Kin-based institutions and economic development

Duman Bahrami-Rad[*] Jonathan Beauchamp[†] Jonathan Schulz[‡]
Joseph Henrich[§]

April 11, 2022

Abstract

What factors explain global differences in economic prosperity? While many theories have been advanced, little attention has been paid to one of the oldest and most fundamental of human institutions: kin-based institutions—the set of social norms governing lineage, marriage, post-marital residence, family organization, and an array of family obligations. We establish a robust and economically significant negative empirical association between the tightness and breadth of kin-based institutions—their kinship intensity—and economic development. To measure kinship intensity and economic development, we combine quantified ethnographic observations on kinship and genotypic measures (from which we estimate inbreeding levels as a proxy for kin marriage rates) with data on satellite nighttime luminosity and regional GDP. Our results are robust to controlling for a suite of geographic and cultural variables and hold across countries, within country at both the regional and ethnolinguistic levels, and within country in a spatial regression discontinuity analysis. We present evidence consistent with kinship intensity indirectly impacting economic development via its effects on the division of labor, cultural psychology, institutions, and innovation.

JEL classifications: D01, J12, J16, N30, Z12, Z13

---

[*] Department of Economics, Bowdoin College.

[†] Interdisciplinary Center for Economic Science and Department of Economics, George Mason University.

[‡] Department of Economics, George Mason University.

[§] Department of Human Evolutionary Biology, Harvard University.

1 Introduction

Understanding the origins of global and regional differences in economic prosperity is among the oldest endeavors in economics, tracing back through Adam Smith (1776) to the likes of Machiavelli (1531) and Ibn Kaldun (1377).[1] In recent decades, the availability of new data sources and improved approaches to causal identification have shed fresh light on the topic. Researchers have argued for the role of a suite of important factors, including climate and geography (Diamond, 1997; Hibbs and Olsson, 2004; Dell et al., 2012), disease (Sarma et al., 2019), political institutions (Acemoglu et al., 2002), colonialism (Dell, 2010), human capital (Glaeser et al., 2004), the slave trade (Nunn, 2008) and culture (Landes, 2000; Tabellini, 2010; Alesina et al., 2013; Henrich, 2020). The emerging picture is one of a complex causal network in which aspects of climate, geography, ecology and endemic disease (e.g., malaria) operate, at least partly, through their impact on the cultural evolution of technologies (e.g., writing), institutions (Spolaore and Wacziarg, 2013; Rodrik et al., 2004), social norms, preferences (e.g., fairness), beliefs (e.g. in particular gods) and aspects of psychology like individualism, patience and trust (e.g., Dohmen et al. 2015; Galor and Özak 2016; Gorodnichenko and Roland 2016; Henrich 2020; Spolaore and Wacziarg 2013). In this paper, we aim to establish a tight empirical relationship between measures of economic prosperity, assessed using satellite nighttime luminosity and regional GDP, and the intensity of traditional kin-based institutions, proxied using both ethnographic (Alesina and Giuliano, 2010; Alesina et al., 2013; Greif, 2006a; Ghosh et al., 2021) and genetic data.

Most economists are unfamiliar with the complexities of kin-based institutions, even though they are among the oldest and most fundamental of human institutions. Kin-based institutions are the diversity of ways in which societies around the world and back into history have extended and re-enforced (or suppressed) blood and affinal ties through social norms that regulate marriage, honor codes, obligations to kinfolk, post-marital residence, inheritance, corporate ownership of land and much more (Murdock, 1949). In most societies, these kin-based institutions operate over generations to continuously weave the threads of each person's most important social connections (Greif and Tabellini, 2010; Henrich, 2020; Alesina and Giuliano, 2015; Bergeron, 2020). Kin-based institutions vary along several important dimensions, but anthropologists have long highlighted the importance of kinship intensity (Walker et al., 2013): the degree to which individuals are enmeshed in broad and tight kin networks that demand their loyalty and prescribe much of their behavior. As we'll illustrate below, societies dominated by small, monogamous nuclear families tend to have low levels of kinship intensity while polygynous societies dominated by endogenous patrilineal clans tend to have rather high levels of kinship intensity. Practices like cousin marriage (Bittles and Black, 2010; Leutenegger et al., 2011), customary inheritance (Bahrami-Rad et al., 2021; Bau, 2021), post-marital residence (Bau, 2021) and polygamy (Fenske, 2015) remain important throughout the world.

A growing body of work suggests that differences in kinship intensity may influence economic growth through multiple causal pathways, through their influence on incentives,

---

[1] Scholars well back into Antiquity have also speculated on the question. In Islamic Spain for example, during the first century of the second millennium, the scholar Said contrasted his own "civilized" society to to the "black" and "white barbarians" to his south (Africans) and north (northern Europeans), respectively (Lewis, 2001). For the "white barbarians," Said suggests that latitude and climate might have left them a bit dull.

constraints, social networks and psychology. For instance, intensive kinship has already been linked to lower levels of impersonal trust, individualism, public goods provision, and effective democratic governance, as well as to higher levels of corruption, nepotism and conformity (Moscona et al., 2020; Enke, 2019; Greif and Tabellini, 2010; Schulz et al., 2019; Henrich, 2020; Alesina and Giuliano, 2015, 2011; Acemoglu et al., 2002; Schulz et al., 2018; Bahrami-Rad et al., 2021; Edlund, 2018; Bergeron, 2020). Each of these patterns has been argued to influence economic growth and/or innovation (Gorodnichenko and Roland, 2016; Algan and Cahuc, 2013; Acemoglu et al., 2019; Mauro, 1995).

