \left(\frac{r}{r + f(e, M) \times M}\right) (1 - p(e)) \tag{A5}$$

[1]For notational clarity, the reference to equations in the main body is made using their sequential number as it appears in the main text while new equations introduced in this appendix are preceded by an "A".

The explicit formulation for the level of population density in the Malthusian stationary state with health can be computed by expanding (A5) as quadratic formulation and solving for the positive root to get,[2]

$$L(G,e)=\left(\frac{\gamma}{\rho}\right)^{\frac{1}{\alpha}}\frac{XA(g)}{2}\left(\frac{r\left(1-p(e)\right)}{r+f(e,M)\times M}\right)\left(1+\left(1+\frac{4d}{(\gamma/\rho)^{\frac{1}{\alpha}}X}\frac{G}{A(g)}\frac{(r+f(e,M)\times M)}{r(1-p(e))}\right)^{\frac{1}{2}}\right) \tag{A6}$$

The levels of endogamy and number of groups, $\{e^*,G^*\}$ (henceforth simply denoted $\{e,G\}$) that maximize Malthusian population density population in the interior optimum are implicitly characterized by the solution to the system,

$$\frac{\partial L(G,e)}{\partial G} = 0 \tag{A7}$$

$$\frac{\partial L(G,e)}{\partial e} = 0 \tag{A8}$$

Optimal Endogamy. Let us start by (A7). Rewrite (A6) as

$$L(G,e)=\Phi(M,e)A(G)\left[1+\left(1+\frac{2d}{\Phi(M,e)}\frac{G}{A(G)}\right)^{\frac{1}{2}}\right] \tag{A9}$$

where,

$$\Phi(e,M)\equiv\left(\frac{\gamma}{\rho}\right)^{\frac{1}{\alpha}}\frac{X}{2}\left(\frac{r\left(1-p(e)\right)}{r+f(e,M)\times M}\right) \tag{A10}$$

Equation (A7) implies,

$$\frac{\partial L(G,e)}{\partial e}=\frac{\partial L(\cdot)}{\partial \Phi(\cdot)}\frac{\partial \Phi(\cdot)}{\partial e(\cdot)}=0$$

[2]Expanding (A5) as quadratic equation we have

$$L^2-cL-cdG=0$$

where,

$$c\equiv\left(\frac{\gamma}{\rho}\right)^{\frac{1}{\alpha}}XA(g)\left(\frac{r}{r+f(e,M)\times M}\right)(1-p(e))$$

Solving for the positive root of the quadratic equation and rearranging we obtain the explicit solution

$$L=\frac{c}{2}\left(1+\left(1+\frac{4dG}{c}\right)^{\frac{1}{2}}\right)$$

which gives (A6) when substituting for c.

Notice that $\partial L(G,e)/\partial \Phi(M,e) > 0$.[3] Hence the characterization of optimal endogamy requires,

$$\frac{\partial \Phi(M,e)}{\partial e} = 0 \Leftrightarrow \frac{\partial}{\partial e}\left(\frac{1-p(e)}{r+f(e,M)\times M}\right) = 0 \tag{A11}$$

Notice that the optimal level of endogamy does not depend on G.

To study the change in optimal e in response to an increase in M recall the features of the function $f(\cdot)$ from Section 2.2 and rewrite (A11) explicitly as,

$$\frac{-p'(e)}{f_e(e)(1-p(e))} - \frac{M}{r+f(e,M)M} = 0$$

Applying implicit function theorem to this first order condition, considering that the partial derivative of the implicit function with respect to e is negative from second order condition, and since (from Section 2.2) $f_M(M) > 0$, we finally have

$$sign\left\{\frac{\partial e^*}{\partial M}\right\} = sign\left\{\frac{\partial}{\partial M}\left[\frac{1}{\frac{r}{M}+f(e,M)}\right]\right\} > 0$$

Optimal number of groups, G. Rewrite the stationary state of population density, (A6), as

$$L = \left[\left(\frac{\gamma}{\rho}\right)^{\frac{1}{\alpha}}\frac{X}{2} \times \frac{r\,(1-p(e))}{r+\tilde{M}}\right] \times A(G)\Psi(G,\tilde{M}) \tag{A12}$$

where

$$\Psi(G,\tilde{M}) \equiv 1 + \left(1 + \frac{4d}{(\gamma/\rho)^{\frac{1}{\alpha}}X}\frac{G}{A(G)}\frac{r+\tilde{M}}{r(1-p(e))}\right)^{\frac{1}{2}} \tag{A13}$$

and $\tilde{M} \equiv f(\cdot)\times M$ denotes the effective malaria exposure.

To simplify the exposition, we also use the short notations for the function $\Psi = \Psi(G,\tilde{M})$, for the partial derivatives $\Psi_G = \Psi_G(G,\tilde{M})$, $\Psi_{\tilde{M}} = \Psi_{\tilde{M}}(G,\tilde{M})$, and for the cross derivative $\Psi_{G,\tilde{M}} = \Psi_{G,\tilde{M}}(G,\tilde{M})$. Using (A13), the first order condition for the optimal number of groups (A8) can be rewritten as,

$$A'(G)\Psi + A(G)\Psi_G = 0 \tag{A14}$$

Notice that from second order condition the partial derivative of this implicit function with respect

[3]Since, rewriting (A9),

$$L(G,e) = \Phi(M,e)^{\frac{1}{2}}A(G)\left[\Phi(M,e)^{\frac{1}{2}} + \left(\Phi(M,e) + \frac{2dG}{A(G)}\right)^{\frac{1}{2}}\right]$$

which is a strictly increasing function of $\Phi(M,e)$.

to G is negative and that $\partial\tilde{M}/\partial M > 0$ from Section 2.2. By implicit function theorem, we have[4]

$$sign\left\{\frac{\partial G}{\partial M}\right\} = sign\left\{\Psi_{G,\tilde{M}}\Psi - \Psi_G\Psi_{\tilde{M}}\right\}$$

which is equivalent to studying the monotonicity of the function Ψ_G/Ψ in $\tilde{M}$.[5] Denoting by

$$x(\tilde{M}) \equiv \frac{r+\tilde{M}}{r(1-p(e))} \tag{A15}$$

which is an increasing linear function of $\tilde{M}$ and denoting by

$$\nu \equiv \frac{4d}{(\gamma/\rho)^{\frac{1}{\alpha}}X}\frac{G}{A(G)} \tag{A16}$$

Recalling that $\frac{\partial}{\partial G}\left[\frac{G}{A(G)}\right] > 0$, we have,[6]

$$\frac{\partial}{\partial\tilde{M}}\left[\frac{\Psi_G}{\Psi}\right] = \frac{\partial}{\partial\tilde{M}}\left[\frac{x(\tilde{M})}{1+[1+\nu x(\tilde{M})]^{(1/2)}+\nu x(\tilde{M})}\right] > 0. \tag{A17}$$

From the previous discussion and from (A15), the optimal number of groups therefore increases in the level of malaria exposure, $\partial G/\partial M$.

□

[4]The partial derivative of the first order condition (A14) with respect to M is given by,

$$\left[A'(G)\Psi_{\tilde{M}}(G,\tilde{M}) + A(G)\Psi_{G,\tilde{M}}(G,\tilde{M})\right]\frac{\partial\tilde{M}}{\partial M}$$

Condition (A15) is obtained substituting $A'(G) = -A(G)\Psi_G/\Psi$ derived from (A14), in the expression above.

[5]Notice that,

$$sign\left\{\frac{\partial}{\partial\tilde{M}}\left[\frac{\Psi_G}{\Psi}\right]\right\} = sign\left\{\frac{\Psi_{G,\tilde{M}}\Psi - \Psi_G\Psi_{\tilde{M}}}{\Psi^2}\right\}$$

[6]Given $x(\tilde{M})$ and ν, rewrite $\Psi = 1+(1+\nu x(\tilde{M}))^{\frac{1}{2}}$ and $\Psi_G = \frac{1}{2}(1+\nu x(\tilde{M})^{\frac{1}{2}-1}x(\tilde{M})\frac{\partial}{\partial G}\frac{G}{A(G)}$. Taking the ratio between these functions implies (A17) since the resulting function is strictly increasing in $x(\tilde{M})$ which, in turns, is increasing in $\tilde{M}$.

A2 Data

A2.1 Alternative Indexes of Malaria Transmission

Our baseline measure of malaria transmission is the malaria stability index produced by Kiszewski et al. (2004), which combines information on characteristics of prevalent mosquitoes in the region and climate features facilitating mosquitoes' activities related to malaria transmission. We verify our findings using two alternative indexes of malaria transmission that ignore variation in mosquitoes' vector across regions and that predict mosquitoes' activity in the location solely based on long-term climatic averages.

Temperature-based Predicted Malaria Stability (falciparum) We reconstruct a malaria stability index that follows Kiszewski et al. (2004) but that ignores the variation of mosquitoes vector across locations.[7] The mosquitoes characteristics in the index, namely proportion biting people and daily survival rate, are predicted as functions of long-term temperatures following the epidemiological literature.

More precisely, we compute a malaria index following the equation:

$$\sum_{m=1}^{12} \frac{a_m^2 p_m^E}{-lnp_m}$$

where m stands for month. E is the length of extrinsic incubation period in days ($E = \frac{111}{t-16}$ for falciparum). The parameter a represents the proportion biting people and p the mosquito daily survival rate. We calibrate the parameter p, following McCord (2017), as $p(T) = exp^{-1/(-4.4+(1.31*T)-(0.03*T^2))}$ where T is temperature. To parameterize the mosquito biting rate we follow Garske et al. (2013), who estimate $a(T)$ looking at *Anopheles maculipennis*, a vector similar to *Anopheles gambiae*, and find a biting rate dependence on temperature of $a(T) = max(0, (1 + \frac{36.5}{T-9.9}))$. To render this measures better comparison with the Malaria Stability of Kiszewski et al. (2004), we impose the same minimum lagged threshold of monthly precipitation (10 mm) (that they introduce as a pre-condition for malaria transmission).

Plasmodium Falciparum Suitability To verify the validity of this approach, in terms of parametrization of temperature dependence, we verify our findings with a very similar index taken from the epidemiological literature, and namely the index of the suitability of malaria transmission devised and estimated by Gething et al. (2011). Exactly like the Temperature-based Predicted Malaria Stability index that we created - described in the paragraph above - this index is constructed based

[7] See McCord (2017).

on a model attempting to incorporate all the principal mechanisms of temperature dependency in malaria transmission. While very close in spirit, the two indexes present small differences mostly related to modelling choices and parametrizations.

Comparison across measures. We plot the distribution of the different measures in Figure D1. Panel A of Figure D1 shows the correlation between our own Malaria Stability (Temperature) and the Plasmodium Falciparum Suitability from Gething et al., 2011. The two measures are correlated at 92%. Panel B and Panel C plots the distribution between our baseline Malaria Stability and its two temperature-based proxy. Finally, Figure D2 plots the spatial distribution of Malaria Stability (Temperature) and the Malaria Transmission (Temperature) across our baseline sample of 1 x 1 degree cells.

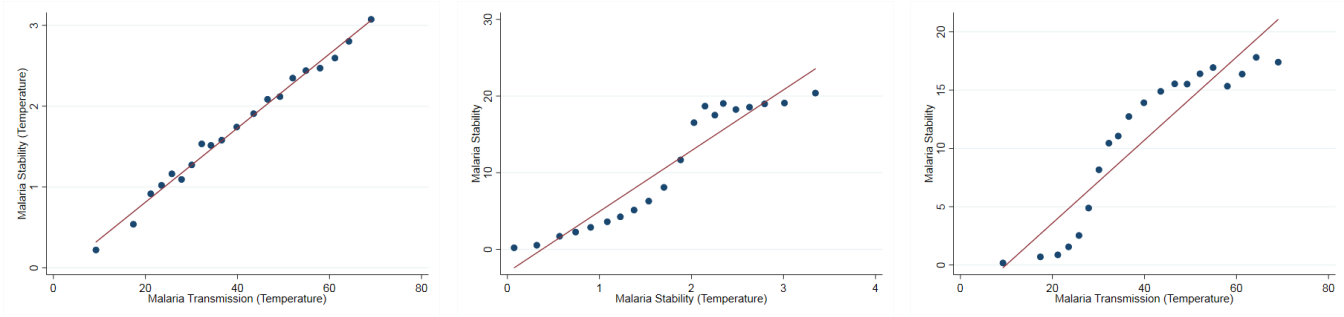
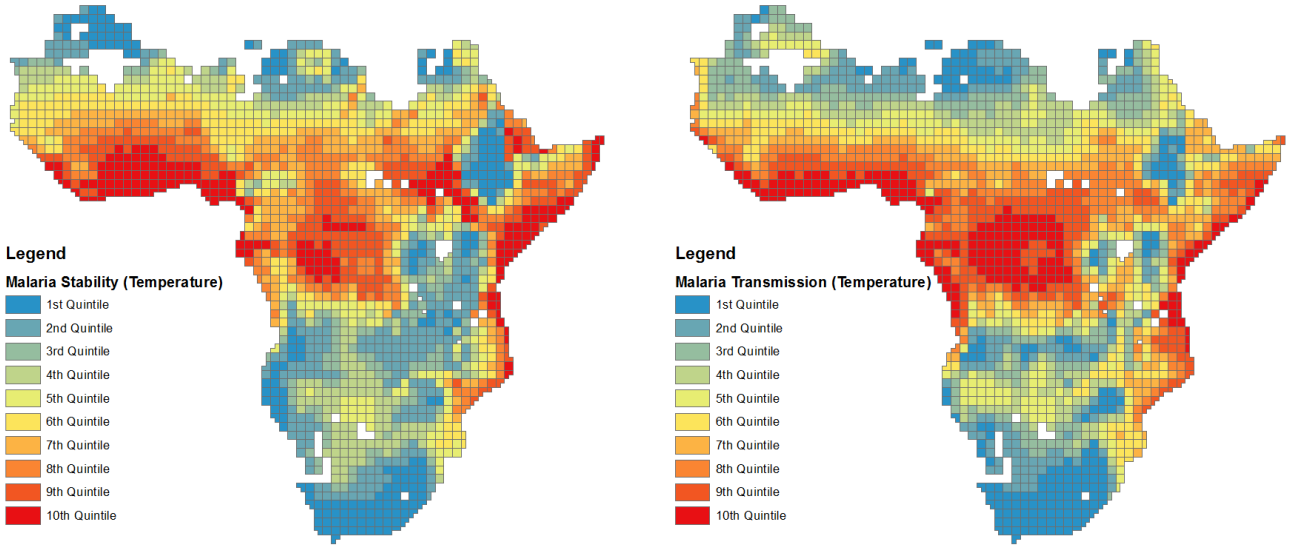


Figure D1: **Bin Scatter - Malaria Indexes**
The graph on the left is a binned scatterplot between Malaria Stability (Temperature) and the Plasmodium Falciparum Suitability index from Gething et al. (2011). The graph in the center is a binned scatterplot between Malaria Stability from Kiszewski et al. (2004) and our index of Malaria Stability (Temperature). The graph on the right is a binned scatterplot between Malaria Stability from Kiszewski et al. (2004) and the Plasmodium Falciparum Suitability index from Gething et al. (2011).



Panel A: Malaria Stability (Temperature) Panel B: Plasmodium Falciparum Suitability (Gething et al. 2011)

Figure D2: **Spatial Distribution of Temperature-Based Malaria Indexes**

Malaria Endemicity 1900 This measure of historical Malaria Endemicity attempts to map the levels of people parasitization at the beginning of the twentieth century. The original map was produced from Lysenko and Semashko (1968), and we use the version recently digitized by Hay et al. (2004). The advantage of this population-based measure of malaria is to be the only available dissaggregate measure of actual historical malaria prevalence. The main limitation, from the perspective of disaggregate analysis, is the low spatial resolution of the data. The information is categorical and, being assembled from scattered medical studies performed across different locations, the construction of continuous surfaces of data required heavy spatial interpolation. 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. Endemicity is defined as the parasitization rate (PR) in the 2-10 year age cohort. 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. A map of the measure aggregated at 1x1 degree cell level is reported on the left panel of Figure D3.

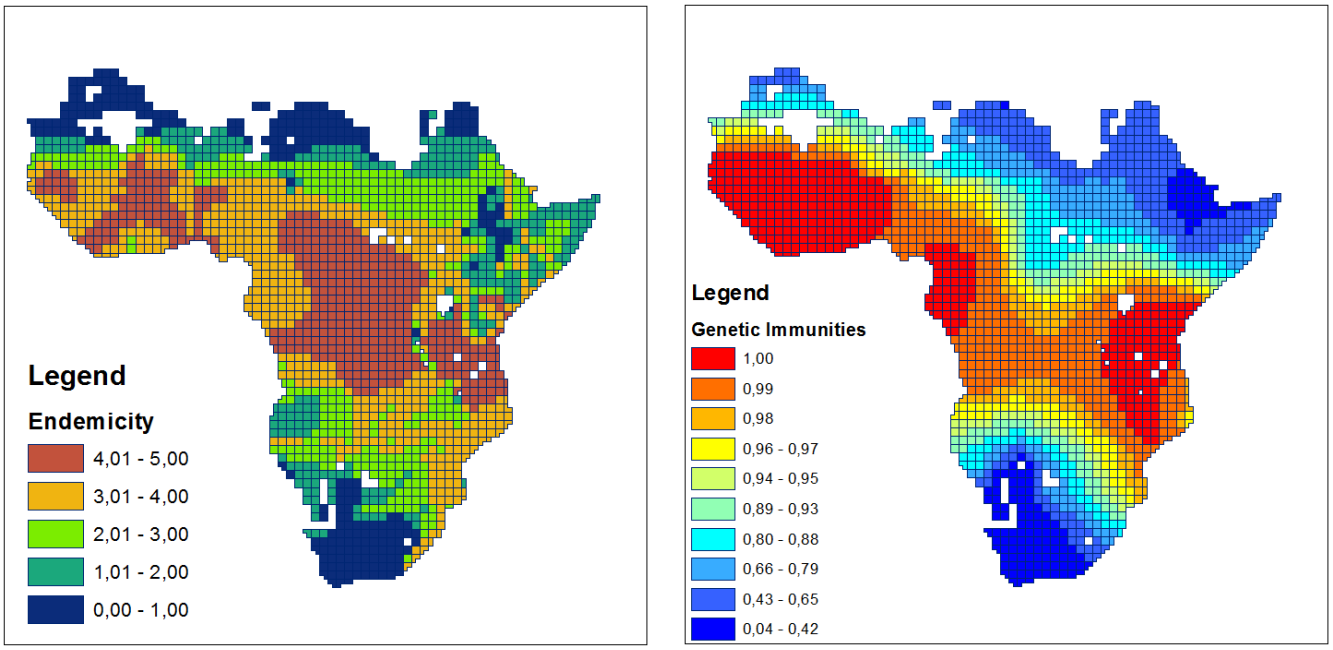


Figure D3: **Malaria Endemicity in 1900 and Duffy Antigen Distribution**

Duffy Negative Phenotype The Duffy antigen is a trait that protects against malaria by preventing the disease from entering the red blood cells. This mechanism takes place through the alteration of the structure of the Duffy glycoprotein on the surface of the red blood cell. The Duffy antigen confers an almost complete resistance from infection with malaria plasmodium vivax, but it also protects against the more severe falciparum variant.

We exploit information on genetic immunities in Africa that are built using information on blood

tests in the period between 1950 and 2010 (Howes et al. 2011). The data are an interpolated raster on the prevalence of the Duffy antigen. A map of the measure aggregated at 1x1 degree cell level is reported on the right panel of Figure D3. Note that information on this trait is likely affected by population movements during colonial and post-colonial times, recent admixing of groups, and migrations over the last decades.

A2.2 Murdock Ethnicity America

We retrieve information on the spatial distribution of ethnic groups in the Americas from Murdock's Outline of South American Cultures (1951) and Ethnographic Bibliography of North America, digitized by Chiovelli (2016). These maps represent the best attempt to depict the spatial distribution of the pre-colonial ethnic homeland at the eve of European colonization in North and South Americas. As for the widely-known Africa's map, data always pertain to the period of earliest European contact. Similarly to the Murdock map for Africa, the ethnic groups usually consist of a single tribe with a specific culture or to a group of tribes sharing common cultural traits.

In his work, Murdock (1951) maps the spatial distribution of pre-colonial ethnic homelands in Central and South America. For each country in Central and South America, we have a map depicting the geographical distribution of its corresponding pre-colonial ethnic groups. Figures D4 shows the original ethnolinguistic maps for Bolivia, Colombia, and Venezuela. We derive information on North America and Mexico from Murdock (1959), where the whole ethnic homeland distribution of continental North America is mapped.

Figure D5 shows the results of the digitization process (Chiovelli, 2016). The final dataset contains 488 ethnic homelands. In Latin America, we count 330 ethnic groups before the arrival of Columbus. The ethnic groups charted spans from Araucanians (Chile), Aztec (Mexico), Charrua (Uruguay), Cherokee (USA), Chibcha (Colombia), Inca (Peru), Guarani (Brazil and Paraguay), Tehuelche (Argentina), and Tupinamba (Brazil).