In this paper, we cut through such intricacies and focus on empirically establishing a tight and robust reduced-form link between kinship intensity and economic prosperity. We also explore and discuss potential mechanisms and present new evidence suggesting that lower levels of kinship intensity may foster a broader division of labor and greater occupational specialization.

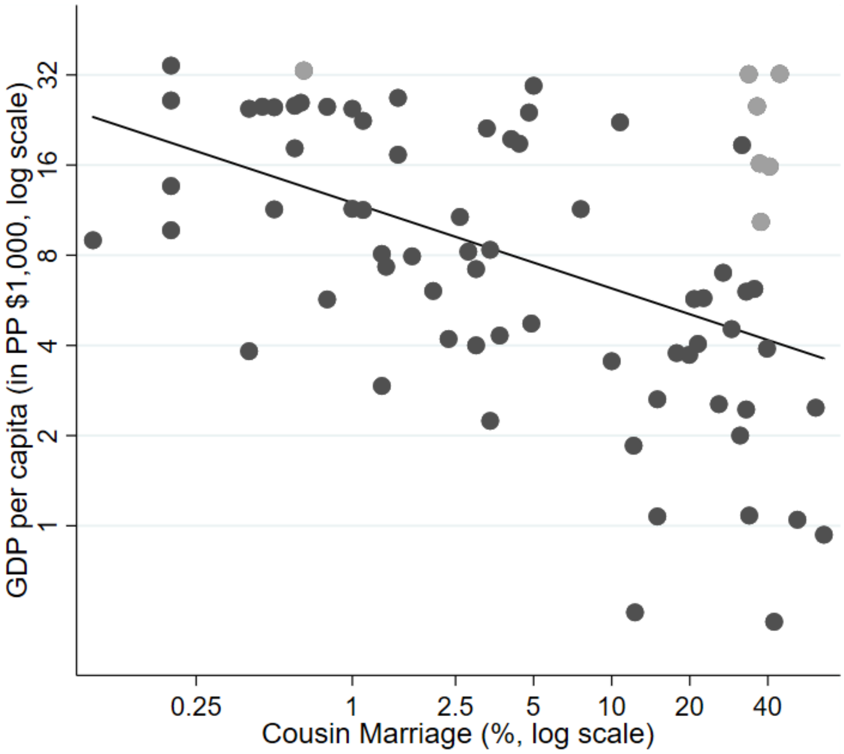


Figure 1: GDP per capita in 2000 vs. prevalence of marriage among kin (second cousins or closer) across countries. The light gray dots denote oil-rich countries with oil output above 250 barrels per day per capita in 2000. The GDP data comes from Ashraf and Galor (2013) and the cousin marriage data comes from Schulz (2022).

Motivating our investigation, Figure 1 shows the cross-country correlation between national GDP per capita in 2000 and the prevalence of marriages among kin up to and including second cousins, a simple proxy for kinship intensity. The cross-country Spearman's correlation is negative and significant ($\hat{\rho} = -0.45$, $p < 0.001$). Below, we present an array of more detailed analyses that interrogate the relationship between kinship intensity and economic prosperity.

To measure economic prosperity, we rely primarily on the logarithm of pixel-level satellite nighttime luminosity data (Michalopoulos and Papaioannou, 2013; Donaldson and Storeygard, 2016), but always hold constant the logarithm of populations density. This allows us to interpret our estimated associations as being between kinship intensity and nighttime luminosity *per capita*.[2] As a supplemental measure of economic prosperity, we also analyze regional measures of GDP per capita (Gennaioli et al., 2014).

To measure kinship intensity, we deploy an existing approach from our previous work (Schulz et al., 2019) that combines data from the Ethnographic Atlas (EA), a compilation of anthropological observations from before industrialization and European colonization based on coding ethnographies from over 1,200 societies, and global language phylogenies from over 5,000 ethno-linguistic groups (Schulz et al., 2019; Enke, 2019; Giuliano and Nunn, 2018; Bahrami-Rad et al., 2021; Kirby et al., 2016). We construct a Kinship Intensity Index (KII) that aggregates measures of societies' intensive kinship practices across five dimensions that capture preferences for cousin marriage, norms regarding polygamy, co-residence of extended families, lineage organization, and community organization.

As an alternative measure of kinship intensity, we use genetic data from the Human Origins (HO) dataset (Reich Lab, 2020) to estimate the average inbreeding coefficient for a diverse swath of populations. As far as we know, both the HO dataset and our genetic-data-based estimator of the inbreeding coefficient are novel in the economics literature. The HO dataset contains genetic data from nearly 10,000 modern-day individuals from across the world. The inbreeding coefficient is a fundamental variable in the field of population genetics that measures the relatedness of one's parents. We estimate it using genetic data with an estimator, F_{ROH}, that captures the share of one's genome that is in runs of homozygosity (ROHs)—long genomic segments where the paternally and maternally inherited DNA is identical.[3] One important advantage of this alternative measure of kinship intensity is that it is measured precisely with 21st-century genetic data and does not rely on ethnographic observations or self-reports.

We show that populations' average F_{ROH}, which we denote $\overline{F_{ROH}}$, robustly correlates with cousin marriage practices across populations (consistent with previous work—e.g., Pemberton and Rosenberg 2014; Sahoo et al. 2021) as well as with the KII and its other component measures. These relationships not only confirm the ground truth of the KII, but also imply that it captures enduring and important cultural practices—important enough to be detectable in contemporary genetic data.

Across an array of analyses, we establish a statistically significant and economically important negative association between kinship intensity and economic prosperity. We begin by estimating OLS regressions to document associations between the KII and nighttime luminosity across geographic pixels and across ethnicities, and then between the KII and regional GDP per capita across subnational regions. To test for the robustness of these associations, we present an extensive battery of OLS regressions that include country fixed effects and control for a rich range of geographic, cultural, and ecological variables.

To address concerns that unobservable omitted factors that vary smoothly across space may drive the association, we then conduct a spatial regression discontinuity (RD) analysis at the boundaries between pairs of neighboring ethnicities from the same countries, with nighttime luminosity and the KII as the dependent and explanatory variables and using geographic pixels as the units of observation. We adjust for the same suite of control variables as in our earlier OLS regressions. To thoroughly account for any factors that may vary smoothly across space, we estimate specifications that control for each geographic pixels' distance to the ethnic boundaries in a flexible way, allowing for

---

[2] In most specifications, we estimate the coefficient on log population density to be close to unity.

[3] As we explain below, though measured with genetic data, the inbreeding coefficient and F_{ROH} capture *cultural practices* and are not genetically determined.

a different coefficient on the distance-to-the-boundary term for each ethnicity in each ethnicity pair, or interacting latitude and longitude with each ethnicity pair. We estimate the RD specification in a sample that also includes pairs of neighboring ethnicities from different countries. We further verify the robustness of our spatial RD results to using pixels within different ranges of distance to the ethnic boundary, thus allowing for spillover effects of light and economic activity and for gradual changes in the faction of the population that belongs to each neighboring ethnicity at the ethnic boundary. Lastly, we estimate placebo spatial RD regressions, replacing nighttime luminosity as the dependent variable with each of a series of geographic controls, and find no association between the KII and these placebo dependent variables.

Across all of these analyses, the association between the KII and economic prosperity is robust. Further, the estimated coefficient on the KII remains remarkably stable, implying that a one-standard deviation increase in the KII is associated with a $\sim 30-50\%$ decrease in per capita economic output across the globe and a $\sim 7-11\%$ decrease within country; this holds both with nighttime luminosity and with regional GDP per capita as the measure of economic output.

Next, we examine the association between $\overline{F_{ROH}}$ estimated using genetic data from HO and nighttime luminosity across geographic pixel, in a series of OLS regressions that parallel those we ran earlier with the KII instead of $\overline{F_{ROH}}$. Again, we document a robust and economically significant association. Our estimates imply that nighttime luminosity is $\sim 30\%$ lower for an ethnicity in which everyone is the offspring of second cousins vs. an ethnicity in which everyone is the offspring of unrelated parents. (The difference in $\overline{F_{ROH}}$ between two such ethnicities is slightly smaller than the standard deviation of $\overline{F_{ROH}}$ across ethnicities.)

We then examine the causal pathways that may account for the link between kinship intensity and economic prosperity. First, since we have previously argued that historical exposure to the Western Church causally decreased kinship intensity in European ancestry societies (Schulz et al., 2019), we verify that neither historical exposure to Christianity nor European ancestry fully drive our results. Second, we examine whether reverse causality from economic development to kinship intensity could account for our results, but conclude that this is unlikely to be the case. Third, gesturing to future work, we examine possible mechanisms that could account for a causal effect of kinship intensity on economic development. We discuss the existing literature and present new analyses that provide preliminary support for pathways that involve the negative effects of kinship intensity on the division of labor and trade; on cultural psychological variables such as trust, impersonal cooperation, impartiality, and individualism; on the quality of formal institutions; and on innovation.

Related literature

This paper builds on, and contributes to, a number of interrelated lines of research within economics. First, our demonstration of a robust link between kinship intensity and economic prosperity contributes to the central questions in economic history related to the origins of the industrial revolution and the "rise of the West" (Mokyr, 2016; Clark, 2007). Greif and Tabellini, along with many historians and anthropologists, have argued that the transformation of European kinship into what has been dubbed the "European Marriage Pattern" (Hajnal, 1982), a virtually unique form of family organization built around monogamous nuclear family households, was central to emergence of the political and economic institutions that underpinned the European expansion after 1500 and eventually the industrial revolution (Tabellini, 2010; Greif, 2006b,a; Mitterauer, 2010; Goody, 1983; Henrich, 2020). More recently, Ghosh et al. (2021) take advantage of the checkered state-level imposition of laws prohibiting cousin marriage in the United States. Their analysis shows that declines in cousin marriage precipitated by the new laws fostered urbanization and economic growth in the 20th century. Our results extend these insights into the contemporary era and around the globe.

Second, a closely related literature examines the factors that drive contemporary differences in economic prosperity. By revealing a robust, and potentially causal, relationship between kinship intensity and economic prosperity, we place kin-based institutions into the cluster of other important institutions and cultural practices that have been linked to economic growth (Alesina and Giuliano, 2015; Nunn, 2012). Our focus on kin-based institutions can contribute to explaining the long-term persistence of economic prosperity (Comin et al., 2010; Edlund, 2018), the influence of both early European settlements (Easterly and Levine, 2016) and Christian missions (Calvi and Mantovanelli, 2016; Bai and Kung, 2015), and the diffusion of innovations from the technological frontier (Spolaore and Wacziarg, 2013). Unlike the commonly used measures of formal institutions such as constraints on the executive, kin-based institutions are generally vertically culturally-transmitted at the group-level and strikingly persistent across time: analyses from both economics and anthropology using diverse measures of kin-based institutions demonstrate that these cultural traits are among the most persistent (Guglielmino et al., 1995; Alesina and Giuliano, 2014; Bahrami-Rad et al., 2021). The stickiness of these institutions may contribute to the persistence of economic outcomes.