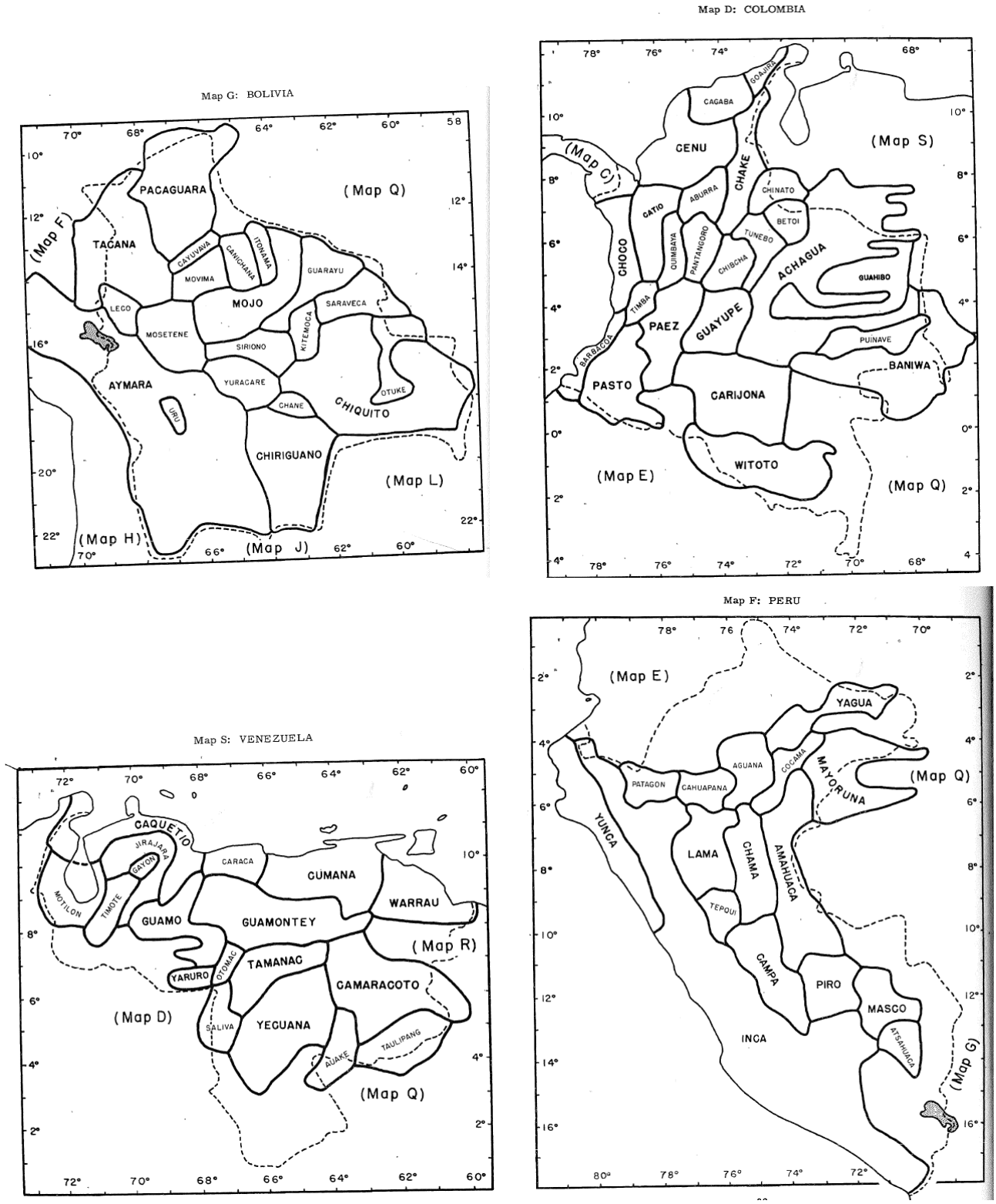


Figure D4: **Original Maps of Pre-Colonial Ethnic Homelands in Bolivia, Colombia, Venezuela, and Peru**

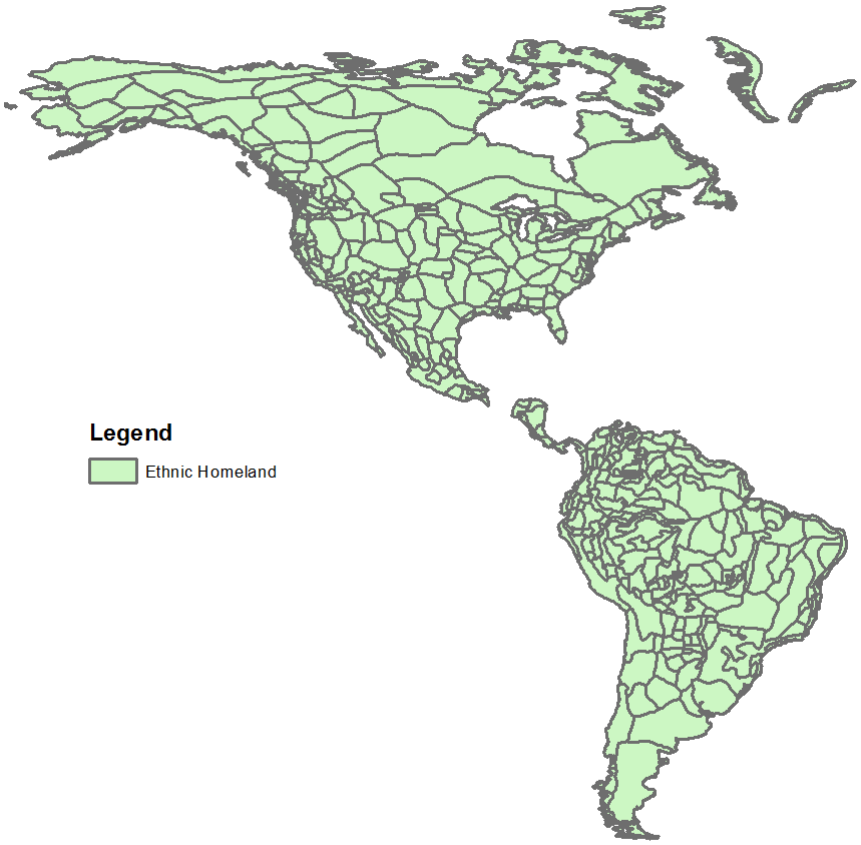


Figure D5: **Digitized Map of Murdock's Americas (Chiovelli, 2016)**

A3 Cell-level Analysis: Further Results and Robustness

We have performed a plethora of sensitivity checks to assess the robustness of the association between malaria stability and the number of ethnic groups.

A3.1 Robustness of Baseline

A3.1.1 Index of Land Fractionalization

We examine whether the results are robust to an alternative definition of ethnic diversity. We exploit the non-overlapping nature of GREG ethnolinguistic polygons and construct an index of territorial ethnic fractionalization. The index can be interpreted as the probability that two randomly drawn portion of the cell belonging to two different ethnic groups. Figure D6 illustrates the spatial distribution of the fractionalization index at cell level. Results are presented in Table A1. A one (cross-cells) standard deviation increase in malaria stability increases the fractionalization by 0.28 standard deviation. The effect implies an increase of the average fractionalization index by about 30 percent of the sample mean (0.233).

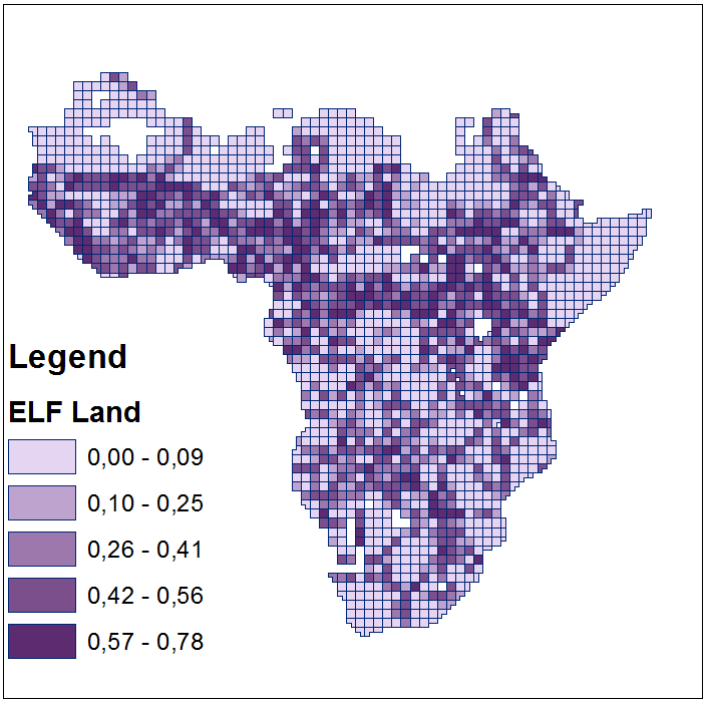


Figure D6: **Territorial Ethnic Fractionalization Distribution**

Table A1: **Land Fractionalization Index**

	Ethno-Lingustic Land Fractionalization				
	(1)	(2)	(3)	(4)	(5)
Malaria Stability	0.007***	0.007***	0.005***	0.005***	0.004**
	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)
	[0.275]	[0.278]	[0.210]	[0.184]	[0.174]
Geographic Controls:					
Ln(Cell Area)	Yes	Yes	Yes	Yes	Yes
Soil Suitability and Elevation (Mean and Std.)	No	No	Yes	Yes	Yes
Avg. Temperature and Precipitation	No	No	Yes	Yes	Yes
Ruggedness and Caloric Suitability Pre 1500	No	No	Yes	Yes	Yes
Location Controls:					
Distances (equator, coast, river, border, capital)	No	No	No	Yes	Yes
Number of Countries and Within Country	No	No	No	Yes	Yes
TseTse suit. & Pr. Genetic Distance	No	No	No	No	Yes
Country FE	No	Yes	Yes	Yes	Yes
Observations	1,976	1,976	1,976	1,976	1,976
R-squared	.0999	.207	.24	.274	.28

Notes: The Table reports the OLS specification estimates associating the ethnic fractionalization index with the level of Malaria Stability in Africa. In all specifications, the dependent variable is the territorial fractionalization index computed using GREG data at cell level; Malaria Stability is the average level of malaria suitability in the cell; see text for details. The geographic variables include the natural logarithm of the cell land area, soil suitability and elevation (mean and standard deviation), average temperature and precipitation, terrain ruggedness and caloric suitability before 1500. The distance variables include distance from the equator, from the coast, from the river, from the country border and from the country capital. Number of countries is the number of countries whose land falls within the cell. The within-country is a dummy variable taking value 1 if the cell is fully contained in a country. TseTse suitability measures the geographic suitability for Trypanosomiasis transmission. Predicted Genetic distance is computed as the log of migratory distance from East Africa. Variable description, data sources and summary statistics are reported in Tables E1, E2, and S1, respectively. Beta coefficient in square brackets. Robust standard errors clustered by country are reported in round brackets. ***, **, and * indicate significance at the 1-, 5-, and 10-% levels, respectively.

A3.1.2 Within Malarial Areas

We also examine the stability of our results when restricting the sample to those cells with a strictly positive level of Malaria Stability. The number of cells with positive malaria is 1,584 (80% of the baseline sample). In this sub-sample, both the average malaria stability and the average number of ethnic groups is higher than in the baseline sample (13.27 and 2.32, respectively). The results suggest the existence of an intensive margin effect, that goes beyond the extensive margin effect comparing places with and without malaria transmission.

Table A2: **Sample of Cells with positive Malaria Stability**

	Ln (Number of Groups - GREG)				
	(1)	(2)	(3)	(4)	(5)
Malaria Stability	0.015***	0.012**	0.014***	0.014***	0.013***
	(0.003)	(0.005)	(0.004)	(0.004)	(0.004)
	[0.235]	[0.186]	[0.226]	[0.222]	[0.209]
Geographic Controls:					
Ln(Cell Area)	Yes	Yes	Yes	Yes	Yes
Soil Suitability and Elevation (Mean and Std.)	No	No	Yes	Yes	Yes
Avg. Temperature and Precipitation	No	No	Yes	Yes	Yes
Ruggedness and Caloric Suitability Pre 1500	No	No	Yes	Yes	Yes
Location Controls:					
Distances (equator, coast, river, border, capital)	No	No	No	Yes	Yes
Number of Countries and Within Country	No	No	No	Yes	Yes
TseTse suit. & Pr. Genetic Distance	No	No	No	No	Yes
Country FE	No	Yes	Yes	Yes	Yes
Observations	1,584	1,584	1,584	1,584	1,584
R-squared	.0849	.222	.275	.304	.307

Notes: The Table reports the OLS specification estimates associating the log number of ethnic groups with the level of Malaria Stability, restricting the sample to cells where the Malaria Stability is larger than zero. In all specifications, the dependent variable is the natural logarithm of the number of ethnic groups in the cell; Malaria Stability is the average level of malaria suitability in the cell. The geographic variables include the natural logarithm of the cell land area, soil suitability and elevation (mean and standard deviation), average temperature and precipitation, terrain ruggedness and caloric suitability before 1500. The distance variables include distance from the equator, from the coast, from the river, from the country border and from the country capital. Number of countries is the number of countries whose land falls within the cell. The within-country is a dummy variable taking value 1 if the cell is fully contained in a country. TseTse suitability measures the geographic suitability for Trypanosomiasis transmission. Predicted Genetic distance is computed as the log of migratory distance from East Africa. Variable description, data sources and summary statistics are reported in Tables E1, E2, and S1, respectively. Beta coefficient in square brackets. Robust standard errors clustered by country are reported in round brackets. ***, **, and * indicate significance at the 1-, 5-, and 10-% levels, respectively.

A3.1.3 Within cells with more than one group

We also estimate a quite restrictive specification looking only at the sample of those cells with more than one ethnic group. By focusing on the intensive margin, we are left with 1,223 cells (62% of the total sample). The average malaria stability is 12.96 and the average number of ethnic group in the cell is 2.82. As shown in Table A3, the estimate is smaller than the baseline ones, reported in Table 1, though precisely estimated. In spite of relying solely on variation in the intensive margin, the link between the number of ethnic groups and malaria stability retains economic and statistical significance.

Table A3: **Sample of Cells with More Than One Ethnic Group**

	Ln (Number of Groups - GREG)				
	(1)	(2)	(3)	(4)	(5)
Malaria Stability	0.010***	0.008***	0.009***	0.009***	0.009***
	(0.001)	(0.002)	(0.003)	(0.003)	(0.003)
	[0.289]	[0.231]	[0.257]	[0.271]	[0.260]
Geographic Controls:					
Ln(Cell Area)	Yes	Yes	Yes	Yes	Yes
Soil Suitability and Elevation (Mean and Std.)	No	No	Yes	Yes	Yes
Avg. Temperature and Precipitation	No	No	Yes	Yes	Yes
Ruggedness and Caloric Suitability Pre 1500	No	No	Yes	Yes	Yes
Location Controls:					
Distances (equator, coast, river, border, capital)	No	No	No	Yes	Yes
Number of Countries and Within Country	No	No	No	Yes	Yes
TseTse suit. & Pr. Genetic Distance	No	No	No	No	Yes
Country FE	No	Yes	Yes	Yes	Yes
Observations	1,223	1,223	1,223	1,223	1,223
R-squared	.0977	.193	.226	.263	.264

Notes: The Table reports the OLS specification estimates associating the log number of ethnic groups with the level of Malaria Stability, restricting the sample to those cells with more than one ethnolinguistic group. In all specifications, the dependent variable is the natural logarithm of the number of ethnic groups in the cell; Malaria Stability is the average level of malaria suitability in the cell; see text for details. The geographic variables include the natural logarithm of the cell land area, soil suitability and elevation (mean and standard deviation), average temperature and precipitation, terrain ruggedness and caloric suitability before 1500. The distance variables include distance from the equator, from the coast, from the river, from the country border and from the country capital. Number of countries is the number of countries whose land falls within the cell. The within-country is a dummy variable taking value 1 if the cell is fully contained in a country. TseTse suitability measures the geographic suitability for Trypanosomiasis transmission. Predicted Genetic distance is computed as the log of migratory distance from East Africa. Variable description, data sources and summary statistics are reported in Tables E1, E2, and S1, respectively. Beta coefficient in square brackets. Robust standard errors clustered by country are reported in round brackets. ***, **, and * indicate significance at the 1-, 5-, and 10-% levels, respectively.

A3.1.4 Old World

We test whether our results can be extended to all those regions where malaria was present before 1500. We thus expanded the sample to 9,566 cells in Africa, Europe, and Asia (so-called Old World). We reconstructed all the available controls and replicated the analysis of Table 1. We detect a strong and significant positive effect of Malaria Stability and the number of ethnic groups in the Old World sample.

Table A4: **Extended Sample: Old World**

	Ln (Number of Groups - GREG)				
	(1)	(2)	(3)	(4)	(5)
Malaria Stability	0.013***	0.014***	0.013***	0.011**	0.010**
	(0.003)	(0.005)	(0.004)	(0.004)	(0.004)
	[0.150]	[0.169]	[0.151]	[0.126]	[0.117]
Geographic Controls:					
Ln(Cell Area)	Yes	Yes	Yes	Yes	Yes
Soil Suitability and Elevation (Mean and Std.)	No	No	Yes	Yes	Yes
Avg. Temperature and Precipitation	No	No	Yes	Yes	Yes
Ruggedness and Caloric Suitability Pre 1500	No	No	Yes	Yes	Yes
Location Controls:					
Distances (equator, coast, river, border, capital)	No	No	No	Yes	Yes
Number of Countries and Within Country	No	No	No	Yes	Yes
TseTse suit. & Pr. Genetic Distance	No	No	No	No	Yes
Country FE	No	Yes	Yes	Yes	Yes
Observations	9,566	9,566	9,566	9,566	9,566
R-squared	.0623	.251	.293	.364	.366

Notes: The Table reports the OLS specification estimates associating the log number of ethnic groups with the level of Malaria Stability in Africa, Europe, and Asia. In all specifications, the dependent variable is the natural logarithm of the number of ethnic groups in the cell; Malaria Stability is the average level of malaria suitability in the cell; TseTse Suitability is the average level of the suitability for the vector of trypanosomiasis disease in the cell; see text for details. The geographic variables include the natural logarithm of the cell land area, soil suitability and elevation (mean and standard deviation), average temperature and precipitation, terrain ruggedness and caloric suitability before 1500. The distance variables include distance from the equator, from the coast, from the river, from the country border and from the country capital. Number of countries is the number of countries whose land falls within the cell. The within-country is a dummy variable taking value 1 if the cell is fully contained in a country. TseTse suitability measures the geographic suitability for Trypanosomiasis transmission. Predicted Genetic distance is computed as the log of migratory distance from East Africa. Variable description, data sources and summary statistics are reported in Tables E1, E2, and S2, respectively. Beta coefficient in square brackets. Robust standard errors clustered by country are reported in round brackets. ***, **, and * indicate significance at the 1-, 5-, and 10-% levels, respectively.

A3.1.5 Conditioning on Night lights and Population density today

We also estimate a specification including contemporary *proxies* of development, like light density at night and log of population density in 2000. We trade the inclusion of potentially bad controls with the possibility of controlling for levels of development. Table A5 replicates Table 1 including the development proxy in all columns. In spite of the reduction in coefficient magnitude, the effect remains always statistically significant and sizable.

Table A5: **Cross-cell Analysis: Endogenous Controls**

	Ln (Number of Groups - GREG)				
	(1)	(2)	(3)	(4)	(5)
Malaria Stability	0.016***	0.017***	0.016***	0.014***	0.014***
	(0.002)	(0.003)	(0.003)	(0.003)	(0.004)
	[0.273]	[0.287]	[0.274]	[0.240]	[0.240]
Geographic Controls:					
Ln(Cell Area)	Yes	Yes	Yes	Yes	Yes
Soil Suitability and Elevation (Mean and Std.)	No	No	Yes	Yes	Yes
Avg. Temperature and Precipitation	No	No	Yes	Yes	Yes
Ruggedness and Caloric Suitability Pre 1500	No	No	Yes	Yes	Yes
Location Controls:					
Distances (equator, coast, river, border, capital)	No	No	No	Yes	Yes
Number of Countries and Within Country	No	No	No	Yes	Yes
TseTse suit. & Pr. Genetic Distance	No	No	No	No	Yes
Country FE	No	Yes	Yes	Yes	Yes
Endogenous Controls:					
Ln(Night Lights) and Ln(Population)	Yes	Yes	Yes	Yes	Yes
Observations	1,976	1,976	1,976	1,976	1,976
R-squared	.209	.31	.352	.379	.381

Notes: The Table reports the OLS specification estimates associating the log number of ethnic groups with the level of Malaria Stability, controlling for contemporary development proxy. In all specifications, the dependent variable is the natural logarithm of the number of ethnic groups in the cell; Malaria Stability is the average level of malaria suitability in the cell; see text for details. The geographic variables include the natural logarithm of the cell land area, soil suitability and elevation (mean and standard deviation), average temperature and precipitation, terrain ruggedness and caloric suitability before 1500. The distance variables include distance from the equator, from the coast, from the river, from the country border and from the country capital. Number of countries is the number of countries whose land falls within the cell. The within-country is a dummy variable taking value 1 if the cell is fully contained in a country. TseTse suitability measures the geographic suitability for Trypanosomiasis transmission. Predicted Genetic distance is computed as the log of migratory distance from East Africa. The endogenous controls are the natural logarithm (0.01 + average luminosity) in the cell and log of population density in 2000 (CESIN). Variable description, data sources and summary statistics are reported in Tables E1, E2, and S1, respectively. Beta coefficient in square brackets. Robust standard errors clustered by country are reported in round brackets. ***, **, and * indicate significance at the 1-, 5-, and 10-% levels, respectively.