Finally, our analyses contribute to the growing literature on the role of social norms, culture, historical legacies and economic development (for overviews, see Spolaore and Wacziarg, 2013; Nunn, 2010; Michalopoulos and Papaioannou, 2018), which links kin ties to (1) particular aspects of psychology, (2) the functioning of political institutions, educational investments, (3) rates of intergroup violence, and (4) economic development (Edlund, 2018; Fafchamps, 2011; Platteau, 2000; Hoff and Sen, 2011; Henrich, 2020) Pioneering this effort, Alesina et al. (2013, 2015) have shown, using data from the World Value Survey, that the importance of close ties within the nuclear families is associated to female labor force participation, political attitudes, the importance of family businesses and, ultimately, economics prosperity. Complementing this work, Akbari et al. (2019) provide evidence that higher cousin marriage rates are associated with higher frequency of corruption across countries and also across European regions. In the Democratic Republic of Congo, using the historical locations of Christian missions, Bergeron (2020) shows that city dwellers whose home villages were closer to historical missions revealed greater moral universalism and reported social networks with more people outside their families and ethno-linguistic groups. Using global data, both Enke (2019) and Schulz et al. (2019) demonstrate a robust correlation between kinship intensity and measures of impersonal trust, moral universalism, conformity, analytical thinking, individualism and cooperation with strangers.

Some work has also linked various aspects of kin-based institutions to important outcomes which likely have direct or indirect links to economic prosperity. First, testing venerable ideas from anthropology on the origins of honor cultures (Sahlins, 1961), Moscona et al. (2017; 2020) show that segmentary lineage organizations in Africa are associated both to less trust in strangers and more intergroup violence. In Indonesia and Ghana, Bau (2021) shows that traditional norms about post-marital residence—patrilocality or matrilocality—interact with the arrival of pension systems to influence the education of either males or females. Bau's investigation also shows how pension systems undercut one of the primary function of intensive kinship, resulting in the decline of traditional kinship practices. Similarly, focusing also on the impact of different kin-based institutions in the Democratic Republic of Congo, Lowes 2018 shows experimentally that matrilineal wives are less cooperative with their spouses than patrilineal wives—as long suggested by anthropologists—but they experience less domestic violence and their children are healthier and better educated. Finally, looking both historically within Europe and cross-nationally today, van Zanden et al. (2019), de Moor and van Zanden (2010), and Carmichael and Rijpma (2017) investigate the effect of family systems on women's agency and labor market participation.

Paper structure

Our approach rolls out as follows. We begin, in Section 2, by introducing kin-based institutions, discussing how we measure their kinship intensity using both the KII and $\overline{F_{ROH}}$, and illustrating how they operate on the ground via selected ethnographies. Next, in Section 3, we describe the main data we use, emphasizing our measures of economic performance and population density and our battery of control variables. In Section 4, we proceed to our analyses, first linking the KII to nighttime luminosity across pixels and across ethnicities, then linking the KII to regional GDP per capita, and finally again linking the KII to luminosity, but across contiguous ethnicities in a spatial RD analysis. In Section 5, to complement the KII analyses, we show $\overline{F_{ROH}}$ relates to the KII and its subcomponents and then repeat the analysis linking the KII to nighttime luminosity, swapping in $\overline{F_{ROH}}$ for the KII. In Section 6, we discuss possible causal pathways, including confounding by Christianity or European ancestry, reverse causality, and mechanisms that may account for a causal impact of kinship intensity on economic prosperity. Section 7 concludes.

2 Kin-based institutions

Representing perhaps the oldest and most fundamental of human institutions, kin-based institutions are packages of social norms that govern marriage and regulate family relationships. Well back into our evolutionary past, these institutions (Murdock, 1949; Parkin, 1997) have played a central role in organizing economic production, distribution, consumption, political decision-making and social insurance, particularly for the injured, infirmed, aged, and orphaned. Kinship norms variously prescribe and prohibit particular inheritance customs (e.g., matrilineal descent of identity or land), polygamy (e.g., polygyny), cousin marriage, arranged marriage, clan membership, corporate ownership of land, household organizations (e.g., extended households) and post-marital residence (e.g., patrilocality, where the bride resides with the husband's family). The durability of kin-based institutions, along with their universality, likely arises from their anchoring in several well-established aspects of human nature (Henrich, 2016),including our inclinations for kin-based altruism, incest avoidance, and pair-bonding (Henrich, 2020).

Nevertheless, despite many similarities, kin-based institutions vary considerably across societies, having evolved culturally in response to diverse ecologies, novel technologies, new religious beliefs, and state policies (Dalton and Leung, 2014; Daynes, 2001; Fenske, 2015; Henrich, 2020; Schulz et al., 2019; Bau and Fernández, 2021; Holden and Mace, 2003; Bahrami-Rad et al., 2021; Tène, 2021). Today, for example, roughly 1.1 billion people in Africa, the Middle East, and parts of Asia reside in regions in which between

20 to 60 percent of marriages are between second cousins or closer kin (Romeo and Bittles, 2014), and cousin marriage rates are stable or increasing in some populations, including in Pakistan (Bittles, 2022), Iran (Abbasi-shavazi and Mcdonald, 2008), Oman (Islam, 2022) and Yemem (Jurdi and Saxena, 2003). Similarly, many populations still trace descent primarily through either the male line or the female line (Moscona et al., 2020; Lowes, 2018; Tène, 2021). This contrasts with Western societies, where social structure is characterized by a strong emphasis on the monogamous nuclear family, kin marriages are virtually absent, and descent is traced bilaterally through both the mother's and the father's side (Greif, 2006a; Goody, 1983; Mitterauer, 2010). Interestingly, despite the parallels with genetic inheritance, anthropological data suggest that fewer than half of societies traced descent bilaterally (Henrich, 2020).

In considering how kinship shapes people's lives, minds and societies, anthropologists have characterized cross-societal variation in kin-based institutions according to their kinship intensity (Walker et al., 2013; Walker and Hill, 2014; Henrich, 2020; Schulz et al., 2019). Intensive kinship norms foster tight, dense and overlapping relationships, which often create essentialized 'corporate groups' such as clans or lineages that are relationally isolated from other such groups (Enke (2019) calls this "kinship tightness"). Cousin marriage, for example, weaves families together into dense kin-based networks, which provide preferred and privileged partners for economic exchanges, mutual aid, insurance and political alliances. Cousins, of course, are already kinfolk, but cousin marriage reenforces and tightens these bonds with additional ties and, perhaps more importantly, helps guarantee that these relationships endure into later generations. Similarly, norms that promote co-residence, where children grow up in extended families or clans living in the same dwelling, strengthen the cohesion, interdependence and loyalty within such groups. Further, norms that establish unilineal descent, prescribe post-marital residence, favor arranged marriage, and encourage polgynyous arrangements all also intensify kinship by building larger networks of dense, overlapping and enduring kin ties.

By contrast, extensive kinship systems, such as those found among mobile huntergather populations, are characterized by marriage to non-kin (incest taboos often prohibit cousin marriage), bilateral descent, limited polygyny and flexible residence norms. Marrying and residing with non-kin creates large and more diverse kin networks, as does bilateral descent, where people trace relatedness through both their mothers and fathers. Such norms creates large, interconnected, and non-exclusive kin networks in which everyone except siblings has a unique combination of relatives.

From an economic perspective, kin-based institutions dramatically shape people's social networks, relationships, loyalties, obligations, responsibilities, incentives, constraints and, as we have argued elsewhere, their motivations and ways of thinking, feeling and reasoning (Schulz et al., 2019; Henrich, 2020). As we examine in Section 6, these differences may influence economic specialization, the division of labor, impersonal trust, exchange, distribution, political behavior and innovation as well as, ultimately, economic prosperity. Given the profound ways in which kinship intensity shapes people's lives and decisions, it is worth considering whether it can be robustly linked to economic growth.

Efforts to understand why populations vary in their kin-based institutions and the cultural evolution of these institutions are just beginning. The most important determinant of a population's current kin-based institutions are the kin-based institutions of their forebears—cultural persistence (Guglielmino et al., 1995; Bahrami-Rad et al., 2021; Alesina and Giuliano, 2014; Jordan et al., 2009)—but several ecological, economic, epidemiological, legal and religious factors have been shown to shape kinship. First, the spread of sedentary farming, with the consequent need to invest in and defend territory, likely fostered an intensification of kinship (Henrich, 2020; Jones, 2011; Ember and Ember, 1971). Focusing on China, Noblit (2021) has recently shown that both a county's suitability for paddy rice agriculture and its susceptibility to rainfall shocks (associated with typhoons) favored the gradual diffusion of lineage organizations over centuries after the year 1000 CE. Focusing on matrilineal inheritance in Africa, Tene (2021) points to a role for ecological factors that favor hoe agriculture (done by women) and those that inhibit large-animal pastoralism (done by men). Consistent with this, Holden and Mace argue that the spread of cattle across Africa led to the decline of matrilineal descent (Holden and Mace, 2003).

Considering a role for disease, Enke (2019) links kinship intensity to the ecological potential for malaria and tsetse flies (which cause sleeping sickness and kill livestock). Focusing on policy, Bahrami-Rad (2021)shows that laws in India that prohibited unilineal inheritance (effectively allowing women to inherit equally) fostered more arranged marriages to cousins (and reduced gender equality). Similarly, scholars have argued that Islam's prescriptions regarding inheritance by daughters (they get half of what sons get) fostered more marriages of daughters to their brothers' fathers' sons (patrilateral parallel cousin marriage)—a type of cousin marriage rarely observed outside of Islamic societies (Korotayev, 2015).

Finally, a number of economists, historians, and anthropologists have argued that the branch of Christianity that evolved into the Roman Catholic Church has dramatically transformed kinship around the world with its prohibitions and prescriptions regarding polygyny, cousin marriage, bilateral inheritance and other practices related to marriage and the family (Korotayev, 2003; Goody, 1983; Henrich, 2020; Mitterauer, 2010; Greif, 2006a; Schulz et al., 2019). For instance, Schulz et al. (2019) link historical exposure to the Church to the rate of cousin marriage across European regions and to kinship intensity across countries. Focusing on Africa, Fenske (2015) and Bergeron (2020) show that distance from historical Christian missions predicts greater polygyny and stronger kin ties, respectively.