A3.2 Alternative Measures of Predicted Malaria: ITT and IV

We check whether our findings are robust to alternative climatic-based measures of malaria exposure. The dependent variable in all specification is the log number of ethnic groups in the cell. In all specification we control for geographic and location controls, and country fixed effects. Column (1) is equivalent to column (5) in Table 1. Column (2) presents the OLS estimate of the temperature-based Malaria Stability. The effect is precisely estimated and the temperature-based proxy is positively related to the number of groups. In column (3), we perform a 2SLS estimation using the temperature-based proxy as an instrument for Malaria Stability. The magnitude of the coefficient is almost identical to that of column (1). A very same pattern emerges in column (5) and (6) where we introduce the Plasmodium Falciparum Suitability. The fact that the difference between the OLS and IV estimates is small seems to suggest that the potential endogeneity problem in the Malaria Stability index is not so severe. Moreover, results from a Hausman test do not allow us to reject the hypothesis of both the OLS and IV estimators being consistent.

Table A6: **Predicted Malaria Stability: Alternative Measures.**

Measure of Malaria:	Ln (Number of Groups - GREG)				
	Baseline	Malaria Stability Temp.		Falciparum Suitability	
		ITT	IV	ITT	IV
	(1)	(2)	(3)	(4)	(5)
Malaria	0.015***	0.093**	0.016***	0.006**	0.032**
	(0.004)	(0.035)	(0.006)	(0.003)	(0.013)
	[0.257]	[0.156]	[0.276]	[0.186]	[0.547]
Geographic Controls:					
Ln(Cell Area)	Yes	Yes	Yes	Yes	Yes
Soil Suitability and Elevation (Mean and Std.)	Yes	Yes	Yes	Yes	Yes
Avg. Temp. and Prec.	Yes	Yes	Yes	Yes	Yes
Ruggedness and Caloric Suit. Pre 1500	Yes	Yes	Yes	Yes	Yes
Location Controls:					
Distances	Yes	Yes	Yes	Yes	Yes
Number of Countries and Within Country	Yes	Yes	Yes	Yes	Yes
TseTse suit. & Pr. Genetic Distance	Yes	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes	Yes
Observations	1,976	1,971	1,971	1,976	1,976
R-squared	.379	.366	.378	.367	.358
F-Stat			188		18.6

Notes: The Table reports the OLS and 2SLS specification estimates associating the log number of ethnic groups with the level of Malaria Stability and temperature-based measures of malaria stability/suitability in Africa. The geographic variables include the natural logarithm of the cell land area, soil suitability and elevation (mean and standard deviation), average temperature and precipitation, terrain ruggedness and caloric suitability before 1500. The distance variables include distance from the equator, from the coast, from the river, from the country border and from the country capital. Number of countries is the number of countries whose land falls within the cell. The within-country is a dummy variable taking value 1 if the cell is fully contained in a country. TseTse suitability measures the geographic suitability for Trypanosomiasis transmission. Predicted Genetic distance is computed as the log of migratory distance from East Africa. Variable description, data sources and summary statistics are reported in Tables E1, E2, and S1, respectively. Beta coefficient in square brackets. Robust standard errors clustered by country are reported in round brackets. ***, **, and * indicate significance at the 1-, 5-, and 10-% levels, respectively.

A3.3 Measures of Historical Malaria Incidence in the Population

We verify our baseline findings using two alternative proxies of historical malaria exposure. First, we exploit information on levels of people parasitization at the beginning of the twentieth century from Lysenko and Semashko (1968). The advantage of this population-based measure of malaria is to be informative on malaria prevalence in the African population at an early phase of the process of European colonization. The main limitation, from the perspective of disaggregate analysis, is that the data have a lower spatial resolution.

Second, we use the frequency of genetic immunities to malaria in terms of the share of individuals with the Duffy negative phenotype. The results in Table A7 replicate the baseline analysis of Table 1 with the alternative measures of long-term exposure to malaria. The two panels report the results for malaria endemicity in 1900 (Panel A) and Duffy Antigen (Panel B), respectively. These results confirm the baseline findings and the (non-standardized) point estimates are comparable also in magnitude. A one cross-cell standard deviation increase in malaria endemicity and genetic immunities to malaria decrease the standard deviation in the size of historical ethnic groups by a 0.26 and a 0.33 standard deviation, respectively. The magnitude of the coefficients tends to decrease more sharply and to lose significance with the inclusion of geographic controls.

Table A7: **Cross-cell Analysis: Malaria Endemicity & Duffy Antigen**

	Ln (Number of Groups - GREG)				
	(1)	(2)	(3)	(4)	(5)
Malaria Endemicity	0.120***	0.119***	0.063**	0.052*	0.054**
	(0.022)	(0.033)	(0.028)	(0.028)	(0.026)
	[0.353]	[0.348]	[0.185]	[0.154]	[0.159]
Geographic Controls:					
Ln(Cell Area)	Yes	Yes	Yes	Yes	Yes
Soil Suitability and Elevation (Mean and Std.)	No	No	Yes	Yes	Yes
Avg. Temperature and Precipitation	No	No	Yes	Yes	Yes
Ruggedness and Caloric Suitability Pre 1500	No	No	Yes	Yes	Yes
Location Controls:					
Distances (equator, coast, river, border, capital)	No	No	No	Yes	Yes
Number of Countries and Within Country	No	No	No	Yes	Yes
TseTse suit. & Pr. Genetic Distance	No	No	No	No	Yes
Country FE	No	Yes	Yes	Yes	Yes
Observations	1,976	1,976	1,976	1,976	1,976
R-squared	.165	.284	.335	.363	.367

	Ln (Number of Groups - GREG)				
	(1)	(2)	(3)	(4)	(5)
Duffy Antigen	0.607***	0.742**	0.209	0.264	0.220
	(0.155)	(0.292)	(0.171)	(0.221)	(0.212)
	[0.267]	[0.326]	[0.092]	[0.116]	[0.097]
Geographic Controls:					
Ln(Cell Area)	Yes	Yes	Yes	Yes	Yes
Soil Suitability and Elevation (Mean and Std.)	No	No	Yes	Yes	Yes
Avg. Temperature and Precipitation	No	No	Yes	Yes	Yes
Ruggedness and Caloric Suitability Pre 1500	No	No	Yes	Yes	Yes
Location Controls:					
Distances (equator, coast, river, border, capital)	No	No	No	Yes	Yes
Number of Countries and Within Country	No	No	No	Yes	Yes
TseTse suit. & Pr. Genetic Distance	No	No	No	No	Yes
Country FE	No	Yes	Yes	Yes	Yes
Observations	1,976	1,976	1,976	1,976	1,976
R-squared	.118	.27	.329	.36	.364

Notes: The Table reports the OLS specification estimates associating the log number of ethnic groups with the level of Malaria endemicity in 1900s (Upper Panel), and the level of Duffy Antigen in the population (Bottom Panel). In all specification, the dependent variable is the natural logarithm of the number of ethnic groups in the cell; Malaria Endemicity is the parasite rate in the cell measured at the beginning of the twentieth century; see text for details. The geographic variables include the natural logarithm of the cell land area, soil suitability and elevation (mean and standard deviation), average temperature and precipitation, terrain ruggedness and caloric suitability before 1500. The distance variables include distance from the equator, from the coast, from the river, from the country border and from the country capital. Number of countries is the number of countries whose land falls within the cell. The within-country is a dummy variable taking value 1 if the cell is fully contained in a country. TseTse suitability measures the geographic suitability for Trypanosomiasis transmission. Predicted Genetic distance is computed as the log of migratory distance from East Africa. Variable description, data sources and summary statistics are reported in Tables E1, E2, and S1, respectively. Beta coefficient in square brackets. Robust standard errors clustered by country are reported in round brackets. ***, **, and * indicate significance at the 1-, 5-, and 10-% levels, respectively.

A3.4 Multi-Host Vector Transmitted Pathogens: Placebo Diseases

As a placebo for the predicted channel, we consider the role of other vector-borne multi-host diseases on the log number of ethnic groups. We look at the effect of exposure to trypanosomiasis (transmitted by the *TseTse* fly) which, as shown by Alsan (2015), shaped the productive and social organization of the different ethnic groups in Africa and exposure to Dengue and Yellow fever (transmitted by Aedes mosquitoes).

We predict the intensity of transmission of Trypanosomiasis using a measure of suitability to *TseTse* fly produced by the Animal Health and Production Division and DFID - Animal Health Programme by Environmental Research Group Oxford (ERGO Ltd) in collaboration with the Trypanosomiasis and Land Use in Africa (TALA) research group at the Department of Zoology, University of Oxford. As a proxy of transmission of Dengue, we use an index predicting suitability to *Aedes aegypti*, the principal mosquito vector for Dengue fever, using data produced by Kraemer *et al.* (2015). Note that *Aedes aegypti* is also a vector for yellow fever and it is therefore related to both diseases. Figure D7 reports the binned scatter plots between Malaria Stability and Trypanosomiasis suitability (on the left) Malaria Stability and Dengue suitability (on the left), after controlling for geographic and location controls, and country fixed effects. Interestingly, once controlling for geography, the relationship between Malaria Stability and Trypanosomiasis suitability turns negative. On the contrary, the relationship between Malaria Stability and Dengue suitability is consistently positive and tight.

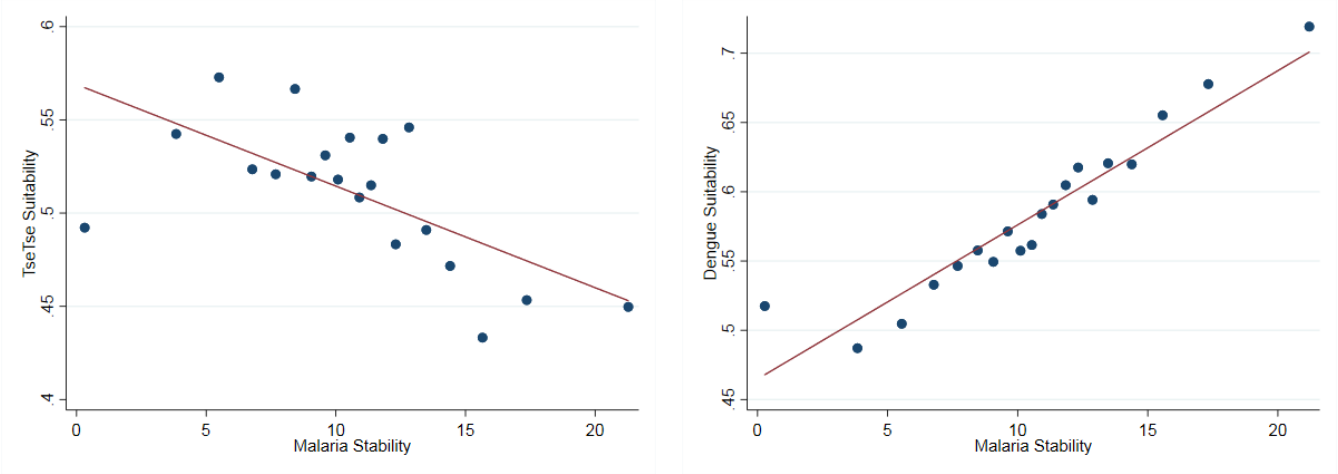


Figure D7: **Bin Scatter - Malaria Stability, Trypanosomiasis and Dengue Suitability**

The baseline empirical specification is therefore extended to the consideration of other diseases *vis-a-vis* the effect of predicted malaria exposure in Table A9 and to both in A10. Across the different specifications, exposure to Trypanosomiasis, Dengue and Yellow fever do not play a significant systematic role in explaining the log number of ethnic groups. Conversely, the effect of malaria is always positive and significant and little affected in terms of magnitude of the point estimates.

Table A8: **Placebo Vector-Borne Disease: Trypanosomiasis**

	Ln (Number of Groups - GREG)				
	(1)	(2)	(3)	(4)	(5)
Malaria	0.017***	0.020***	0.017***	0.015***	0.015***
	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)
	[0.298]	[0.339]	[0.287]	[0.256]	[0.257]
Trypanosomiasis	0.112**	0.029	-0.070	-0.096	-0.093
	(0.047)	(0.090)	(0.072)	(0.066)	(0.062)
	[0.134]	[0.035]	[-0.084]	[-0.115]	[-0.112]
Geographic Controls:					
Ln(Cell Area)	Yes	Yes	Yes	Yes	Yes
Soil Suitability and Elevation (Mean and Std.)	No	No	Yes	Yes	Yes
Avg. Temperature and Precipitation	No	No	Yes	Yes	Yes
Ruggedness and Caloric Suitability Pre 1500	No	No	Yes	Yes	Yes
Location Controls:					
Distances (equator, coast, river, border, capital)	No	No	No	Yes	Yes
Number of Countries and Within Country	No	No	No	Yes	Yes
Pr. Genetic Distance	No	No	No	No	Yes
Country FE	No	Yes	Yes	Yes	Yes
Observations	1,976	1,976	1,976	1,976	1,976
R-squared	.184	.299	.352	.379	.379

Notes: The Table reports the OLS specification estimates associating the log number of ethnic groups with the level of Malaria Stability and Trypanosomiasis in Africa. In all specifications, the dependent variable is the natural logarithm of the number of ethnic groups in the cell; Malaria Stability is the average level of malaria suitability in the cell; Trypanosomiasis measures the average level of suitability for the TseTse fly vector in the cell; see text for details. The geographic variables include the natural logarithm of the cell land area, soil suitability and elevation (mean and standard deviation), average temperature and precipitation, terrain ruggedness and caloric suitability before 1500. The distance variables include distance from the equator, from the coast, from the river, from the country border and from the country capital. Number of countries is the number of countries whose land falls within the cell. The within-country is a dummy variable taking value 1 if the cell is fully contained in a country. Variable description, data sources and summary statistics are reported in Tables E1, E2, and S1, respectively. Beta coefficient in square brackets. Robust standard errors clustered by country are reported in round brackets. ***, **, and * indicate significance at the 1-, 5-, and 10-% levels, respectively.

Table A9: **Placebo Vector-Borne Disease: Dengue/Yellow Fever**

	Ln (Number of Groups - GREG)				
	(1)	(2)	(3)	(4)	(5)
Malaria	0.017***	0.017***	0.014***	0.013***	0.013***
	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)
	[0.298]	[0.287]	[0.243]	[0.222]	[0.222]
Dengue/Yellow Fever	0.166	0.149	0.207*	0.217	0.219
	(0.123)	(0.093)	(0.123)	(0.164)	(0.162)
	[0.087]	[0.078]	[0.108]	[0.113]	[0.114]
Geographic Controls:					
Ln(Cell Area)	Yes	Yes	Yes	Yes	Yes
Soil Suitability and Elevation (Mean and Std.)	No	No	Yes	Yes	Yes
Avg. Temperature and Precipitation	No	No	Yes	Yes	Yes
Ruggedness and Caloric Suitability Pre 1500	No	No	Yes	Yes	Yes
Location Controls:					
Distances (equator, coast, river, border, capital)	No	No	No	Yes	Yes
Number of Countries and Within Country	No	No	No	Yes	Yes
TseTse suit. & Pr. Genetic Distance	No	No	No	No	Yes
Country FE	No	Yes	Yes	Yes	Yes
Observations	1,976	1,976	1,976	1,976	1,976
R-squared	.173	.301	.353	.378	.379

Notes: The Table reports the OLS specification estimates associating the log number of ethnic groups with the level of Malaria Stability and Dengue/Yellow Fever Suitability in Africa. In all specifications, the dependent variable is the natural logarithm of the number of ethnic groups in the cell; Malaria Stability is the average level of malaria suitability in the cell; Dengue/Yellow Fever is the average level of suitability for the vector of dengue and yellow fever in the cell; see text for details. The geographic variables include the natural logarithm of the cell land area, soil suitability and elevation (mean and standard deviation), average temperature and precipitation, terrain ruggedness and caloric suitability before 1500. The distance variables include distance from the equator, from the coast, from the river, from the country border and from the country capital. Number of countries is the number of countries whose land falls within the cell. The within-country is a dummy variable taking value 1 if the cell is fully contained in a country. Predicted Genetic distance is computed as the log of migratory distance from East Africa. Variable description, data sources and summary statistics are reported in Tables E1, E2, and S1, respectively. Beta coefficient in square brackets. Robust standard errors clustered by country are reported in round brackets. ***, **, and * indicate significance at the 1-, 5-, and 10-% levels, respectively.

Table A10: **Placebo Vector-Borne Disease: Trypanosomiasis and Dengue/Yellow Fever**

	Ln (Number of Groups - GREG)				
	(1)	(2)	(3)	(4)	(5)
Malaria	0.015***	0.016***	0.014***	0.013***	0.013***
	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)
	[0.255]	[0.281]	[0.243]	[0.219]	[0.219]
Dengue/Yellow Fever	0.128	0.150	0.194	0.190	0.193
	(0.122)	(0.093)	(0.121)	(0.165)	(0.162)
	[0.067]	[0.078]	[0.101]	[0.099]	[0.100]
Trypanosomiasis	0.106**	0.029	-0.063	-0.085	-0.082
	(0.048)	(0.089)	(0.071)	(0.067)	(0.063)
	[0.128]	[0.035]	[-0.076]	[-0.103]	[-0.098]
Geographic Controls:					
Ln(Cell Area)	Yes	Yes	Yes	Yes	Yes
Soil Suitability and Elevation (Mean and Std.)	No	No	Yes	Yes	Yes
Avg. Temperature and Precipitation	No	No	Yes	Yes	Yes
Ruggedness and Caloric Suitability Pre 1500	No	No	Yes	Yes	Yes
Location Controls:					
Distances (equator, coast, river, border, capital)	No	No	No	Yes	Yes
Number of Countries and Within Country	No	No	No	Yes	Yes
Pr. Genetic Distance	No	No	No	No	Yes
Country FE	No	Yes	Yes	Yes	Yes
Observations	1,976	1,976	1,976	1,976	1,976
R-squared	.186	.301	.354	.381	.381

Notes: The Table reports the OLS specification estimates associating the log number of ethnic groups with the level of Malaria Stability, Trypanosomiasis and Dengue/Yellow Fever in Africa. In all specifications, the dependent variable is the natural logarithm of the number of ethnic groups in the cell; Malaria Stability is the average level of malaria suitability in the cell; Trypanosomiasis measures the average level of suitability for the TseTse fly vector in the cell; Dengue/Yellow Fever is the average level of suitability for the vector of dengue and yellow fever in the cell; see text for details. The geographic variables include the natural logarithm of the cell land area, soil suitability and elevation (mean and standard deviation), average temperature and precipitation, terrain ruggedness and caloric suitability before 1500. The distance variables include distance from the equator, from the coast, from the river, from the country border and from the country capital. Number of countries is the number of countries whose land falls within the cell. The within-country is a dummy variable taking value 1 if the cell is fully contained in a country. TseTse suitability measures the geographic suitability for Trypanosomiasis transmission. Predicted Genetic distance is computed as the log of migratory distance from East Africa. Variable description, data sources and summary statistics are reported in Tables E1, E2, and S1, respectively. Beta coefficient in square brackets. Robust standard errors clustered by country are reported in round brackets. ***, **, and * indicate significance at the 1-, 5-, and 10-% levels, respectively.

A3.5 Pre-Colonial Characteristics

Our theory in Section 2 does not deliver specific predictions on the role of malaria for the organization of ethnic groups. Still, it is interesting to check the robustness of the baseline patterns explicitly accounting for, potentially omitted, pre-colonial characteristics of these ethnic groups and also exploring the potential role of malaria in shaping the organization of pre-colonial groups.

A3.5.1 Pre-colonial Characteristics - Murdock's Ethnographic Atlas

We extend the baseline specification to account for pre-colonial ethnographic characteristics from the Ethnographic Atlas (1967).[8] This anthropological database contains detailed information for 534 ethnic groups in Africa. Exploiting the distribution of group as mapped by the Murdock (1957) map of Africa, we create a set of variables measuring the average pre-colonial subsistence patterns, settlement characteristics and institutional and cultural features in the cell.[9] The list of variables employed in the analysis is the following. Subsistence variables include: Gathering, Hunting, Fishing, Animal Husbandry, Agricultural Dependence, Agricultural Types, and Milking; settlement pattern variables include: Population Density, Settlement complexity, Political Complexity at local level; institutional and cultural characteristics include: Polygyny, Clan Communities, Slavery, Property right, and Political Complexity beyond the local level.