2.1 Measuring kinship intensity

Our analyses rely on two main measures of kinship intensity. The first measure, which we use in most of our analyses, is the Kinship Intensity Index (KII) from Schulz et. al. (2019). The KII is an omnibus measure of the overall strength of kin-based institutions constructed using anthropological data and available for nearly 1,000 societies in the Ethnographic Atlas (EA) (as curated and expanded at D-Place (Kirby et al., 2016)). To complement this ethnographic measure and verify the robustness of our main results to using a contemporaneous, on-the-ground measure of kinship intensity, we used genetic data from the Human Origins (HO) dataset (Reich Lab, 2020) to compute the average inbreeding coefficient for a few hundred populations.

The Kinship Intensity Index (KII)

The Kinship Intensity Index (KII), which we developed in previous work (Schulz et al., 2019), is calculated using quantified anthropological observations from over 1,200 populations from the Ethnographic Atlas and, for a version of our analyses, is extended to populations around the world using language phylogenies.[4] With the average observation

---

[4] We used the same approach to calculate the KII as Schulz et al. (2019), but used data from the

occurring around 1900 CE, the EA is based on ethnographies written by anthropologists aiming to reconstruct people's lifeways prior to European colonization, global market integration and industrialization.[5] The KII is thus a "deep" historical measure that aims to capture enduring norms and practices before modernization.

The KII aggregates five sub-indicators that capture key dimensions of kin-based organization:

1. *Cousin marriage preference* captures the intensity of the norms about marrying cousins. Preference for cousin marriage inhibits the formation of extensive ties among previously unconnected families or clans, encourages the creation of additional links among already related families and households, and increases the genetic relatedness of family members.

2. *Polygynous marriage norms* permit men to marry multiple wives. This results in larger and more extended households and introduces social and economic interdependence among co-wives and half-siblings. At a societal level, polygynous marriage norms result in fewer fathers, larger reproductive skew, and greater genetic relatedness.

3. *Co-residence of extended families* captures the degree to which several generations of a family, each with their own spouses and children, co-reside. Such residential norms create stronger emotional bonds and greater economic interdependence among the co-residing individuals. This contrasts with the neolocal nuclear family, where only the two spouses with their children live together, separate from other relatives.

4. *Lineage organization* captures the social norms governing descent and identity. In societies with unilineal descent, people reckon descent and social identity primarily or entirely through either their mother's or father's side. The exclusive membership on one side determines social identity and increases cohesion and interdependence within the lineage. This contrasts with bilateral descent where membership is non-exclusive and everyone except siblings have a unique combination of relatives, resulting in more diverse and diffuse kin-networks and lower kinship intensity.

5. *Community organization* captures whether extended family and clan members reside within the same localized area of a settlement (e.g., a neighborhood), and whether there is community-level endogamy (e.g., people can only marry co-villagers). Localization decreases the interaction with outsiders while endogamous marriages mean that communities form denser clusters (since no outsiders from different villages join the community through marriage), thereby increasing kinship intensity.

As we further describe in Section 3.3, we matched the EA data (including the KII sub-indicators) to languages from the Ethnologue (Gordon, 2005; Lewis, 2009), a comprehensive map of the world's languages (these languages are in turn associated with "country-ethnicities").[6] To compute the KII for each Ethnologue language, we first stan-

---

latest versions of the EA (Kirby et al., 2016), Ethnologue (Gordon, 2005; Lewis, 2009) and Glottolog (Hammarström et al., 2016).

[5] 8 of the 1,291 societies in the EA were coded based on pre-1500 CE historical observations, so we excluded them from our analyses.

[6] We used two alternative matching methods: the direct matching method and the language-tree matching method. When using the language-tree matching method, we imputed missing observations for the cousin-marriage preference and co-residence-of-extended-families sub-indicators—see Schulz et al. (2019) for details.

dardized the five sub-indicators using their means and standard deviations across the EA societies; then, we took the average of the language's five matched sub-indicators and standardized it using its mean and standard deviation computed across the EA societies. The resulting KII is defined for the Ethnologue's languages but is standardized based on the EA societies.[7] Figure 2 shows the distribution of the KII across the ethnicities around the world, and Appendix A.1 provides further details on the construction of the five sub-indicators and of the KII.