Pre-Colonial Characteristics as Covariates. Table A11 extends the baseline results including pre-colonial characteristics. Each of the columns reports the point estimate for Malaria Stability, controlling for each of the Murdock pre-colonial variables listed on the heading of the respective columns. The specification includes the (log) area of the cell as well country fixed effect and is thereby comparable to the baseline within country estimates reported in Table 1 Column (2). The samples of different columns differ slightly since the ethnographic characteristics are missing for some of the cells. In spite of the changing size of the sample across different specifications due to limited data availability, the effect of the exposure to malaria is very stable to the inclusion of each pre-colonial co-variate in terms of both magnitude and statistical significance.

Pre-Colonial Characteristics as Dependent Variables. Table A12 explores the role of malaria for pre-colonial characteristics as dependent variables. With the exception of Fishing and Animal Husbandry, Malaria Stability displays no systematic patterns of statistical and economic significance on these pre-colonial characteristics.

[8] We extract data from the digitized and corrected version by Gray (1999).

[9] Since more than one ethnic homeland might be contained in a single cell, we average out the value of each ethnographic characteristic, if available, across the ethnic homelands existing in each cell.

Table A11: **Pre-Colonial Controls**

	Ln (Number of Groups - GREG)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Malaria Stability	0.021***	0.021***	0.021***	0.020***	0.019***	0.022***	0.022***
	(0.004)	(0.004)	(0.004)	(0.003)	(0.004)	(0.004)	(0.004)
	[0.369]	[0.374]	[0.375]	[0.357]	[0.327]	[0.388]	[0.396]
Pre-Colonial Control	Gathering	Hunting	Fishing	Animal Husb.	Agri. Dep.	Agri. Type	Milking
Ln(Cell Area)	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,703	1,703	1,703	1,703	1,703	1,648	1,648
R-squared	.303	.297	.295	.296	.306	.294	.296

	Ln (Number of Groups - GREG)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Malaria Stability	0.017***	0.021***	0.021***	0.020***	0.022***	0.020***	0.022***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
	[0.310]	[0.371]	[0.374]	[0.354]	[0.386]	[0.334]	[0.397]
Pre-Colonial Control	Pop. Dens.	Complex Settl.	Polyginy	Clans	Slavery	Property Rights	Jurisd. Hierarchy
Ln(Cell Area)	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,304	1,648	1,687	1,565	1,670	820	1,624
R-squared	.306	.298	.298	.287	.309	.423	.309

Notes: The Table reports the OLS specification estimates associating the log number of ethnic groups with the level of Malaria Stability, controlling for several socio-economics pre-colonial controls from the Ethnographic Atlas. In all specification, the dependent variable is the natural logarithm of the number of ethnic groups in the cell; Malaria Stability is the average level of malaria suitability in the cell; see text for details. Each column includes the corresponding Murdock variable (Murdock, 1967; Gray, 1999). All regressions include the natural logarithm of cell land area and country fixed effects. Variable description, data sources and summary statistics are reported in Tables E1, E2, E3, S1, and S3, respectively. Beta coefficient in square brackets. Robust standard errors clustered by country are in parentheses. ***, **, and * indicate significance at the 1-, 5-, and 10-% levels, respectively.

Table A12: **Pre-Colonial Characteristics as Outcomes**

	Gathering (1)	Hunting (2)	Fishing (3)	Animal Husb. (4)	Agri. Dep. (5)	Agri. Type (6)	Milking (7)
Malaria Stability	0.010*	-0.006**	0.013***	-0.049*	0.008	-0.018	-0.004
	(0.005)	(0.003)	(0.004)	(0.027)	(0.021)	(0.013)	(0.005)
	[0.220]	[-0.150]	[0.290]	[-0.201]	[0.036]	[-0.185]	[-0.084]
Ln(Cell Area)	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,703	1,703	1,703	1,703	1,703	1,648	1,648
R-squared	.358	.448	.483	.708	.706	.606	.712

		Complex Settl. (2)	Polygyny (3)	Clans (4)	Slavery (5)	Property Rights (6)	Jurisd. Hierarchy (7)
Malaria Stability		0.005	0.008	0.000	-0.001	-0.004	0.006
		(0.005)	(0.005)	(0.004)	(0.002)	(0.003)	(0.011)
		[0.110]	[0.331]	[0.003]	[-0.030]	[-0.083]	[0.061]
Ln(Cell Area)		Yes	Yes	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations		1,648	1,687	1,565	1,670	820	1,624
R-squared		.508	.405	.375	.564	.678	.411

Notes: The Table reports the OLS specification estimates associating several socio-economics pre-colonial variables from the Ethnographic Atlas with the level of Malaria Stability. Malaria Stability is the average level of the malaria suitability in the cell; see text for details. All regressions include the natural logarithm of cell land area and country fixed effects. Variable description, data sources and summary statistics are reported in Tables E1, E2, E3, S1, and S3, respectively. Beta coefficient in square brackets. Robust standard errors clustered by country are in parentheses. ***, **, and * indicate significance at the 1-, 5-, and 10-% levels, respectively.

A3.5.2 Pre-colonial Population Density and Malaria Stability

We also empirically investigated the relationship between Malaria Stability and pre-colonial population density. According to Section 2, the role of malaria on population density is theoretically ambiguous. Figure D8 plots the local polynomial of malaria stability on the log of pre-colonial population density. The fitting line appears to be quite flat. In Table B1 we perform a regression analysis, employing both the Malaria Stability index as well as its temperature-based proxies. In the unconstrained specifications of column (1), (3), and (5), the coefficient of malaria is mainly positive and significant. The coefficient flips sign once geographical, climatic, and location controls are taken into account in column (4) and (6). Moreover, the coefficient becomes statistically indistinguishable from zero once we controlled for our baseline set of controls.

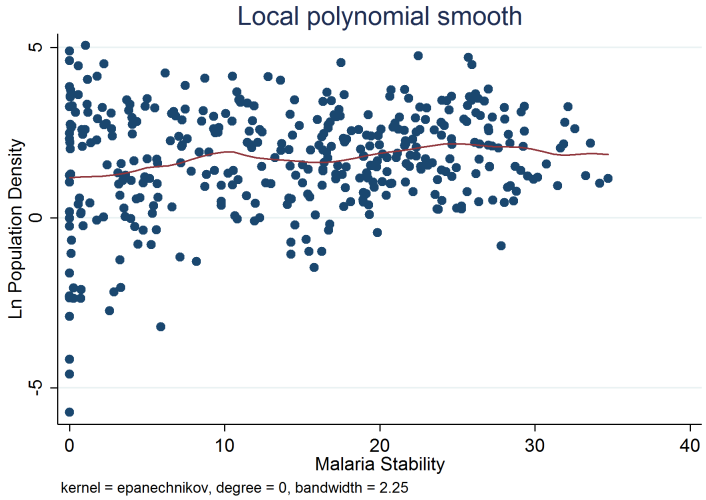


Figure D8: **Local Polynomial Pre-Colonial Pop Density and Malaria Stability**

Table A13: **Malaria Stability and Pre-Colonial Population Density**

	Ln (Pre-Colonial Population Density)					
	Baseline		Malaria Stability Temp.		Falciparum Suitability	
	(1)	(2)	(3)	(4)	(5)	(6)
Malaria Stability	0.088***	0.035***				
	(0.026)	(0.011)				
	[0.422]	[0.167]				
Malaria Stability (Temperature)			0.674**	-0.258		
			(0.302)	(0.180)		
			[0.305]	[-0.117]		
Malaria Transmission (Temperature)					0.051***	-0.011
					(0.019)	(0.009)
					[0.419]	[-0.093]
Geographic Controls:						
Ln(Cell Area)	No	Yes	No	Yes	No	Yes
Soil Suitability and Elevation (Mean and Std.)	No	Yes	No	Yes	No	Yes
Avg. Temperature and Precipitation	No	Yes	No	Yes	No	Yes
Ruggedness and Caloric Suitability Pre 1500	No	Yes	No	Yes	No	Yes
Location Controls:						
Distances (equator, coast, river)	No	Yes	No	Yes	No	Yes
TseTse suit. & Pr. Genetic Distance	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,304	1,304	1,302	1,302	1,304	1,304
R-squared	.181	.755	.106	.751	.187	.75

Notes: The Table reports the OLS specification estimates associating the log of pre-colonial population density with the level of Malaria Stability and its temperature-based proxies. In all specifications, the dependent variable is the natural logarithm of the pre-colonial population density in the cell (Alsan, 2014); Malaria Stability is the average level of malaria suitability in the cell; Malaria Stability (Temperature) replicates the Malaria Stability index where mosquitoes characteristics are a function of temperature; Malaria Transmission (Temperature) is an index predicting the intensity of malaria transmission based on temperature; see text for details. The geographic variables include the natural logarithm of the cell land area, soil suitability and elevation (mean and standard deviation), average temperature and precipitation, terrain ruggedness and caloric suitability before 1500. The distance variables include distance from the equator, from the coast, from the river, from the country border and from the country capital. Number of countries is the number of countries whose land falls within the cell. The within-country is a dummy variable taking value 1 if the cell is fully contained in a country. TseTse suitability measures the geographic suitability for Trypanosomiasis transmission. Predicted Genetic distance is computed as the log of migratory distance from East Africa. Variable description, data sources and summary statistics are reported in Tables E1, E2, and S1, respectively. Beta coefficient in square brackets. Robust standard errors clustered by country are reported in round brackets. ***, **, and * indicate significance at the 1-, 5-, and 10-% levels, respectively.

A4 Pre-Colonial Ethicities Africa and the Americas (Placebo): Further Results and Robustness

A4.1 Alternative Specification - Africa

Table B1 looks at the effect of malaria stability on the log of the number of groups in Africa using Murdock data, including additional controls not available for South America, such as the TseTse suitability. Tables B2, B3, and B4 look at the relationship between pre-colonial ethnic diversity and other vector diseases, such as Trypanosomiasis and Dengue/Yellow Fever. As for historical diversity, Trypanosomiasis, Dengue and Yellow fever do not play a significant systematic role in explaining the log number of pre-colonial ethnic groups.

Table B1: **Pre-Colonial Ethnicities: Murdock Africa**

	Ln (Number of Groups - Murdock)			
	(1)	(2)	(3)	(4)
Malaria Stability	0.021***	0.014***	0.013***	0.013***
	(0.001)	(0.002)	(0.002)	(0.002)
	[0.357]	[0.235]	[0.216]	[0.214]
Geographic Controls:				
Ln(Cell Area)	Yes	Yes	Yes	Yes
Soil Suitability and Elevation (Mean and Std.)	No	Yes	Yes	Yes
Avg. Temperature and Precipitation	No	Yes	Yes	Yes
Ruggedness and Caloric Suitability Pre 1500	No	Yes	Yes	Yes
Location Controls:				
Distances (equator, coast, river)	No	No	Yes	Yes
TseTse suit. & Pr. Genetic Distance	No	No	No	Yes
Observations	1,976	1,976	1,976	1,976
R-squared	.155	.286	.292	.297

Notes: The Table reports the OLS specification estimates associating the number of pre-colonial ethnic groups with the level of Malaria Stability in Africa. The dependent variable is the natural logarithm of the number of pre-colonial groups (Murdock 1959) in the cell. All dependent variables are constructed using the Murdock maps for Africa. Malaria Stability is the average level of the malaria suitability in the cell. The geographic variables include the natural logarithm of the cell land area, soil suitability and elevation (mean and standard deviation), average temperature and precipitation, terrain ruggedness and caloric suitability before 1500. The distance variables include distance from the equator, from the coast, from the river, from the country border and from the country capital. Number of countries is the number of countries whose land falls within the cell. The within-country is a dummy variable taking value 1 if the cell is fully contained in a country. TseTse suitability measures the geographic suitability for Trypanosomiasis transmission. Predicted Genetic distance is computed as the log of migratory distance from East Africa. Variable description, data sources and summary statistics are reported in Tables E1, E2, E3 and S4, and S1, respectively. Beta coefficient in square bracket. Robust standard errors clustered by country are in parentheses. ***, **, and * indicate significance at the 1-, 5-, and 10-% levels, respectively.

Table B2: **Pre-Colonial Ethnicities: Murdock Africa - Trypanosomiasis**

	Ln (Number of Groups - GREG)			
	(1)	(2)	(3)	(4)
Malaria	0.016***	0.014***	0.013***	0.013***
	(0.001)	(0.002)	(0.002)	(0.002)
	[0.279]	[0.235]	[0.215]	[0.214]
Trypanosomiasis	0.147***	-0.037	-0.040	-0.062**
	(0.020)	(0.030)	(0.031)	(0.032)
	[0.172]	[-0.043]	[-0.047]	[-0.073]
Geographic Controls:				
Ln(Cell Area)	Yes	Yes	Yes	Yes
Soil Suitability and Elevation (Mean and Std.)	No	Yes	Yes	Yes
Avg. Temperature and Precipitation	No	Yes	Yes	Yes
Ruggedness and Caloric Suitability Pre 1500	No	Yes	Yes	Yes
Location Controls:				
Distances (equator, coast, river)	No	No	Yes	Yes
Pr. Genetic Distance	No	No	No	Yes
Observations	1,976	1,976	1,976	1,976
R-squared	.178	.287	.293	.297

Notes: The Table reports the OLS specification estimates associating the number of pre-colonial ethnic groups with the level of Malaria Stability in Africa. The dependent variable is the natural logarithm of the number of pre-colonial groups (Murdock 1959) in the cell. All dependent variables are constructed using the Murdock maps for Africa. Malaria Stability is the average level of the malaria suitability in the cell. Trypanosomiasis measures the average level of suitability for the TseTse fly vector in the cell. The geographic variables include the natural logarithm of the cell land area, soil suitability and elevation (mean and standard deviation), average temperature and precipitation, terrain ruggedness and caloric suitability before 1500. The distance variables include distance from the equator, from the coast, from the river, from the country border and from the country capital. Number of countries is the number of countries whose land falls within the cell. The within-country is a dummy variable taking value 1 if the cell is fully contained in a country. Predicted Genetic distance is computed as the log of migratory distance from East Africa. Variable description, data sources and summary statistics are reported in Tables E1, E2, E3, and S1, respectively. Beta coefficient in square bracket. Robust standard errors clustered by country are in parentheses. ***, **, and * indicate significance at the 1-, 5-, and 10-% levels, respectively.

Table B3: **Pre-Colonial Ethnicities: Murdock Africa - Dengue and Yellow Fever**

	Ln (Number of Groups - GREG)			
	(1)	(2)	(3)	(4)
Malaria Stability	0.014***	0.013***	0.012***	0.011***
	(0.002)	(0.002)	(0.002)	(0.002)
	[0.235]	[0.219]	[0.196]	[0.193]
Dengue/Yellow Fever	0.334***	0.069	0.090	0.099
	(0.056)	(0.067)	(0.070)	(0.071)
	[0.169]	[0.035]	[0.046]	[0.050]
Geographic Controls:				
Ln(Cell Area)	Yes	Yes	Yes	Yes
Soil Suitability and Elevation (Mean and Std.)	No	Yes	Yes	Yes
Avg. Temperature and Precipitation	No	Yes	Yes	Yes
Ruggedness and Caloric Suitability Pre 1500	No	Yes	Yes	Yes
Location Controls:				
Distances (equator, coast, river)	No	No	Yes	Yes
TseTse suit. & Pr. Genetic Distance	No	No	No	Yes
Observations	1,976	1,976	1,976	1,976
R-squared	.169	.287	.293	.296

Notes: The Table reports the OLS specification estimates associating the number of pre-colonial ethnic groups with the level of Malaria Stability in Africa. The dependent variable is the natural logarithm of the number of pre-colonial groups (Murdock 1959) in the cell. All dependent variables are constructed using the Murdock maps for Africa. Malaria Stability is the average level of the malaria suitability in the cell. Dengue/Yellow Fever is the average level of suitability for the vector of dengue and yellow fever in the cell. The geographic variables include the natural logarithm of the cell land area, soil suitability and elevation (mean and standard deviation), average temperature and precipitation, terrain ruggedness and caloric suitability before 1500. The distance variables include distance from the equator, from the coast, from the river, from the country border and from the country capital. Number of countries is the number of countries whose land falls within the cell. The within-country is a dummy variable taking value 1 if the cell is fully contained in a country. Predicted Genetic distance is computed as the log of migratory distance from East Africa. Variable description, data sources and summary statistics are reported in Tables E1, E2, E3, and S1, respectively. Beta coefficient in square bracket. Robust standard errors clustered by country are in parentheses. ***, **, and * indicate significance at the 1-, 5-, and 10-% levels, respectively.

Table B4: **Pre-Colonial Ethnicities: Murdock Africa - Trypanosomiasis, Dengue and Yellow Fever**

	Ln (Number of Groups - GREG)			
	(1)	(2)	(3)	(4)
Malaria	0.011***	0.013***	0.012***	0.012***
	(0.002)	(0.002)	(0.002)	(0.002)
	[0.181]	[0.222]	[0.199]	[0.197]
Trypanosomiasis	0.135***	-0.034	-0.035	-0.057*
	(0.020)	(0.030)	(0.031)	(0.032)
	[0.158]	[-0.040]	[-0.041]	[-0.067]
Dengue and Yellow Fever	0.284***	0.057	0.074	0.075
	(0.056)	(0.068)	(0.071)	(0.072)
	[0.144]	[0.029]	[0.038]	[0.038]
Geographic Controls:				
Ln(Cell Area)	Yes	Yes	Yes	Yes
Soil Suitability and Elevation (Mean and Std.)	No	Yes	Yes	Yes
Avg. Temperature and Precipitation	No	Yes	Yes	Yes
Ruggedness and Caloric Suitability Pre 1500	No	Yes	Yes	Yes
Location Controls:				
Distances (equator, coast, river)	No	No	Yes	Yes
Pr. Genetic Distance	No	No	No	Yes
Observations	1,976	1,976	1,976	1,976
R-squared	.188	.287	.293	.297

Notes: The Table reports the OLS specification estimates associating the number of pre-colonial ethnic groups with the level of Malaria Stability in Africa. The dependent variable is the natural logarithm of the number of pre-colonial groups (Murdock 1959) in the cell. All dependent variables are constructed using the Murdock maps for Africa. Malaria Stability is the average level of the malaria suitability in the cell. Trypanosomiasis measures the average level of suitability for the TseTse fly vector in the cell; Dengue/Yellow Fever is the average level of suitability for the vector of dengue and yellow fever in the cell. The geographic variables include the natural logarithm of the cell land area, soil suitability and elevation (mean and standard deviation), average temperature and precipitation, terrain ruggedness and caloric suitability before 1500. The distance variables include distance from the equator, from the coast, from the river, from the country border and from the country capital. Number of countries is the number of countries whose land falls within the cell. The within-country is a dummy variable taking value 1 if the cell is fully contained in a country. Predicted Genetic distance is computed as the log of migratory distance from East Africa. Variable description, data sources and summary statistics are reported in Tables E1, E2, E3, and S1, respectively. Beta coefficient in square bracket. Robust standard errors clustered by country are in parentheses. ***, **, and * indicate significance at the 1-, 5-, and 10-% levels, respectively.

A4.2 ITT and IV

In Table B5, we replicate the results of the placebo using temperature-based predicted measures of malaria suitability. While somehow less precisely estimated, results confirm a differential effect of malaria suitability in Africa, where the *plasmodium* of malaria was present, with respect to the Americas, where malaria was not present before 1500 AD.