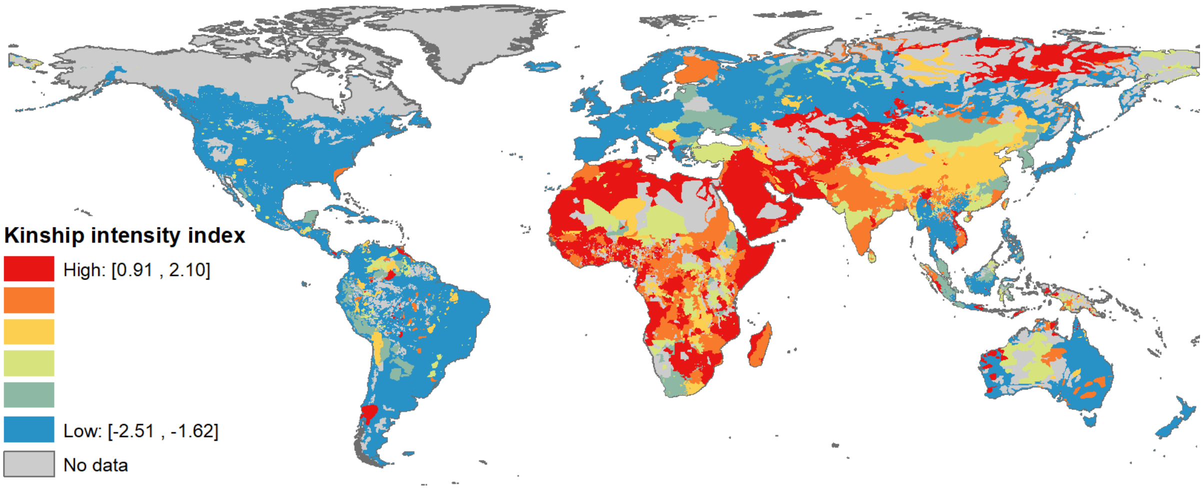


Figure 2: Distribution of the KII around the world, for the 2,352 ethnicities matched with the language-tree matching method (described below).

Although Schulz et al. (2019) selected the five sub-indicators on purely theoretical grounds—aiming to operationalize the anthropologists' concept of kinship intensity—these measures turn out to be positively correlated across the EA societies with nonmissing data (see Appendix Table B.1.1), with the first principal component accounting for 35 percent of the variation.

The inbreeding coefficient (F)

Our second measure of kinship intensity, the inbreeding coefficient (denoted F), is a key variable in the field of population genetics that can be estimated using genetic data. F is an individual-level measure of the relatedness of one's parents, and so a society's average F should in principle correlate with its kinship intensity and in particular with it rate of endogamous marriage. As we describe in more detail in Section 5, we estimated F for contemporary individuals in the Human Origins (HO) dataset (Reich Lab, 2020)—a dataset of genotyped individuals from populations around the world—and then estimated the mean F across the HO individuals matched to each national ethnicity from the Ethnologue.

Because the KII ultimately derives from ethnographic observations, often made over the course of a year or so of field research, one might worry that it represents merely ideal behavior (vs. ground truth of how people live their lives), or a snapshot in time of an otherwise unstable or rapidly changing pattern of kinship (Leach, 1964). Using F

---

[7] We standardized the sub-indicators and the KII based on their means and standard deviations across the EA societies, rather than across the Ethnologue languages, because the different matching methods we use lead to different sets of languages being matched to the EA, and some of these languages correspond to very small groups; standardizing based on the EA societies allows to us to consistently interpret the coefficients we estimate across our analyses.

measured among contemporary individuals as our alternative measure of kinship intensity allows us to verify that the KII captures stable cultural practices. As we report below (see Table 6 and Appendix Tables B.5.2-5.3), F is positively associated with the KII and particularly strongly associated with the cousin-marriage-preference sub-indicator, supporting the view that the KII does indeed capture enduring cultural practices. Consistent with this, Pemberton and Rosenberg (2014) and Sahoo et al. (2021) also report positive associations between F and cousin marriage preferences or practices.

We emphasize that, although we estimate it using genetics data, F is not genetically determined. Rather, F provides us with a proxy that should track marriage practices—cultural traits that contribute to kinship intensity. To the extent it captures the deep-rooted cultural practices, however, F should be rather stable through time.

2.2 Kinship intensity on the ground

To illustrate how kinship intensity shapes individual decision-making and societies, let's take a closer look at three populations with high, medium and low kinship intensity based on both their KII and F values.

2.2.1 The Marri Baluch ($KII = 2.10$; $F = 0.062$)

Exemplifying a society with intensive kinship, the Marri Baluch are agro-pastoralists in Pakistan who live enmeshed in a hierarchy of patrilineages in a mountainous region lying about midway between Islamabad and Karachi (Pehrson, 1977). Traditionally, the nomadic Marri Baluch have relied primarily on herding sheep and goats, though some households (particularly elites) maintain mud houses in villages and engage in cereal farming, growing mostly wheat. The ownership of both land and animals is communal, though animals can be individually owned, and both forms of wealth are transferred corporately by the inheritance of a Marri identity from one's father. All pasture land is held in common (among Marris) but agricultural land is reapportioned every decade or so among large patrilineal groups according to the number of Marri men in each. Interestingly, although paternal kinship is the central factor in economic, political and social life, getting "counted" for land apportionment depends also on not having a low-caste or enslaved mother.

Marriages are typically arranged for adolescent girls, who are "sold" by their fathers for a brideprice (e.g., 80 sheep) to other men. There are strong biases toward marriages within the same small patrilineage (*wari*), a pattern reflected in the lower brideprices associated with such marriages. In one survey, 30 percent of marriages were between patrilateral parallel cousins—i.e., the children of two brothers—and another third were among other patrilateral relatives. In total, 72 percent of marriages were among kin and no marriages occurred outside of the Marri Baluch–none outside the ethnic group. Compared to many societies with intensive kinship, incest taboo are not particularly constraining on spousal choice: beyond primary female relatives, men are only tabooed from having sex with the wives of their fathers, father's brothers, brothers, sons and brothers' sons. Polygynous marriage is permitted up to four wives following Islamic law, though the relative equality among most men means this is largely limited to political elites (whose positions often depend on paternal succession).