Table B5: **Placebo - Temperature-Based Malaria Indexes**

	Ln (Number of Groups - Murdock's Data)					
	Africa			Americas		
	Baseline	Temp.	Suit.	Baseline	Temp.	Suit.
	(1)	(2)	(3)	(4)	(5)	(6)
Malaria Stability	0.013***			0.001		
	(0.002)			(0.010)		
	[0.216]			[0.004]		
Malaria Stability (Temperature)		0.048**			-0.000	
		(0.021)			(0.025)	
		[0.079]			[-0.001]	
Malaria Transmisssion (Temperature)			0.002			-0.001
			(0.001)			(0.002)
			[0.046]			[-0.059]
Ln(Cell Area)	Yes	Yes	Yes	Yes	Yes	Yes
Geographic Controls	Yes	Yes	Yes	Yes	Yes	Yes
Location Controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,973	1,968	1,973	1,503	1,487	1,503
R-squared	.292	.275	.275	.137	.134	.137

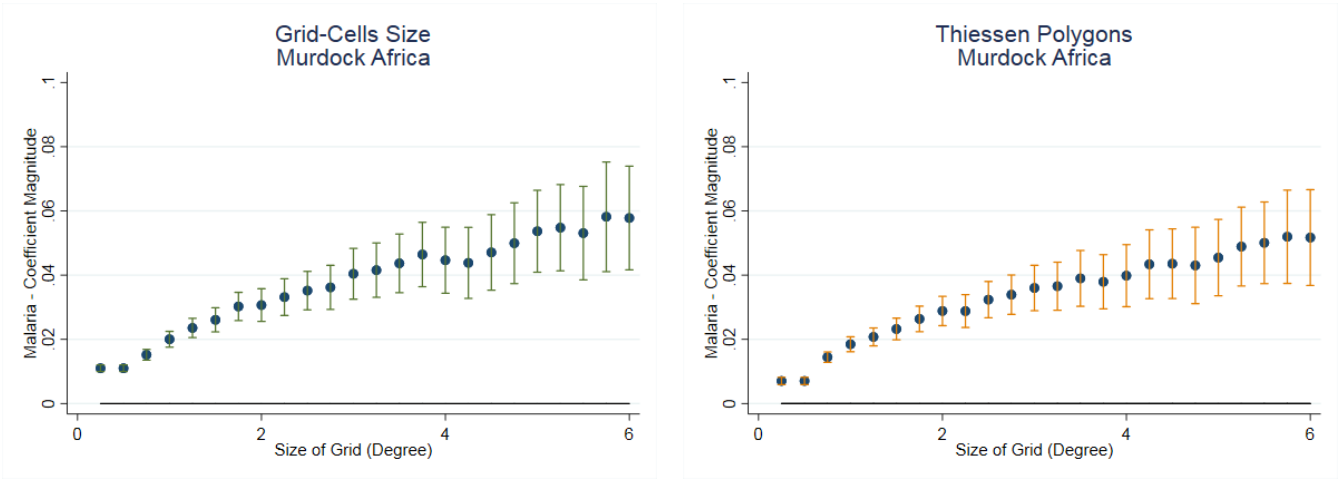
Notes: The Table reports the OLS specification estimates associating the number of pre-colonial ethnic groups with the level of Malaria Stability and temperature-based malaria indexes in Africa and the Americas. The dependent variable is the natural logarithm of the number of pre-colonial groups (Murdock 1951 and 1959) in the cell. All dependent variables are constructed using the Murdock maps for Africa and the Americas (see Appendix A2.2 for further details). Malaria Stability is the average level of malaria suitability in the cell; Malaria Stability (Temperature) replicates the Malaria Stability index where mosquitoes characteristics are a function of temperature; Falciparum Suitability is an index predicting the intensity of malaria transmission based on temperature; see text for details. The geographic variables include the natural logarithm of the cell land area, soil suitability and elevation (mean and standard deviation), average temperature and precipitation, terrain ruggedness and caloric suitability before 1500. The distance variables include distance from equator, from coast and from rivers. Variable description, data sources and summary statistics are reported in Tables E1, E2, E3, S1, and S4 respectively. Beta coefficient in square bracket. Robust standard errors clustered by country are in parentheses. ***, **, and * indicate significance at the 1-, 5-, and 10-% levels, respectively.

A4.3 Cell Sizes and Thiessen Polygons

We explore the role of varying the size of grid cells as units of observation, using Murdock data, both in Africa and the Americas. As in Figure 3 in the main body, we reconstruct the database and replicate the empirical results using a full range of alternative cells size as units of observation ranging from 0.25 degrees (about 28km x 28km) to 6 degrees (about 666km x 666km), in steps of 0.25 degrees. Two patterns emerge.

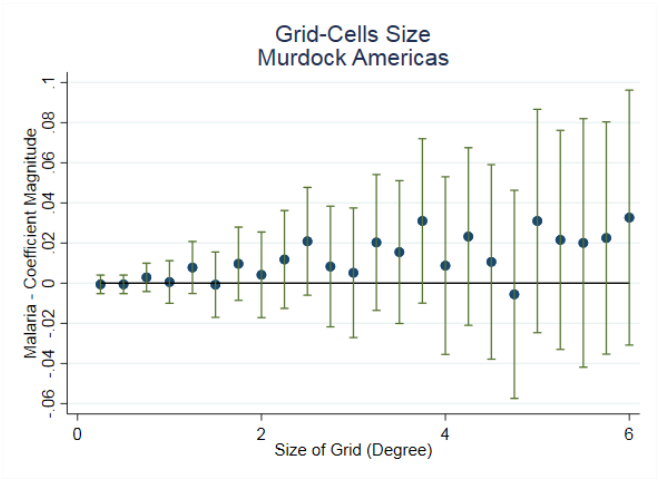
The association between malaria stability and the number of pre-colonial ethnic groups in Africa is confirmed across all spectrum of cell size considered. As Figure D9 Panel A shows, the baseline effects obtained with 1 degree cells are, if anything, conservative. In Panel B of Figure D9 we replicate the analysis using Thiessen polygon transformation of the original map to account for potential (non-random) errors in the drawing of borders of historical ethnicities. The effect of malaria is confirmed and very similar in absolute magnitude to that of Panel A.

In the Americas, the evidence is quite different. Firstly, the sign of the coefficients depend on the size of the cell, making the results bumpy. Secondly, the effect (positive and negative) are always indistinguishable from zero in terms of statistical significance. Finally, accounting for (non-random) errors in the drawing of borders does not change the broad picture. No clear and precise effect of malaria on the log number of ethnic groups can be detected in the Americas.

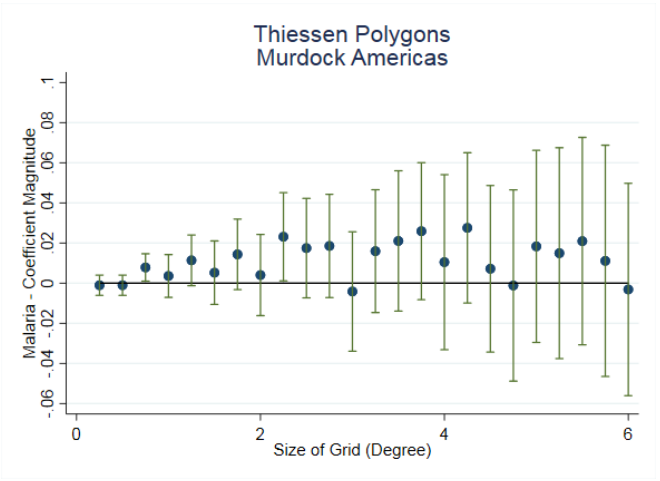


Panel A Panel B

Figure D9: **Grid-Cells Size and Thiessen Polygon Transformations - Murdock Africa**



Panel A



Panel B

Figure D10: **Grid-Cells Size and Thiessen Polygon Transformations - Murdock America**

A5 The Legacy of Malaria for Ethnic Diversity Today

This section studies the effect of long-term exposure to malaria on ethnic diversity today, looking at patterns of ethnic admixing in settlements, at the strength of ethnic identification with the group and at the modern distribution of ethnolinguistic groups.

A5.1 Spatial Clustering and Ethnic Admixing (Village level data)

Our hypothesis predicts that malaria exposure fosters the emergence of behavioral isolation in terms of in-group cultures that favor both strong interactions within groups and limited interactions across groups. Lack of systematic information on the patterns of differential interactions and the strength of ethnic identities prevent empirical tests on historical data. Under the assumption of (some) persistence, we can attempt to track some of these implications using contemporaneous data. In malarial areas, ethnic groups should be expected to display limited admixing in the same territory.

To test this prediction we construct measures of ethnic admixing at the local level. We use data from the Demographic and Health Survey (DHS), which samples households at the village level. We exploit all available waves, spanning 22 African countries, for which information on the ethnic group of the respondent is available.[10] We use information on the ethnic composition at the local level to retrieve the number of ethnic groups in 14,202 villages.

Table C1 and C2 summarize main results. Higher exposure to malaria reduces ethnic fractionalization and the number of ethnic groups at the cluster (village) level within countries. The pattern is confirmed when controlling for geographic characteristics of the village. Results also hold when conditioning for the level of economic development and population density and when accounting for the degree of ethnic fractionalization and the number of groups in the region. The magnitude of the coefficient implies that a standard deviation increase in malaria stability brings about a 0.11 decrease in the level of fractionalization of the village.

[10]The data provide detailed information on representative samples of women ranging in age from 15 to 49 years. Countries in the sample are Benin, Burkina Faso, Cameroon, Central African Republic, Ethiopia, Gabon, Ghana, Guinea, Ivory Coast, Kenya, Liberia, Malawi, Mali, Mozambique, Namibia, Niger, Nigeria, Senegal, Sierra Leone, Togo, Uganda, and Zambia.

Table C1: **Ethinc Admixing at Village Level: Number of Ethnic Groups**

Dependent Variable	Ln(Number of Ethnic Groups)				
	(1)	(2)	(3)	(4)	(5)
Malaria Stability	-0.011***	-0.007***	-0.008***	-0.008***	-0.008***
	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)
Geographic Controls:					
Soil Suitability and Elevation (Mean and Std.)	No	Yes	Yes	Yes	Yes
Avg. Temperature and Precipitation	No	Yes	Yes	Yes	Yes
Ruggedness and Caloric Suitability Pre 1500	No	Yes	Yes	Yes	Yes
Location Controls:					
Distances (equator, coast, river, border, capital)	No	Yes	Yes	Yes	Yes
TseTse suit. & Pr. Genetic Distance	No	Yes	Yes	Yes	Yes
Group Fractionalization in Region	No	No	Yes	Yes	Yes
Number of Groups in Region	No	No	No	Yes	Yes
Population and Night Lights	No	No	No	No	Yes
Country FE	Yes	Yes	Yes	Yes	Yes
Observations	13,179	13,179	13,179	13,179	13,179
R-squared	0.184	0.278	0.338	0.342	0.349

Notes: The Table reports the OLS specification estimates associating the number of ethnic group in a village with the level Malaria Stability. In all specification, the dependent variable is the log number of ethnic group in the village, based on DHS respondents; Malaria Stability measures the level of malaria suitability in a radius of 10 km around the centroid of the village. The geographic variables include the natural logarithm of the cell land area, soil suitability and elevation (mean and standard deviation), average temperature and precipitation, terrain ruggedness and caloric suitability before 1500. The distance variables include distance from the equator, from the coast, from the river, from the country border and from the country capital. TseTse suitability measures the geographic suitability for Trypanosomiasis transmission. Predicted Genetic distance is computed as the log of migratory distance from East Africa. Group fractionalization in the region measures the ethnic polarization at region level, Number of groups in the Region is the number of ethnic group in the region based on DHS respondents. Population and Night Lights measure population density and night luminosity. Variable description, data sources and summary statistics are reported in Tables E1, E2, and S5, respectively. Beta coefficient in square brackets. Robust standard errors clustered by country are reported in round brackets. ***, **, and * indicate significance at the 1-, 5-, and 10-% levels, respectively.

Table C2: **Ethnic Admixing at Village Level: ELF Index**

Dependent Variable	Ethnolinguistic Fractionalization in Village				
	(1)	(2)	(3)	(4)	(5)
Malaria Stability	-0.005***	-0.003**	-0.003***	-0.003***	-0.003***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Geographic Controls:					
Soil Suitability and Elevation (Mean and Std.)	No	Yes	Yes	Yes	Yes
Avg. Temperature and Precipitation	No	Yes	Yes	Yes	Yes
Ruggedness and Caloric Suitability Pre 1500	No	Yes	Yes	Yes	Yes
Location Controls:					
Distances (equator, coast, river, border, capital)	No	Yes	Yes	Yes	Yes
TseTse suit. & Pr. Genetic Distance	No	Yes	Yes	Yes	Yes
Group Fractionalization in Region	No	No	Yes	Yes	Yes
Number of Groups in Region	No	No	No	Yes	Yes
Population and Night Lights	No	No	No	No	Yes
Country FE	Yes	Yes	Yes	Yes	Yes
Observations	13,179	13,179	13,179	13,179	13,179
R-squared	0.108	0.223	0.293	0.294	0.304

Notes: The Table reports the OLS specification estimates associating ethno-liguistic fractonalization in a village with the level Malaria Stability. In all specification, the dependent variable is an index of ethnic fractionalization in the village, based on DHS respondents; Malaria Stability measures the level of malaria stability in a radius of 10 km around the centroid of the village. The geographic variables include the natural logarithm of the cell land area, soil suitability and elevation (mean and standard deviation), average temperature and precipitation, terrain ruggedness and caloric suitability before 1500. The distance variables include distance from the equator, from the coast, from the river, from the country border and from the country capital. TseTse suitability measures the geographic suitability for Trypanosomiasis transmission. Predicted Genetic distance is computed as the log of migratory distance from East Africa. Variable description, data sources and summary statistics are reported in Tables E1, E2, and S5, respectively. Beta coefficient in square brackets. Robust standard errors clustered by country are reported in round brackets. ***, **, and * indicate significance at the 1-, 5-, and 10-% levels, respectively.

A5.2 Ethnic Identification with own Group

A history of isolation and limited admixing may bring about a stronger contemporary ethnic identification. The conceptual framework provides a rationale for this since ethnic identities in malaria areas should be shaped in the context of the emergence of (and are instrumental to) the enforcement of cultures of geographic and behavioral isolation in each location. As a further qualification of the role of malaria on African ethnic diversity, we test whether malaria exposure strengthens the degree with which an individual identifies with his or her ethnic group.

To measure the intensity of identification with own ethnic group, we use the third wave of Afrobarometer which reports interviews conducted across 17 countries.[11] More precisely, the question attempts to measure ethnic identification versus national identification, it is therefore not ideal for our setting but still potentially informative.[12]

Table C3 summarizes the main findings. All specifications include country fixed effects to account for the specificity of the national states. Column (1) reports the unconditional (within country) results. Column (2) extends the specification to the inclusion of bio-climatological and geographic controls and location controls. Columns (3) further includes individual controls (living conditions, education, religion, occupation, rural or urban) while Column (4) also accounts for various potentially relevant, but also potentially endogenous, measures of socio-economic development (in terms of population and night lights) and proxies for the presence of different ethnic groups in terms of the respondents' group size (and share) in the region, the total number of ethnic groups, and the index of ethnic fractionalization in the region.

All specifications highlight a positive and statistically significant correlation between malaria stability and the strength of ethnic identification. Throughout specifications, the coefficient remains stable, precisely estimated and sizable in terms of magnitude.[13] The coefficient suggests that a

[11] Countries in the sample are Benin, Botswana, Ghana, Kenya, Lesotho, Madagascar, Malawi, Mali, Mozambique, Namibia, Nigeria, Senegal, South Africa, Tanzania, Uganda, Zambia, and Zimbabwe. In the sample lightly less than one-third of respondents identify only with their own country, while only around 5% identify only with their ethnic group. We assemble a dataset that associates with each respondent (or cluster) the (average) Malaria Stability and other geographic characteristics using a circle with a 10 km radius as baseline. The final sample contains 19,809 individuals.

[12] Take, for instance, the case of Kenya. The question reads "Let us suppose that you had to choose between being a Kenyan and being a [respondent's identity group]. Which of these two groups do you feel most strongly attached to?" This is a categorical variable with higher values associated with more ethnic group identification. This measure is not ideal to test our hypothesis which relates to the intensity of ethnic feelings, or even better ethnic enmities, between ethnic groups rather than between one own ethnic group and the national state. The "metric" of ethnic distances and the relationship between groups and the state may not be straightforward. For instance, in certain countries, the state can be controlled by a ruling (majority) group while in others the state nations encompass a set of related ethnicities. For these reasons, we think that, while potentially informative, these results should be interpreted as mainly suggestive.

[13] In the most extensive specification of Column (4) a one-standard deviation increase in Malaria Stability is associated with a 0.09 standard deviations increase in the strength of group identification. The (unconditional) effect is little affected by the inclusion of the covariates and standard tests suggest that the effect of malaria is unlikely to be driven by the unobserved individual, or location-specific, characteristics.

standard deviation increase in malaria stability lead to a 0.12 standard deviations increase in the strength of ethnic identification.

Table C3: **Ethnic Identification Today - Afrobarometer Data**

Dependent Variable	Identification with own Ethnic Group			
	(1)	(2)	(3)	(4)
Malaria Stability	0.022***	0.017***	0.016***	0.015***
Cluster s.e. Ethnologue Name	(0.004)	(0.005)	(0.004)	(0.004)
Cluster s.e. Ethnic Homeland	(0.003)	(0.004)	(0.003)	(0.003)
Geographic Controls	No	Yes	Yes	Yes
Location Controls	No	Yes	Yes	Yes
TseTse suit. & Pr. Genetic Distance	No	Yes	Yes	Yes
Individual Controls	No	No	Yes	Yes
Population and Night Lights	No	No	No	Yes
Group Size in Region	No	No	No	Yes
Number of Groups in Region	No	No	No	Yes
Group Fractionalization in Region	No	No	Yes	Yes
Ethnic Group Share in Region	No	No	No	Yes
Country FE	Yes	Yes	Yes	Yes
Observations	19,809	19,809	19,809	19,809
R-squared	0.114	0.121	0.143	0.145

Notes: The Table reports the OLS specification estimates associating the strength of individual identification with the ethnic group with the level Malaria Stability. In all specification, the dependent variable is the strength of ethnic identification with the group, based on Afrobarometer respondents (individual data), and 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 measures the level of malaria suitability in a radius of 10 km around the centroid of the village; see text for details. The geographic controls include soil suitability and elevation (mean and standard deviation), average temperature and precipitation, terrain ruggedness and caloric suitability before 1500 and TseTse suitability. The distance variables include distance from the equator, from the coast, from the river, from the country border, from the country capital and from East Africa. "Individual" controls include living condition FE (q4b), education FE (q90), religion FE (q91), occupation FE (q95), and rural/urban FE (q113). The unit of observation is the Afrobarometer respondent. Variable description, data sources and summary statistics are reported in Tables E5 and S6 respectively. Standard errors are double clustered by ethnic group and district of the respondent, reported in parentheses. ***, **, and * indicate significance at the 1-, 5-, and 10-% levels, respectively.

A5.3 Ethnolinguistic Diversity Today (Cell Level)

In this section, we turn our attention to ethnolinguistic diversity at the grid cell level using contemporaneous data for the distribution of the population (and not the historical homeland) today. Like Michalopoulos (2012), we employ the World Language Mapping System (WLMS) data (Lewis, 2005) portraying the spatial distribution of contemporary ethno-linguistc group. WLMS contains

information on all ethnolinguistic groups in Africa (including, for some countries like South Africa, also the descendants of European colonizers) and delivers the best available representation of the distribution of the ethnolinguistic population today at the disaggregate level.[14]

Table C4 reports the results (with standardized beta coefficients in square brackets). The replication of the baseline analysis confirms the existence of a significant, and quantitatively relevant, role of malaria. An increase of 1 cross-cells standard deviation of the Malaria Stability Index is associated with a 33% increase in the log of ethnolinguistic groups per cell.

Column (6) and (7) extends the specification by including measures of historical and pre-colonial ethnic diversity (in terms of the number of ethnic groups in a cell). This allows us to study the existence of a (further) direct effect on diversity today of the historical forces that shaped the emergence of ethnic groups in the past. In line with the view that the role of malaria in the emergence of ethnic groups was mostly confined to pre-modern societies, the results show that exposure to the pathogen has no significant effect on today's diversity once controlling for past diversity.

It is interesting to point that a similar pattern emerges, and a similar interpretation can be offered, for the disappearing role of the standard deviation of agricultural suitability, which can be taken as a proxy for the historical emergence of groups with productive specialization. Interestingly, the role of differences in elevation, which very likely still represents a barrier to the movements of the population still today, remains significant above and beyond the role of past diversity.

[14]The data conceptually differ with respect to the historical and pre-colonial data in some important dimensions. First, the WLMS accounts the migration of Africans during the colonial and post-colonial period. Together with historical and pre-colonial data, this allows us to explore the persistence of the geographic location of ethnic groups in the last century and the role of malaria. Second, unlike the Murdock maps, which track the ancestral homelands of ethnic groups, the WLMS delivers information on the actual distribution of the population today (in this respect being closer in concept to the GREG database). Differently from the pre-colonial maps, which aimed at locating the traditional ethnic homelands, the distribution of modern groups involves a substantial overlapping of areas occupied by the different linguistic groups. For this reason, we follow Michalopoulos closely and consider the number of ethnic groups in a cell (rather than the average size of each group) as the main variable of interest in the empirical analysis. For historical and pre-colonial ethnic diversity the overlap between homelands is not an issue.

Table C4: **World Language Mapping System Evidence**

	Ln (Number of Groups - WLMS)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Malaria Stability	0.035***	0.026***	0.012**	0.009*	0.008*	0.001	0.005
	(0.007)	(0.006)	(0.005)	(0.005)	(0.005)	(0.004)	(0.004)
	[0.427]	[0.322]	[0.153]	[0.108]	[0.098]	[0.013]	[0.063]
Ln(Number of Ethnic Groups GREG)						0.437***	
						(0.064)	
						[0.311]	
Ln(Number of Ethnic Groups, Pre-Colonial Africa)							0.441***
							(0.061)
							[0.322]
Geographic Controls:							
Ln(Cell Area)	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Soil Suitability and Elevation (Mean and Std.)	No	No	Yes	Yes	Yes	Yes	Yes
Avg. Temperature and Precipitation	No	No	Yes	Yes	Yes	Yes	Yes
Ruggedness and Caloric Suitability Pre 1500	No	No	Yes	Yes	Yes	Yes	Yes
Location Controls:							
Distances (equator, coast, river, border, capital)	No	No	No	Yes	Yes	Yes	Yes
Number of Countries and Within Country	No	No	No	Yes	Yes	Yes	Yes
TseTse suit. & Pr. Genetic Distance	No	No	No	No	Yes	Yes	Yes
Country FE	No	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,906	1,906	1,906	1,906	1,906	1,906	1,906
R-squared	.199	.418	.501	.521	.526	.587	.592

Notes:The Table reports the OLS specification estimates associating the log number of ethnic groups with the level of Malaria Stability in Africa. In all specification, the dependent variable is the natural logarithm of the number of ethnic groups in the cell measured using the World Language Mapping System (WLMS); Malaria Stability is the average level of malaria suitability in the cell; see text for details. The geographic variables include the natural logarithm of the cell land area, soil suitability and elevation (mean and standard deviation), average temperature and precipitation, terrain ruggedness, caloric suitability before 1500, within country and number of countries in the cell. The distance variables include distance from the equator, from the coast, from the river, from the country border and from the country capital. TseTse suitability measures the geographic suitability for Trypanosomiasis transmission. Predicted Genetic distance is computed as the log of migratory distance from East Africa. Ln(Number of Ethnic Groups GREG) is the log of the number of historical ethnic group measured by GREG; Ln(Number of Ethnic Groups, Pre-Colonial Africa) is the log of the number of historical ethnic group measured with the Murdock maps. Variable description, data sources and summary statistics are reported in Tables E1, E2, and S1, respectively. Beta coefficient in square brackets. Robust standard errors clustered by country are reported in round brackets. ***, **, and * indicate significance at the 1-, 5-, and 10-% levels, respectively.

A6 Endogamy: Further Results and Robustness

This section provides additional information regarding the data construction detailed in the text section 4, provides an additional exploration of the main data and includes several robustness checks and additional results complementing main results reported in the text.

A6.1 Additional Details on the Measurement of Ethnic Endogamy and Exogamy

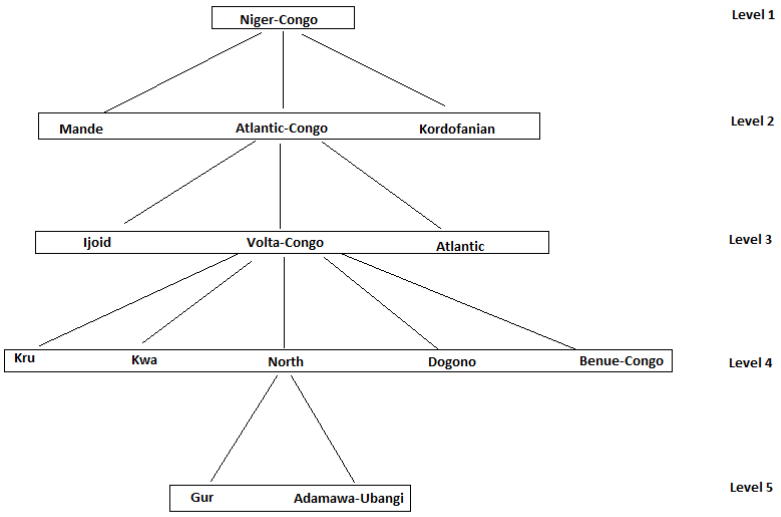
A crucial part of the data construction relates to the harmonization of ethnic names across DHS waves and to the matching of DHS self-reported ethnic identity with the ethnolinguistic trees from the Ethnologue.

Harmonization is necessary because using directly DHS reported ethnicity may lead to mistakes, as the degree of endogamy may be artificially generated by a change in DHS ethnic labelling of the same group across different waves and countries. For instance, for the same country, there might be one wave reporting identification of individual i and individual j with ethnicity x, while other waves may report ethnicity $x1$ for the individual a and ethnicity $x2$ for individual b (where $x1$ and $x2$ are sub-families of ethnicity x). Not accounting for the different classification, the same couple formed by individual i and individual j would be considered endogamous in the first wave, while the couple a and b would not be considered endogamous.

Matching DHS Ethnic Names to Ethnologue Tree. To deal with this issue, we matched each self-reported ethnicity of the DHS with the corresponding ethnicity in the Ethnologue tree. This offers two advantages: i) it permits to compare marriages across countries and across different waves; ii) it allows us to obtain a flexible definition of endogamy that varies at different levels of the linguistic tree. In fact, certain marriages are coded as exogamous at the more disaggregated branches of the Ethnologue tree, but the same marriage would be defined as endogamic at upper branches. Let us look at Figure D11 for an application. Suppose that a female respondent belongs to the Kru ethnic group and her spouse belong to the Dogon ethnic group. The couple would be considered as exogamous at level 4 of the Ethnologue; while, it will be listed as endogamous at level 3, as both groups pertain to the Volta-Congo family.

A6.2 Movers Distribution by Distance

The main analysis in Section 4 focuses on "movers", i.e. people that are currently residing outside their ethnic homeland. We plot the distribution of movers by the distance from their own ethnic homeland. Figure D13 Panel A reports the breakdown of the number of movers by i) less than



Panel A

Figure D11: **Ethnologue Tree: An Example**

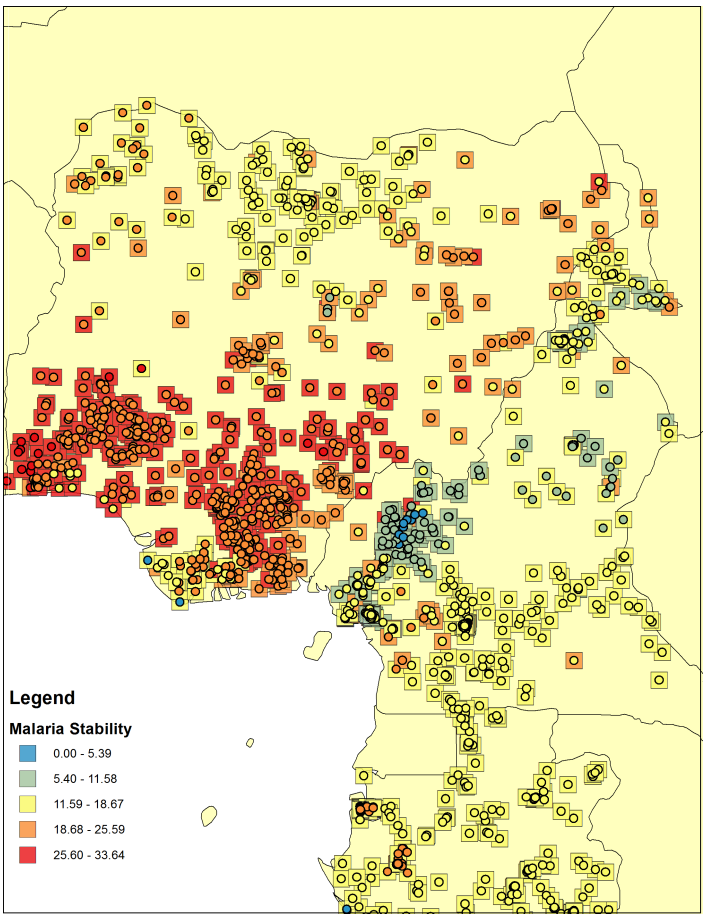
ten kilometres; ii) between ten and a hundred kilometers; iii) between a hundred and five hundred kilometers; iv) and more than five hundred kilometers. We find that 9.8% (1,210) of the movers are found living less than 10 kilometres away from their original ethnic homeland. Given the displacement in DHS data and the potential mistakes in border drawing of ethnic homelands, we do not include these movers in our regression sample. Between 10 and 100 km, there are 6,015 movers (49% of the total movers); while between 100 and 500 km, we spot around 35% (4,294) movers. The remaining 6.2% (762) lives more than 500 km away from their traditional homeland. In Panel B, we further disaggregate the distance categories according to the distance cut-offs used in Table D4.

A6.3 Different Ethnologue Levels

Table D1 we report the regression coefficients for some level of ethnic endogamy shown in Figure D14. Additionally, in Column (5) we show the robustness of the results to the alternative parametrization of the linguistic distance in the exogamy index. All results are confirmed. Figure D14 summarizes size and confidence intervals of coefficients at various level of the Ethnologue tree and show that the effect is present throughout different branches of the tree.

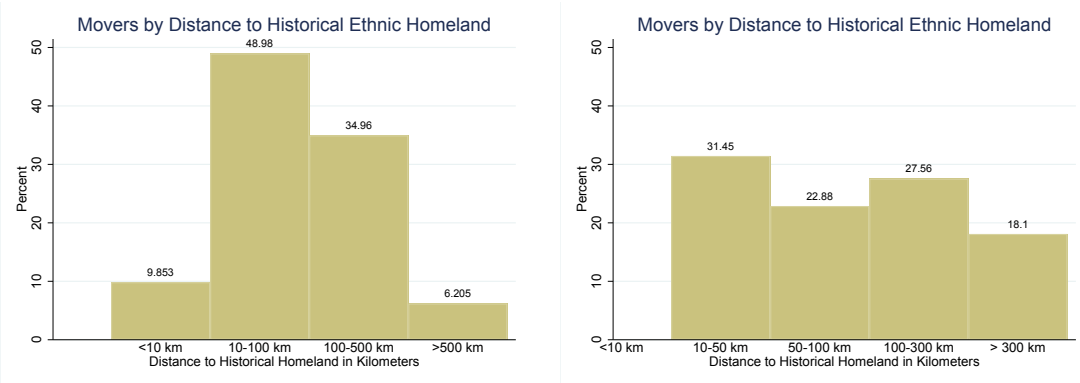
A6.4 Correlates of Movers

According to the results in Table D2, neither ancestral malaria nor the malaria of the location of residence are systematic significant predictors of the probability of being a mover. Being a mover is mostly predicted by location-specific characteristics as documented by the sharp and large rise in the



Panel A

Figure D12: **Location and Mover's Ancestral Malaria**
Figure D12 depicts the spatial distribution of couples in Nigeria and Cameroon. The Malaria stability in the location is represented by the color of the circle, the average ancestral malaria stability of respondents is represented by the color of the square.



Panel A Panel B

Figure D13: **Movers by Distance**

R-square following the inclusion of location fixed effects in Column (5), an increase from around 0.1 to around 0.85. In particular (unreported) results investigating the role of local characteristics show that higher soil suitability and lower average elevation of the location are associated with a higher

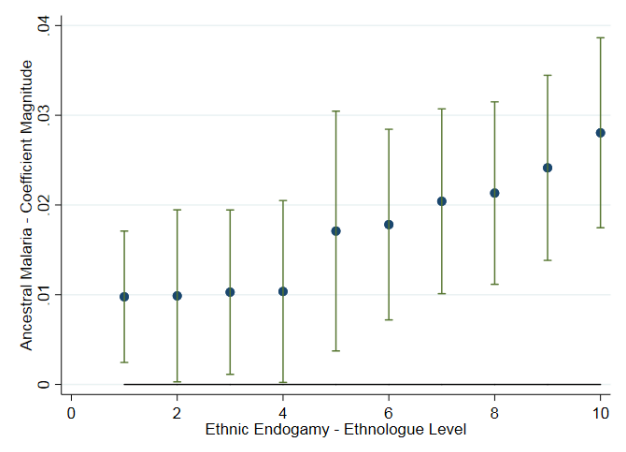


Figure D14: **Ancestral Malaria and Endogamy: Levels of Ethnologue Tree**

The graph plots the estimated coefficients (*y-axis*) obtained by regressing the ethnic endogamy dummy of "movers" at various level of the Ethnologue Tree (*x-axis*) on the Ancestral Malaria index. All regressions include country, year, and village fixed effect, "DHS Individual" controls, "Ancestral Geographic" controls and size of ethnic group in the region. The sample and specification correspond to the ones of Column (4) in Table 3.

likelihood of having movers. Individual characteristics do not appear to increase explicative power much either; see Column (6).

Table D1: **Ancestral Malaria and Endogamy: Different Ethnologue Level**

	Endogamy (Dummy) Different Levels Ethnologue				Exogamy
	(1) Lvl 2)	(2) Lvl 4)	(3) Lvl 8)	(4) Lvl 10)	(5) δ=0.05
	(1)	(2)	(3)	(4)	(5)
Ancestral Malaria Stability	0.013***	0.013***	0.020***	0.022***	-0.016***
	(0.004)	(0.005)	(0.005)	(0.005)	(0.004)
	[0.517]	[0.461]	[0.535]	[0.425]	[-0.722]
Ancestral Controls	Yes	Yes	Yes	Yes	Yes
Individual Controls	Yes	Yes	Yes	Yes	Yes
Village FE	Yes	Yes	Yes	Yes	Yes
Country FE	No	No	No	No	No
Wave FE	Yes	Yes	Yes	Yes	Yes
Size Group	Yes	Yes	Yes	Yes	Yes
Observations	9400	9400	9400	9400	9400
R-squared	0.47	0.48	0.55	0.53	0.48

Notes: The table reports the OLS estimates associating the probability of being endogamous (or the exogamy index) with the location and ancestral level of Malaria Stability, focusing on i) different level of the Ethnologue tree and ii) an alternative weighting of distance for the index of exogamy. The dependent variable - in Column 1 to 4 - is an a binary indicator variable taking value 1 if the marriage is between two people from the same linguistic family at levels of the Ethnologue tree 2, 4, 6 and 10 respectively. The dependent variable in Column 5 is the Exogamy Index (from Desmet et al.), measuring the linguistic distance between the spouses; Location Malaria is the average level of the Malaria Stability in the respondent's location, and Ancestral Malaria is the average level of the Malaria Stability in the Murdock ethnic homeland of the respondent's ethnicity; see text for details. Ancestral controls include soil suitability and elevation (mean and standard deviation), average temperature and precipitation, terrain ruggedness, caloric suitability before 1500, distance from the equator, from the coast, from the river, from the country border and from the country capital, TseTse suitability and Predicted Genetic distance. The DHS Individual controls include urban dummy, years of education, age, religion fixed effects, and dummies of relative wealth (poorest, poorer, middle, richer, richest) for each respondent. Size of Group is the number of individuals belonging to the ethnic group of the respondent in the location region. The unit of observation is the female DHS respondent. Variable description, data sources, and summary statistics are reported in Tables E6, E7, S7, and S8 respectively. Beta coefficient in square brackets. Robust standard errors clustered by ethnic group (DHS) are in parentheses. ***, **, and * indicate significance at the 1-, 5-, and 10-% levels, respectively.

Table D2: **Selection into Movers**

	Probability of Being a Mover					
	(1)	(2)	(3)	(4)	(5)	(6)
Local Malaria Stability	0.003		0.001			
	(0.004)		(0.004)			
	[0.062]		[0.026]			
Ancestral Malaria Stability		0.003	0.008	-0.017*	-0.016*	-0.016*
		(0.006)	(0.010)	(0.009)	(0.008)	(0.009)
		[0.057]	[0.145]	[-0.309]	[-0.293]	[-0.297]
Ancestral Controls	No	No	Yes	Yes	Yes	Yes
Individual Controls	No	No	No	No	No	No
Village FE	No	No	No	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	No	No	No
Wave FE	Yes	Yes	Yes	Yes	Yes	Yes
Size Group	No	No	No	No	Yes	Yes
Observations	19414	19414	19414	19414	19414	19414
R-squared	0.05	0.05	0.13	0.83	0.86	0.86

Notes: The table reports the Linear Probability Model (LPM) estimates associating the probability of being a mover with the location and ancestral level of Malaria Stability. In all specification, the dependent variable is a binary indicator variable taking value 1 if the DHS respondent lives in the ancestral location of her ethnic group, 0 otherwise. Location Malaria is the average level of the Malaria Stability in the respondent's location, and Ancestral Malaria is the average level of the Malaria Stability in the Murdock ethnic homeland of the respondent's ethnicity; see text for details. Ancestral controls include soil suitability and elevation (mean and standard deviation), average temperature and precipitation, terrain ruggedness, caloric suitability before 1500, distance from the equator, from the coast, from the river, from the country border and from the country capital, TseTse suitability and Predicted Genetic distance. The DHS Individual controls include urban dummy, years of education, age, religion fixed effects, and dummies of relative wealth (poorest, poorer, middle, richer, richest) for each respondent. The unit of observation is the female DHS respondent. Variable description, data sources, and summary statistics are reported in Tables E6, E7, S7, and S8 respectively. Beta coefficient in square brackets. Robust standard errors clustered by ethnic group (DHS) are in parentheses. ***, **, and * indicate significance at the 1-, 5-, and 10-% levels, respectively.

A6.5 Defining Homeland Using Thiessen Polygons

Errors and inaccuracies in the drawing of borders in the Murdock map might lead to errors in the identification of movers. To make sure that results are not driven by these inaccuracies, we examine robustness to the alternative definition of movers, performing a re-drawing of borders of the original Murdock map using a Thiessen polygon transformation. The mover sample (9569) is almost identical to the baseline one in Table 3 (9398). Results are identical to those of the baseline specification. Reassuringly, potential systematic bias in homeland's border drawing does not affect the validity of our findings.

Table D3: **Ancestral Malaria and Endogamy: Thiessen Polygon**

	Endogamy (Dummy)				Exogamy (Ethnolinguistic Distance)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Local Malaria Stability	0.002*	0.002	0.001		-0.000	-0.000	0.000	
	(0.001)	(0.001)	(0.001)		(0.000)	(0.000)	(0.000)	
	[0.087]	[0.066]	[0.042]		[-0.030]	[-0.043]	[0.027]	
Ancestral Malaria Stability			0.019***	0.028***			-0.008***	-0.010***
			(0.003)	(0.005)			(0.001)	(0.003)
			[0.549]	[0.819]			[-0.623]	[-0.774]
Ancestral Controls	No	No	Yes	Yes	No	No	Yes	Yes
Individual Controls	No	No	No	Yes	No	No	No	Yes
Village FE	No	No	No	Yes	No	No	Yes	Yes
Country FE	Yes	Yes	Yes	No	Yes	Yes	No	No
Wave FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Size Group	No	No	No	Yes	No	No	No	Yes
Observations	19416	9571	9571	9571	19416	9571	9571	9571
R-squared	0.12	0.17	0.20	0.53	0.02	0.02	0.06	0.43

Notes: The table reports the OLS estimates associating the probability of being endogamous (or the exogamy index) with the location and the ancestral level of Malaria Stability, defining mover using Thiessen polygons borders. The dependent variable in Columns (1)-(4) 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. The dependent variable in Columns (5)-(8) index of exogamy measuring the linguistic distance between the spouses, constructed following Desmet et. al (2011). Location Malaria is the average level of the Malaria Stability in the respondent's location, and Ancestral Malaria is the average level of the Malaria Stability in the Murdock ethnic homeland of the respondent's ethnicity; see text for details. Ancestral controls include soil suitability and elevation (mean and standard deviation), average temperature and precipitation, terrain ruggedness, caloric suitability before 1500, distance from the equator, from the coast, from the river, from the country border and from the country capital, TseTse suitability and Predicted Genetic distance. The DHS Individual controls include urban dummy, years of education, age, religion fixed effects, and dummies of relative wealth (poorest, poorer, middle, richer, richest) for each respondent. Size of Group is the number of individuals belonging to the ethnic group of the respondent in the location region. The unit of observation is the female DHS respondent. Variable description, data sources, and summary statistics are reported in Tables E6, E7, S7, and S8 respectively. Beta coefficient in square brackets. Robust standard errors clustered by ethnic group (DHS) are in parentheses. ***, **, and * indicate significance at the 1-, 5-, and 10-% levels, respectively.

A6.6 Heterogeneity on Movers by Distance

To further explore the potential role of selection into movers we explore the possible heterogeneous effect of ancestral malaria allowing for heterogeneous effects depending on the distance from the homeland. Results are reported in Table D4 The findings suggest that the effect of ancestral malaria on endogamy is significant across all distance-cutoffs, but its economic significance starts to decline after 300km.

Table D4: **Ancestral Malaria and Endogamy: Mover and Distance**

	Endogamy (Dummy)				Exogamy (Ethnolinguistic Distance)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Local Malaria Stability	0.002*	0.003	0.002		-0.000	-0.000	0.000	
	(0.001)	(0.002)	(0.002)		(0.000)	(0.000)	(0.000)	
	[0.087]	[0.085]	[0.067]		[-0.030]	[-0.041]	[0.018]	
Ancestral Malaria Stability								
× Less than 50km			0.021***	0.032***			-0.008***	-0.010***
			(0.003)	(0.005)			(0.001)	(0.003)
			[0.557]	[0.828]			[-0.551]	[-0.678]
× Between 50km and 100km			0.019***	0.028***			-0.007***	-0.009***
			(0.003)	(0.005)			(0.001)	(0.003)
			[0.438]	[0.665]			[-0.477]	[-0.596]
× Between 100km and 300km			0.017***	0.028***			-0.007***	-0.010***
			(0.004)	(0.005)			(0.001)	(0.003)
			[0.409]	[0.652]			[-0.457]	[-0.627]
× More than 300km			0.015***	0.027***			-0.006***	-0.009***
			(0.003)	(0.005)			(0.001)	(0.003)
			[0.228]	[0.401]			[-0.250]	[-0.378]
Ancestral Controls	No	No	Yes	Yes	No	No	Yes	Yes
Individual Controls	No	No	No	Yes	No	No	No	Yes
Village FE	No	No	No	Yes	No	No	Yes	Yes
Country FE	Yes	Yes	Yes	No	Yes	Yes	No	No
Wave FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Size Group	No	No	No	Yes	No	No	No	Yes
Observations	19416	9400	9400	9400	19416	9400	9400	9400
R-squared	0.12	0.17	0.21	0.53	0.02	0.02	0.06	0.43

Notes: The table reports the OLS estimates associating the probability of being endogamous (or the exogamy index) with the location and ancestral level of the Malaria Stability, exploring heterogeneity with respect to mover distance. The dependent variable in Columns (1)-(4) 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. The dependent variable in Columns (5)-(8) index of exogamy measuring the linguistic distance between the spouses, constructed following Desmet et. al (2011); Location Malaria Stability is the average level of the Malaria Stability in the respondent's location. In columns (3)-(4) and (7)-(8), we interacted the Ancestral Malaria Stability, which is the average level of the Malaria Stability in the Murdock ethnic homeland of the respondent's ethnicity, with a battery of dummy variables taking value one if the respondent reside i) less than 50km ii) between 50km and 100km iii) between 100km and 300km iv) and more than 300km from her homeland; see text for details. Ancestral controls include soil suitability and elevation (mean and standard deviation), average temperature and precipitation, terrain ruggedness, caloric suitability before 1500, distance from the equator, from the coast, from the river, from the country border and from the country capital, TseTse suitability and Predicted Genetic distance. The DHS Individual controls include urban dummy, years of education, age, religion fixed effects, and dummies of relative wealth (poorest, poorer, middle, richer, richest) for each respondent. Size of Group is the number of individuals belonging to the ethnic group of the respondent in the location region. The unit of observation is the female DHS respondent. Variable description, data sources, and summary statistics are reported in Tables E6, E7, S7, and S8 respectively. Beta coefficient in square brackets. Robust standard errors clustered by ethnic group (DHS) are in parentheses. ***, **, and * indicate significance at the 1-, 5-, and 10-% levels, respectively.

A6.7 Temperature-based Measure of Ancestral Malaria

We replicate the baseline analysis using the temperature-based version of the Malaria Stability index. Results, summarized in Table D5, follow closely the pattern emerged with our baseline measures, however, coefficients in certain specifications are less precisely estimated.

Table D5: **Ancestral Malaria and Endogamy: Temperature-based Measure of Ancestral Malaria**

	Endogamy (Dummy)				Exogamy (Ethnolinguistic Distance)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Local Malaria Stability (Temperature)	-0.006	0.007	0.003		-0.001	-0.005	0.001	
	(0.007)	(0.009)	(0.009)		(0.003)	(0.004)	(0.003)	
	[-0.026]	[0.025]	[0.013]		[-0.017]	[-0.057]	[0.011]	
Ancestral Malaria Stability (Temperature)			0.084**	0.087*			-0.036**	-0.049*
			(0.039)	(0.052)			(0.015)	(0.026)
			[0.271]	[0.277]			[-0.314]	[-0.434]
Ancestral Controls	No	No	Yes	Yes	No	No	Yes	Yes
Individual Controls	No	No	No	Yes	No	No	No	Yes
Village FE	No	No	No	Yes	No	No	Yes	Yes
Country FE	Yes	Yes	Yes	No	Yes	Yes	No	No
Wave FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Size Group	No	No	No	Yes	No	No	No	Yes
Observations	19416	9400	9400	9400	19416	9400	9400	9400
R-squared	0.12	0.17	0.19	0.52	0.02	0.03	0.04	0.43

Notes: The table reports the OLS estimates associating the probability of being endogamous (or the exogamy index) with the location and the ancestral level of Malaria Stability, defining mover using Thiessen polygons borders. The dependent variable in Columns (1)-(4) 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. The dependent variable in Columns (5)-(8) index of exogamy measuring the linguistic distance between the spouses, constructed following Desmet et. al (2011). Location Malaria Stability (Temperature) is the average level of the malaria stability predicted based on temperature in the respondent's location, and Ancestral Malaria Stability (Temperature) is the average level of malaria stability predicted based on temperature in the Murdock ethnic homeland of the respondent's ethnicity; see text for details. Ancestral controls include soil suitability and elevation (mean and standard deviation), average temperature and precipitation, terrain ruggedness, caloric suitability before 1500, distance from the equator, from the coast, from the river, from the country border and from the country capital, TseTse suitability and Predicted Genetic distance. The DHS Individual controls include urban dummy, years of education, age, religion fixed effects, and dummies of relative wealth (poorest, poorer, middle, richer, richest) for each respondent. Size of Group is the number of individuals belonging to the ethnic group of the respondent in the location region. The unit of observation is the female DHS respondent. Variable description, data sources, and summary statistics are reported in Tables E6, E7, S7, and S8 respectively. Beta coefficient in square brackets. Robust standard errors clustered by ethnic group (DHS) are in parentheses. ***, **, and * indicate significance at the 1-, 5-, and 10-% levels, respectively.

A6.8 Temperature-based Measure IV

In Table D6 we replicate baseline findings on ancestral malaria and endogamy instrumenting Malaria Stability with the predicted temperature-based measure of malaria stability. While not precisely estimated across specifications, coefficients are close to the baseline ones.

Table D6: **Ancestral Malaria and Endogamy: Temperature-based Measure IV**

	Endogamy (Dummy)				Exogamy (Ethnolinguistic Distance)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Local Malaria Stability	-0.001	0.002	-0.001		-0.000	-0.001	0.001	
	(0.002)	(0.002)	(0.002)		(0.001)	(0.001)	(0.001)	
	[-0.053]	[0.049]	[-0.031]		[-0.033]	[-0.112]	[0.089]	
Ancestral Malaria Stability			0.021**	0.023*			-0.009***	-0.013**
			(0.008)	(0.013)			(0.003)	(0.006)
			[0.597]	[0.648]			[-0.695]	[-1.012]
Ancestral Controls	No	No	Yes	Yes	No	No	Yes	Yes
Individual Controls	No	No	No	Yes	No	No	No	Yes
Village FE	No	No	No	Yes	No	No	Yes	Yes
Country FE	Yes	Yes	Yes	No	Yes	Yes	No	No
Wave FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Size Group	No	No	No	Yes	No	No	No	Yes
Observations	19416	9400	9400	9400	19416	9400	9400	9400
R-squared	0.12	0.17	0.20	0.53	0.02	0.02	0.05	0.43

Notes: The table reports the 2SLS estimates associating the probability of being endogamous (or the exogamy index) with the location and ancestral level of Malaria Stability, instrumented by the ancestral Malaria Stability (Temperature) proxy. The dependent variable in Columns (1)-(4) 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. The dependent variable in Columns (5)-(8) index of exogamy measuring the linguistic distance between the spouses, constructed following Desmet et. al (2011). Location Malaria is the average level of the Malaria Stability in the respondent's location, and Ancestral Malaria is the average level of the Malaria Stability in the Murdock ethnic homeland of the respondent's ethnicity; Ancestral Malaria Stability (Temperature) is the average level of malaria stability predicted based on temperature in the Murdock ethnic homeland of the respondent's ethnicity. Ancestral controls include soil suitability and elevation (mean and standard deviation), average temperature and precipitation, terrain ruggedness, caloric suitability before 1500, distance from the equator, from the coast, from the river, from the country border and from the country capital, TseTse suitability and Predicted Genetic distance. The DHS Individual controls include urban dummy, years of education, age, religion fixed effects, and dummies of relative wealth (poorest, poorer, middle, richer, richest) for each respondent. Size of Group is the number of individuals belonging to the ethnic group of the respondent in the location region. The unit of observation is the female DHS respondent. Variable description, data sources, and summary statistics are reported in Tables E6, E7, S7, and S8 respectively. Beta coefficient in square brackets. Robust standard errors clustered by ethnic group (DHS) are in parentheses. ***, **, and * indicate significance at the 1-, 5-, and 10-% levels, respectively.

A6.9 Genetic Immunities

We experiment with a population-based measure of historical malaria exposure. Looking at ancestral genetic immunities, in terms of the frequency of the Duffy antigen, as an alternative proxy for long-term exposure to malaria confirms the patterns (see Table D7).

Table D7: **Ancestral Malaria and Endogamy: Duffy**

	Endogamy (Dummy)				Exogamy (Ethnolinguistic Distance)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Local Duffy Antigen	0.123**	0.126	0.078		-0.014	-0.020*	-0.001	
	(0.060)	(0.083)	(0.061)		(0.016)	(0.010)	(0.011)	
	[0.073]	[0.074]	[0.046]		[-0.022]	[-0.032]	[-0.002]	
Ancestral Duffy Antigen			0.136	0.105			-0.097***	-0.169***
			(0.105)	(0.109)			(0.030)	(0.040)
			[0.077]	[0.060]			[-0.151]	[-0.263]
Ancestral Controls	No	No	Yes	Yes	No	No	Yes	Yes
Individual Controls	No	No	No	Yes	No	No	No	Yes
Village FE	No	No	No	Yes	No	No	Yes	Yes
Country FE	Yes	Yes	Yes	No	Yes	Yes	No	No
Wave FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Size Group	No	No	No	Yes	No	No	No	Yes
Observations	19416	9400	9400	9400	19416	9400	9400	9400
R-squared	0.12	0.17	0.19	0.52	0.02	0.02	0.04	0.43

Notes: The table reports the OLS estimates associating the probability of being endogamous (or the exogamy index) with the location and ancestral level of the Duffy Antigen. The dependent variable in Columns (1)-(4) 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. The dependent variable in Columns (5)-(8) index of exogamy measuring the linguistic distance between the spouses, constructed following Desmet et. al (2011); Location Duffy Antigen is the average level of the Malaria Stability in the respondent's location, and Ancestral Duffy Antigen is the average level of the Malaria Stability in the Murdock ethnic homeland of the respondent's ethnicity; see text for details. Ancestral controls include soil suitability and elevation (mean and standard deviation), average temperature and precipitation, terrain ruggedness, caloric suitability before 1500, distance from the equator, from the coast, from the river, from the country border and from the country capital, TseTse suitability and Predicted Genetic distance. The DHS Individual controls include urban dummy, years of education, age, religion fixed effects, and dummies of relative wealth (poorest, poorer, middle, richer, richest) for each respondent. Size of Group is the number of individuals belonging to the ethnic group of the respondent in the location region. The unit of observation is the female DHS respondent. Variable description, data sources, and summary statistics are reported in Tables E6, E7, S7, and S8 respectively. Beta coefficient in square brackets. Robust standard errors clustered by ethnic group (DHS) are in parentheses. ***, **, and * indicate significance at the 1-, 5-, and 10-% levels, respectively.

A6.10 Multi-Host vector-borne Diseases: Placebo

Table D8 shows that Ancestral Trypanosomiasis and Ancestral Dengue/Yellow Fever, while being also vector-transmitted diseases, are not related to higher likelihood of endogamy.

Table D8: **Ancestral Malaria and Endogamy: Multi-Host Diseases Placebo**

	Endogamy (Dummy)	Exogamy Index	Endogamy (Dummy)	Exogamy Index	Endogamy (Dummy)	Exogamy Index
	(1)	(2)	(3)	(4)	(5)	(6)
Ancestral Malaria	0.029***	-0.010***	0.026***	-0.010***	0.028***	-0.010***
	(0.005)	(0.003)	(0.005)	(0.004)	(0.005)	(0.004)
	[0.821]	[-0.745]	[0.738]	[-0.749]	[0.789]	[-0.784]
Ancestral Trypanosomiasis	-0.177	0.055			-0.192*	0.049
	(0.109)	(0.034)			(0.111)	(0.034)
	[-0.160]	[0.136]			[-0.173]	[0.120]
Ancestral Dengue/Yellow Fever			-0.108	-0.072	-0.143	-0.063
			(0.117)	(0.062)	(0.122)	(0.061)
			[-0.063]	[-0.115]	[-0.084]	[-0.101]
Ancestral Controls	Yes	Yes	Yes	Yes	Yes	Yes
Individual Controls	Yes	Yes	Yes	Yes	Yes	Yes
Village FE	Yes	Yes	Yes	Yes	Yes	Yes
Country FE	No	No	No	No	No	No
Wave FE	Yes	Yes	Yes	Yes	Yes	Yes
Size Group	Yes	Yes	Yes	Yes	Yes	Yes
Observations	9400	9400	9400	9400	9400	9400
R-squared	0.53	0.43	0.53	0.43	0.53	0.43

Notes: The table reports the OLS estimates associating the probability of being endogamous (or the exogamy index) with the location and the ancestral level of Malaria Stability. Endogamy (Dummy) 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. Exogamy is an index of exogamy measuring the linguistic distance between the spouses, constructed following Desmet et. al (2011). Location Malaria is the average level of the Malaria Stability in the respondent's location, and Ancestral Malaria is the average level of the Malaria Stability in the Murdock ethnic homeland of the respondent's ethnicity; Ancestral Trypanosomiasis is the average level of TseTse suitability in the Murdock ethnic homeland of the respondent's ethnicity while Ancestral Dengue/Yellow Fever measure the average level of Aedes Aegipty suitability in the Murdock ethnic homeland of the respondent's ethnicity; see text for details. Ancestral controls include soil suitability and elevation (mean and standard deviation), average temperature and precipitation, terrain ruggedness, caloric suitability before 1500, distance from the equator, from the coast, from the river, from the country border and from the country capital, TseTse suitability and Predicted Genetic distance. The DHS Individual controls include urban dummy, years of education, age, religion fixed effects, and dummies of relative wealth (poorest, poorer, middle, richer, richest) for each respondent. Size of Group is the number of individuals belonging to the ethnic group of the respondent in the location region. The unit of observation is the female DHS respondent. Variable description, data sources, and summary statistics are reported in Tables E6, E7, S7, and S8 respectively. Beta coefficient in square brackets. Robust standard errors clustered by ethnic group (DHS) are in parentheses. ***, **, and * indicate significance at the 1-, 5-, and 10-% levels, respectively.

A7 Data Sources and Description

Table E1: Main Variables and Data Sources: 1 degree grid cells

Explanatory Variables:

Malaria Stability Average Malaria Stability Index in the 1x1 degree grid cell. Source: Kiszewski et al. (2004).

Malaria Stability (Temperature) Average Malaria Stability Index in the 1x1 degree grid cell predicted using temperature. Source: see Section A2.1 for details.

Falciparum Suitability Average Falciparum Suitability in the 1x1 degree grid cell predicted using temperature. Source: Gething et al. (2011).

Malaria Endemicity Average Historical Malaria Endemicity in the 1x1 degree grid cell. Source: Malaria Endemicity map, devised by Lysenko (1968) and digitized by Hay (2004).

Duffy Antigen Share of Duffy negative phenotype in the 1x1 degree grid cell. Source: Howes et al. (2011).

Dependent Variables:

Number of Ethnic Groups - GREG and Murdock, WLMS Number of languages in the 1x1 degree cell. Source: GREG shapefile, Murdock maps for Africa and the Americas, World Language Mapping System.

Land Fractionalization GREG, Murdock, WLMS Land fractionalization in the 1x1 degree cell. Computed as a standard fractionalization index based on the share of land occupied by each group in the cell. Source: GREG shapefile, Murdock maps for Africa and the Americas, World Language Mapping System.

Table E2: Covariates and Data Sources: 1 degree grid cells

Geographic Covariates:

Average Temperature. Mean annual 1x1 degree cell temperature, baseline period 1901-1960. Source: CRU CL 2.0 data from New (2002).

Average Precipitation. Average 1x1 degree cell monthly precipitation mm/month, baseline period 1901-1960. Source: CRU CL 2.0 data from New (2002).

Land Suitability and Variation in Land Suitability. Average land suitability and standard deviation of land suitability in the 1x1 degree cell. Source: Ramankutty (2002).

Mean Elevation and Variation in Elevation. Average 1x1 degree cell elevation and cell standard deviation in elevations. Source: National Oceanic and Atmospheric Administration (NOAA) and U.S. National Geophysical Data Center, Terrain Base, release 1.0 (CD-ROM), Boulder, Colo.

Ruggedness. Average 1x1 degree cell ruggedness - Terrain Ruggedness Index, 100 m. Source: Terrain Ruggedness Index originally devised by Riley, DeGloria, and Elliot (1999), obtained through http://diegopuga.org.

Caloric Suitability pre-1500. Average caloric suitability pre-1500. Source: Galor and Ozak (2016).

TseTse Fly Suitability. Average predicted suitability for TseTse flies. Source: constructed as the sum of predicted suitability (0 to 1) for the presence of TseTse groups (Fusca, Morsitansand 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.

Dengue/Yellow Fever Suitability. Average predicted suitability for *Aedes Aegypti.* Source: Kraemer *et al.* (2015).

Total Water Area. Total area occupied by water within the 1x1 degree cell. Source: Digital Chart of the World inwater shapefile. We sum up total in-cell water area and the areas of the cell occupied by seas and oceans.

Total Area. Total area of the 1x1 degree cell. We exclude cell parts not covered by ethnic groups.

Number of Countries. Total number of countries in the 1x1 degree cell. Source: Digital Chart of the World.

Within Country. Dummy variable taking value one if the 1x1 degree cell belong to one single country, 0 otherwise. Source: Digital Chart of the World.

Distances:

Predicted Genetic Distance. Ln Migratory distance, on a land path, from Adis Ababa. Source: computed following Ashraf and Galor (2013). The distance of the centroid of 1x1 degree cell from from Adis Ababa 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 connects several obligatory intermediate points, and namely: Cairo, Istambul, Phnom Phen, Anadyr and Prince Rupert.

Ln Distance Coast and *Ln Distance Border.* Source: Digital Chart of the World coastline shapefile.

Ln Distance Capital. Distance to the capital of the country where lies the centroid of the 1x1 degree cell. Source: Digital World Capital shapefile.

Ln Distance River. Distance to closest river. Source: 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.

Proxies for Economic Development:

Night Lights. Average luminosity at night in 1x1 degree cell. Source: NOOA National Geophysical Data Centre for the year 2000.

Ln Population Density. Average population in 1x1 degree cell. Source: NOOA National Geophysical Data Centre for the year 2000. enter 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.

Table E3: Pre-Colonial Covariates and Data Sources: 1 degree grid cells

Pre-Colonial Covariates: Average value of pre-colonial features of the ethnic groups in the cell. Source: variable in the Ethnographic Atlas (Murdock, 1967).

Gathering (V1), *Hunting* (V2), *Fishing* (V3), *Animal Husbandry* (V4) and *Agriculture Dependence* (V5): dummies variable taking value one if dependence on the subsistence practice is above 5%.

Polygyny (V8): dummy variable equal to one if polygyny is present in a society and 0.

Clan Communities (V15): dummy variable taking value 1 if the marriage community system is "clan communities or clan barrios" and 0 otherwise.

Agriculture Type (V28): ordered variable ranging from 0 (no agriculture) to 4 (intensive irrigated agriculture).
Settlements (V30): ordered variable raging from 1 (migratory) to 8 (groups residing in complex settlements) indicating the "settlement pattern of each group".

Complex Settlement (V30): dummy variable taking value 1 if groups reside in permanent settlements (Settlements=7) or complex settlments (Settlements=8) and 0 otherwise.

Local Community (V32): dummy variable taking value 1 if local community is organized in nuclear family, extended family, clan barrio, or village and 0 otherwise.

Milking (V41): dummy variable taking value 1 when animals are "Milked more often than sporadically" and 0 when "Little or no milking".

Slavery (V70): dummy variable taking value 1 when some type of slavery (hereditary, incipient, or significant) is present and 0 otherwise.

Table E4: Ethnic Admixing Today and Data Sources: (Village Level)

Explanatory Variable:

Malaria Stability: Average Malaria Stability in the 10 km radius around the coordinates of each cluster. Source: Malaria Stability index from Kiszewski et al.(2004).

Dependent Variables:

Ethnic Admixing. Number of ethnolinguistic groups (log) and ethnic fractionalization in the village. Constructed using individual survey of both male and female. Sourche: Demographic and Health Survey.

Covariates:

Geographic Controls, Population and Night Lights: Soil suitability and elevation (mean and standard deviation), average temperature and precipitation, terrain ruggedness, caloric suitability before 1500, distance from the equator, from the coast, from the river, from the country border and from the country capital, TseTse suitability and Predicted Genetic distance, average population and average night lights in a radius of 10 km around the centroid of the village of the respondent. Sources and data construction: see Table E2.

Number of Groups in Region. Number of different ethnic groups in the region out of all respondents, constructed using all individual (males and females) available in the surveys. Source: DHS.

Group Fractionalization in Region. Ethnic Fractionalization in the region, constructed using all individual (males and females) available in the surveys. Source: DHS.

Table E5: Ethnic Identification Today: Data Sources (Individual Level)

Explanatory Variable:

Malaria Stability: Average Malaria Stability in the 10 km radius around the coordinates of each cluster. Source: Malaria Stability index from Kiszewski et al. (2004).

Dependent Variable:

Ethnic Identification. Indicating the strength of self-identification with the ethnic group as compared to the state. 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 re-coding, a value of one corresponds 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)'.

Covariates:

Geographic Covariates, Population and Night Lights. Soil suitability and elevation (mean and standard deviation), average temperature and precipitation, terrain ruggedness, caloric suitability before 1500, distance from the equator, from the coast, from the river, from the country border and from the country capital, TseTse suitability and Predicted Genetic distance, average population and average night lights in a radius of 10 km around the centroid of the village of the respondent. Sources and data construction: see Table E2.

Individual Controls. Living conditions (variable q4b), education level (variable q90), religion (variable q91), type of occupation (variable q95), living in a rural/urban area (variable q113). Source: Afrobarometer.

Group Size in Region. Number of respondents of the same group in the region. Source: Afrobarometer.

Number of Groups in Region. Number of different ethnic groups in the region out of all respondents. Source: Afrobarometer.

Group Fractionalization in Region. Ethnic Fractionalization in the region, computed out of respondents. Source: Afrobarometer.

Ethnic Group Share in Region. Respondent's ethnic group share in the region, out of all respondents. Source: Afrobarometer.

Table E6: Ethnic Endogamy Today and Data Sources

Dependent Variables

Endogamous Marriage: Dummy variable equal to one for a marriage between two individuals of an ethnic group belonging to the same linguistic family at level n of the Ethnologue, zero otherwise (we use level 6 as baseline). Since the number of branches varies among linguistic families and subfamilies, we follow the baseline approach from Desmet et al. (2012). Baseline variable is constructed extending the last available branch to the bottom; As robustness, we exclude observations that do not have the same information on linguistic branches. Source: Demographic and Health Survey (DHS) $v131$ for female and $mv131$ for male for most of the waves. As discussed extensively in the text, for each group we matched DHS ethnic group to Ethnologue linguistic group in order to 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 Desmet et al. (2012) . Given that the maximum level of branches is 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 Desmet et al. (2012) and extend the last available branch to the bottom. Source: Demographic and Health Survey (DHS) ($v131$ for female and $mv131$ for male for most of the waves). As discussed extensively in the text, we matched DHS ethnic group to Ethnologue linguistic group in order to have information on the linguistic family at each level of the linguistic tree.

Explanatory Variables:

Malaria Location: Average Malaria Stability Index in the 10 km radius around the coordinates of each cluster. Source: Malaria Stability index from Kiszewski et al. (2004).

Ancestral Malaria: Average Malaria Stability Index in the ethnic homeland of the respondent's ethnic group. Source: Malaria Stability index from Kiszewski et al. (2004).

Malaria Stability (Temperature), Falciparum Suitability, Duffy Location: Average value in the 10 km radius around the coordinates of each cluster. Sources and data construction: see Table E2.

Ancestral Malaria Stability (Temperature), Falciparum Suitability, Duffy: Average value in the ethnic homeland of the respondent's ethnic group. Sources and data construction: see Table E2.

Individual Characteristics:

Movers. Indicator variable taking value 1 if the individual belongs to an ethnic group located in the historical ethnic homeland of the group (or within 10 km from it), and 0 otherwise (constructed using ArcGIS). Source: Demographic and Health Survey (DHS) ($v131$ for female and $mv131$ for male for most of the waves), Murdock map of Africa.

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

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

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).

Female 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, constructed using all individual (males and females) available in the surveys. 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, constructed using all individual (males and females) available in the surveys. Source: Demographic and Health Survey (DHS) ($v131$)

Table E7: Ethnic Endogamy Today and Data Sources

Ancestral Characteristics

Ancestral Covariates:

Geographic Covariates. Soil suitability and elevation (mean and standard deviation), average temperature and precipitation, terrain ruggedness, caloric suitability before 1500, distance from the equator, from the coast, from the river, from the country border and from the country capital, TseTse suitability and Predicted Genetic distance, average population and average night lights in the ethnic homeland of the respondent's ethnic group. Sources and data construction: see Table E2.

Regional Characteristics:

Number of Groups in the Region. Number of different ethnic groups in the region, constructed using all individual (males and females) available in the surveys. Source: Demographic and Health Survey (DHS) ($v131$)

Group Fractionalization in the Region Ethnic fractionalization in the region, constructed using all individual (males and females) available in the surveys. Source: Demographic and Health Survey (DHS) ($v131$)

A7.1 Summary Statistics

Table S1: **Summary statistics - Africa Cell Level Analysis**

Variable	Mean	Std. Dev.	Min.	Max.	N
Main Variables					
Number of Ethnic Groups GREG	2.128	1.187	1	7	1976
Territorial Fractionalization, GREG	0.233	0.241	0	0.784	1976
Number of Ethnic Groups, Pre-Colonial Africa	2.655	1.471	1	13	1976
Territorial Fractionalization, Pre-Colonial Africa	0.316	0.25	0	0.855	1976
Malaria Stability	10.657	9.366	0	34.728	1976
Malaria Stability (Temperature)	1.721	0.904	0	3.95	1971
Falciparum Suitability	39.842	16.41	2.11	73.408	1976
Duffy Antigen	0.821	0.237	0.035	1	1976
Malaria Endemicity	3.031	1.583	0	5	1976
Covariates					
Avg. Temperature	24.354	3.426	11.062	29.655	1976
Avg. Precipitation	681.989	496.906	0	2823.143	1976
Mean Suitability	0.298	0.26	0.001	0.992	1976
Variation Suitability	0.043	0.045	0	0.282	1976
Mean Elevation	656.607	438.611	-0.714	2517.88	1976
Variation Elevation	125.356	124.903	2.963	791.157	1976
Ruggedness	0.606	0.813	0.004	6.905	1976
Caloric Suit. Pre 1500	873.049	611.057	0	2175.441	1976
TseTse Suitability	0.511	0.65	0	2.332	1976
Dengue Suitability	0.583	0.281	0.004	0.952	1976
Number of Country	1.344	0.562	1	4	1976
Within Country	0.699	0.459	0	1	1976
Migratory Distance	2821.984	1431.272	78.073	5959.788	1976
Distance Coast	666997.158	452418.908	224.194	1711029.375	1976
Distance Border	129565.263	113890.768	56.494	596304.563	1976
Distance River	408856.051	313710.515	445.524	1541326.625	1976
Distance Capital	618787.487	386173.844	15530.111	1951377.125	1976
Absolute Latitude	12.785	7.654	0.417	34.417	1976
Light Density	2.62	0.964	2	16.455	1976
Population Density	26.006	57.146	0.019	1215.395	1976

Table S2: **Summary statistics - Old World Cell Level Analysis**

Variable	Mean	Std. Dev.	Min.	Max.	N
Number of Ethnic Groups GREG	1.923	1.157	1	10	9566
Territorial Fractionalization, GREG	0.181	0.223	0	0.859	9566
Malaria Stability	2.504	6.141	0	34.728	9566
Avg. Temperature	9.762	13.854	-22.323	29.655	9566
Avg. Precipitation	554.71	539.695	0	4038.99	9566
Mean Suitability	0.28	0.312	0	0.998	9566
Variation Suitability	0.034	0.048	0	0.409	9566
Mean Elevation	654.507	831.922	-99.081	5725.512	9566
Variation Elevation	140.947	161.742	0	1868.89	9566
Ruggedness	1.011	1.219	0	10.168	9566
Caloric Suit. Pre 1500	752.356	790.424	0	3001.149	9566
Number of Country	1.208	0.471	1	4	9566
Within Country	0.818	0.386	0	1	9566
Migratory Distance	6548.091	3010.624	78.073	14793.704	9566
Distance Coast	480709.893	462407.192	6.51	2075709.875	9566
Distance Border	164026.318	165798.132	6.506	999491.313	9566
Distance River	458697.311	498621.074	48.161	3453667.5	9566
Distance Capital	2414553.51	3301652.354	3818.786	21012714	9566
Absolute Latitude	38.233	20.85	0.417	76.5	9566
Light Density	3.632	2.877	0	36.75	9530
Population Density	59.113	161.169	0	3718.465	9566

Table S3: **Summary statistics - Pre-Colonial Covariates (Cell Level)**

Variable	Mean	Std. Dev.	Min.	Max.	N
Gathering	0.382	0.413	0	1	1703
Hunting	0.717	0.379	0	1	1703
Fishing	0.518	0.435	0	1	1703
Animal husbandry	2.926	2.317	0	9	1703
Agriculture Dep.	4.675	2.081	0	9	1703
Agriculture Type	2.206	0.915	0	4	1648
Milking	0.568	0.467	0	1	1648
Settlements	3.829	2.271	0	7	1648
Complex Settlement	0.417	0.435	0	1	1648
Polygyny	0.934	0.224	0	1	1687
Clan communities	0.358	0.424	0	1	1565
Slavery	0.823	0.348	0	1	1670
Property Rights	0.675	0.457	0	1	820
Jurisdicial Hierarchy	1.321	0.87	0	4	1624

Table S4: **Summary statistics - Pre-Colonial Covariates America (Cell Level)**

Variable	Mean	Std. Dev.	Min.	Max.	N
Number of Ethnic Groups, Pre-Colonial America	2.061	0.940	1	6	1503
Malaria Stability	2.863	2.3	0	6.949	1503
Malaria Stability (Temperature)	2.203	1.167	0	3.92	1487
Falciparum Suitability	49.443	20.814	0.046	75.888	1503
Ln (Land Area)	2.373	0.336	0.086	2.511	1503
Avg. Temperature	23.147	4.336	3.563	28.382	1503
Avg. Precipitation	1331.645	711.494	0	5938.477	1503
Mean Suitability	0.449	0.248	0.001	0.996	1503
Variation Suitability	0.056	0.051	0	0.326	1503
Mean Elevation	556.75	728.731	-4.567	4218.691	1503
Variation Elevation	181.65	253.34	0	1403.824	1503
Ruggedness	0.652	0.962	0.001	6.486	1503
Caloric Suit. Pre-1500	2413.271	1071.994	0	5165.59	1503
Ln(Distance Coast)	5.608	1.463	-0.841	7.34	1503
Ln(Distance River)	5.024	1.295	-0.862	7.231	1503
Absolute Latitude	13.216	9.128	0.5	33.5	1503

Table S5: **Summary statistics: DHS - Ethnic Admixing**

Variable	Mean	Std. Dev.	Min.	Max.	N
Number of Ethnic Group in the Village	3.179	2.136	1	17	13187
Group Fractionalization in the Village	0.312	0.268	0	0.914	13187
Malaria Stability	14.313	9.872	0	38.054	13187
Number of DHS Respondents in the Village	33.148	15.544	1	134	13187
Average Precipitation	999.614	511.221	13.183	2907.742	13187
Average Temperature	24.494	3.151	13.375	29.844	13187
Caloric Suitability pre-1500	1175.392	469.224	0	2258.391	13187
Land Suitability (mean)	0.466	0.257	0.001	0.999	13187
Land Suitability (std)	0.015	0.015	0	0.099	13187
Elevation (std)	77.106	107.361	0	949.784	13187
Elevation (mean)	631.321	611.352	0	3718	13187
Terrain Ruggedness	65599.101	115593.422	0	2173258	13187
TseTse Suitability	0.303	0.314	0	1	13187
Ln (Distance Capital)	0.43	1.448	-7.188	2.807	13187
Ln (Distance River)	-4.029	5.781	-9.210	3.537	13187
Ln (Distance Border)	-2.819	4.003	-9.210	1.372	13187
Ln (Distance Coast)	12.297	1.61	-9.210	14.339	13187
Ln (Distance Adis Ababa)	7.894	0.948	-0.732	8.724	13187
Abs. Latitude	10.378	5.554	0	28.72	13187
Group Fractionalization in the Region	0.551	0.216	0	0.901	13187
Number of Groups in the Region	8.099	4.1	1	26	13187
Night Lights	6.497	11.552	2	63	13187
Population Density	423.957	1199.707	0.133	11133.601	13187

Table S6: **Summary statistics: Afrobarometer - Ethnic Identity**

Variable	Mean	Std. Dev.	Min.	Max.	N
Strength Ethnic Identification	2.551	1.174	1	5	19809
Malaria Stability	10.963	10.12	0	36.04	19809
Night Lights	7.075	12.599	2	63	19809
Population Density	475.81	1388.337	0.004	10124.006	19809
Female Age	27.928	6.103	18	68	19809
Female Respondent Urban	1.634	0.482	1	2	19809
Ethnic 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
Ethnic Group Share in the Region	0.573	0.334	0.002	1	19809
Group Fractionalization in the Region	0.433	0.262	0	0.899	19809
Average Precipitation	846.693	349.984	13.183	2445.047	19809
Average Temperature	22.716	4.12	8.465	29.617	19809
Ruggedness	93272.435	163366.087	0	1641536	19809
Elevation (std)	75.620	100.694	0	1001.294	19809
Land Suitability (std)	0.016	0.015	0	0.087	19809
Elevation (mean)	789.494	634.625	0	2768	19809
Land Suitability (mean)	0.461	0.229	0	0.988	19809
Caloric Suitability pre-1500	1295.396	478.547	0	2289.685	19809
TseTse Suitability	0.21	0.271	0	1	19809
Ln (Distance Capital)	0.401	1.512	-5.234	2.808	19809
Ln (Distance River)	10.528	1.496	-9.210	13.107	19809
Ln (Distance Border)	-2.834	4.047	-9.210	1.318	19809
Ln (Distance Coast)	12.274	1.635	7.288	14.193	19809
Abs. Latitude	13.271	9.026	0	34.317	19809
Ln (Distance Adis Ababa)	7.998	0.523	6.431	8.723	19809

Table S7: **Summary statistics: DHS Full Sample**

Variable	Mean	Std. Dev.	Min.	Max.	N
Endogamy Lev 1	0.99	0.101	0	1	19414
Endogamy Lev 2	0.964	0.185	0	1	19414
Endogamy Lev 3	0.961	0.194	0	1	19414
Endogamy Lev 4	0.954	0.21	0	1	19414
Endogamy Lev 5	0.938	0.241	0	1	19414
Endogamy Lev 6	0.923	0.266	0	1	19414
Endogamy Lev 7	0.915	0.279	0	1	19414
Endogamy Lev 8	0.91	0.286	0	1	19414
Endogamy Lev 9	0.837	0.369	0	1	19414
Endogamy Lev 10	0.78	0.414	0	1	19414
Local Malaria Stability	12.548	10.167	0	36.957	19414
Ancestral Malaria Stability	13.458	9.288	0	33.638	19414
Ancestral Tsetse Suitability	0.223	0.827	-2.474	1.46	19414
Ancestral Dengue Suitability	0.666	0.178	0.034	0.915	19414
Ancestral Pr. Genetic Distance	7.788	0.747	4.806	8.705	19414
Ancestral Precipitation	23.754	2.853	17.344	29.075	19414
Ancestral Temperature	1015.59	419.712	125.095	2773.066	19414
Ancestral Agricultural Suit. (mean)	0.491	0.18	0.015	0.983	19414
Ancestral Agricultural Suit. (std)	0.122	0.048	0.009	0.274	19414
Ancestral Elevation (mean)	749.837	563.449	6.953	2137.922	19414
Ancestral Elevation (std)	232.682	173.437	6.207	673.904	19414
Ancestral Ruggedness	68082.492	71275.694	8666.24	323118.582	19414
Ancestral Caloric Suit. Pre 1500	1288.789	339.432	3.335	1951.036	19414
Ancestral Ln(Distance Coast)	12.716	0.863	9.593	13.971	19414
Ancestral Ln(Distance River)	11.01	0.535	9.226	12.305	19414
Ancestral Abs. Latitude	9.297	5.586	0.017	26.373	19414
Years of Education	4.254	2.057	0	11	19414
Age	29.822	7.932	15	49	19414
Religion	2.448	1.568	0	9	19414
Wealth Index	3.405	1.397	1	5	19414
Size of the Group in the Region	1157.385	1576.861	1	5748	19414

Table S8: **Summary statistics: DHS Movers**

Variable	Mean	Std. Dev.	Min.	Max.	N
Endogamy Lev 1	0.987	0.114	0	1	9398
Endogamy Lev 2	0.947	0.225	0	1	9398
Endogamy Lev 3	0.942	0.234	0	1	9398
Endogamy Lev 4	0.934	0.249	0	1	9398
Endogamy Lev 5	0.91	0.286	0	1	9398
Endogamy Lev 6	0.888	0.316	0	1	9398
Endogamy Lev 7	0.878	0.327	0	1	9398
Endogamy Lev 8	0.873	0.334	0	1	9398
Endogamy Lev 9	0.769	0.421	0	1	9398
Endogamy Lev 10	0.693	0.461	0	1	9398
Local Malaria Stability	11.858	9.936	0	36.296	9398
Ancestral Malaria Stability	12.762	9.002	0	33.638	9398
Ancestral Tsetse Suitability	0.209	0.849	-2.474	1.46	9398
Ancestral Dengue Suitability	0.656	0.184	0.034	0.915	9398
Ancestral Pr. Genetic Distance	7.734	0.836	4.806	8.705	9398
Ancestral Precipitation	23.511	2.854	17.344	29.075	9398
Ancestral Temperature	1002.123	401.062	125.095	2738.653	9398
Ancestral Agricultural Suit. (mean)	0.509	0.18	0.015	0.983	9398
Ancestral Agricultural Suit. (std)	0.119	0.05	0.009	0.247	9398
Ancestral Elevation (mean)	787.397	585.16	6.953	2137.922	9398
Ancestral Elevation (std)	240.948	180.074	6.207	673.904	9398
Ancestral Ruggedness	73807.276	79403.178	8666.24	323118.582	9398
Ancestral Caloric Suit. Pre 1500	1309.072	332.592	3.335	1951.036	9398
Ancestral Ln(Distance Coast)	12.727	0.844	9.613	13.971	9398
Ancestral Ln(Distance River)	10.985	0.547	9.226	12.238	9398
Ancestral Abs. Latitude	9.964	5.839	0.017	26.373	9398
Years of Education	4.212	2.073	0	11	9398
Age	29.849	7.87	15	49	9398
Religion	2.438	1.552	0	9	9398
Wealth Index	3.564	1.408	1	5	9398
Size of the Group in the Region	746.795	1145.012	1	5748	9398

References

Alsan, M. (2015): "The Effect of the TseTse Fly on African Development," *American Economic Review*, 105(1), 382–410.

Ashraf, Q., and O. Galor (2011): "Dynamics and Stagnation in the Malthusian Epoch," *American Economic Review*, 101(5), 2003–2041.

Ashraf, Q., and O. Galor (2013): "The Out of Africa Hypothesis, Human Genetic Diversity and Comparative Development," *American Economic Review*, 103(1), 1–46.

Desmet, K., I. Ortuño Ortín, and R. Wacziarg (2012): "The Political Economy of Ethnolinguistic Cleavages," *Journal of Development Economics*, 97(2), 322–338.

Galor, O., and Ö. Özak (2016): "The Agricultural Origins of Time Preference," *American Economic Review*, 106(10), 3064–3103.

Garske, T., N. M. Ferguson, and A. C. Ghani (2013): "Estimating Air Temperature and Its Influence on Malaria Transmission across Africa," *PloS one*, 8(2), e56487.

Gething, P. W., T. P. Van Boeckel, D. L. Smith, C. A. Guerra, A. P. Patil, R. W. Snow, and S. I. Hay (2011): "Modelling the Global Constraints of Temperature on Transmission of Plasmodium Falciparum and P. vivax," *Parasites & Vectors*, 4(1), 92.

Gray, J. P. (1999): "A Corrected Ethnographic Atlas," *World Cultures*, 10(1), 24–85.

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," *American Journal of Tropical Medicine and Hygiene*, 70(5), 486–498.

Kraemer, M. U., M. E. Sinka, K. A. Duda, A. Q. Mylne, F. M. Shearer, C. M. Barker, C. G. Moore, R. G. Carvalho, G. E. Coelho, W. Van Bortel, et al. (2015): "The Global Distribution of the Arbovirus Vectors Aedes Aegypti and Ae. Albopictus," *elife*, 4, e08347.

Lysenko, A. J., and N. I. Semashko (1968): "Geography of Malaria. A Medico-Geographic Profile of an Ancient Disease," *Itogi Nauki: Medicinskaja Geografija*, pp. 25–146.

McCord, G. C., and J. K. Anttila-Hughes (2017): "A Malaria Ecology Index Predicted Spatial and Temporal Variation of Malaria Burden and Efficacy of Antimalarial Interventions based on African Serological Data," *The American journal of tropical medicine and hygiene*, 96(3), 616–623.

New, M., D. Lister, M. Hulme, and I. Makin (2002): "A High-Resolution Data Set of Surface Climate over Global Land Areas," *Climate research*, 21(1), 1–25.

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.

Riley, S. J., S. D. DeGloria, and R. Elliot (1999): "A Terrain Ruggedness Index that Quantifies Topographic Heterogeneity," *Intermountain Journal of Sciences*, 5(1-4), 23–27.