Curiously, inherited political offices among the Marri have no customary power to tax except for the purposes of paying brideprices for wives—thus, taxation facilitates elite polygyny, allowing such men to weave larger kin networks through marriage. In one study, only 5 percent of Marri households were polygynous at any one time. Upon marriage, new wives move to live with their husband's family: post-martial residence was strongly patrilocal. Women are essentially owned and controlled by either their fathers or husbands. Economically, labor is primarily organized along kinship lines, with a division of labor by age and sex. Locally, decision-making power is vested according to age, sex and lineage position, so fathers dictate to sons, husband to wives, and elder brothers to younger brothers and sisters. The ability of households to protect their domesticated animals and women from theft and kidnapping depends on the honor and reputation (for revenge) of the patrilineage. The honor and shame of men in the same patrileage is intertwined, so a son who acts cowardly stains the honor of his father and brothers (and other more distant paternal relatives to a diminishing degree).

This package of social norms creates substantial economic, social and political interdependence among relatives, especially among close patrilineal men, but also creates zero-sum competitions for farmland, pasture and mates. Several customs help bind together the small patrilineages that form the core of economic life. For example, the extended incest taboos, by applying to precisely the women with whom a man might live together with in a nomadic camp, reduce sexual competition and mitigate concerns that men have about their wives engaging in adulterous affairs (which are rampant). Similarly, unlike many societies with cousin marriage, the preference for endogenous unions within the patrilineage effectively tightens the bonds within the *wari*, though at the same time it forgoes an opportunity to forge marital alliances with other patrilineages or beyond. The combination of incest taboos and patrilineal endogamous marriage further reduces sexual competition within the *wari* and residential camps because it creates a situation in which all women are tabooed for all men except for their wives. Such a situation is particularly important in the nomadic herding context where men often leave their households for weeks at a time for herding, trading, raiding and wage labor.

These kin-based institutions create a tight circle of trust and sense of obligation that declines rapidly outside the *wari* and residential camp. Close matrilineal and affinal ties are recognized, and do matter, but they carry none of the sense of closeness or "duty" found in patrilineal relationships. The divisions created by kinship norms are reflected in frequent contrasts between *jind* ("one's own"), which refers to patrilineal relatives (*wari*), *Azziz*, which captures "kinfolk" (including in-laws), and *seyyal*, referring to strangers of equal status (e.g., the Pashtun or even unrelated Marri). To get a sense of how people think about these different categories, consider the following comment from a Marri regarding affines (from Pehrson (1977)):

> "When you become affinally related to someone, then it is God's command not to do badness or rottenness to them. If they are strangers (*seyyal*), then there is no duty, it does not matter if you steal or fight. But when they become affines, then there is law... There are really no definite duties. But if you are going to do meritorious acts to someone, then you should do it to your affines."

Distrust of strangers pervades many economic interactions among the Marri Baluch. For example, managing farming and herding in small camps inevitably creates labor shortages. To hire a shepherd, a man first goes to his poorer patrilineal relatives, next to the matrilineal relatives, and finally to his affines. If he exhausts these labor sources, he reluctantly turns to strangers, but anticipates problems. As one man explained Pehrson

(1977): “I usually give a shepherd three chances—on theft, lying, etc, and then send him away. For adultery, however, he goes immediately.” Of 35 shepherds in 5 camps, only 3 were unrelated to the owners. Similarly, when conducting trading, Marri men rely on a network of *bradirs* as they move among camps, villages and towns. *Bradirs* represent an institutionalized friendship of sorts and provide primary trading partners, hosts and sources of information. Crucially, *bradirs* are inherited from father to son. Men report not trading in places where they lack *bradirs*—so trade hinges on a network of enduring personal relationships.

2.2.2 The Kurukh ($KII = -0.18$; $F = 0.016$)

Moving to a society with intermediate levels of kinship intensity, the Kurukh or Oraon communities (Roy, 1915) of the Nagpur Plateau (northeastern India) have kin-based institutions that are similar to the Marri Baluch's in important ways, but that have over time been weakened or replaced by non-kin-based institutions. Like Marri Baluch, the Kurukh are patrilineal and patrilocal (brides move to live with their husbands), marriages are arranged, and men maintain patriarchal authority over their wives and children. Traditionally, the Kurukh were organized in patri-clans, but since moving from nomadic pastoralism into settled agriculture, including sustained contact with Hindu communities, Kurukh clans began to operate primarily as exogamous marriage units, considering it incest to marry someone from the same clan. Notably, such clan exogamy is the opposite of the tight clan endogamy preferred by the Marri Baluch. In addition to clan exogamy, the Kurukh also taboo sex or marriage among couples known to be related within three generations—so, anyone who shares a great grandparent is forbidden (thus, second cousins are taboo, but third cousins are not). However, ethnographic data suggest that families possess rather short genealogical memories so some second cousin marriages do occur. Similarly, both polygynous marriage and communal property were probably important among the herding and hunting ancestors of the Kurukh, but monogamous marriage and privately owned land became universal when they became agriculturalists.

Kurukh also possess institutions and public rituals for establishing formal, life-long, friendships. Like the Marri Baluch, new wives arrive from other communities and move into their husband's villages. Usually, these young women do not know anyone in their new home since wives arrive from many villages. However, unlike the Marri Baluch, the Kurukh have rituals that nurture a set of kin-like relationships among women. For example, approximately every three, the word goes out from the female elders that every woman must select a *sahia* or special village friend, from among the other married women in the village. Women may renew a prior *sahia* or add to their social network. At a communal ceremony, the ritual friends publicly greet each other, and begin a series of ritualized and reciprocal exchanges of gifts, food, conversation and fellowship, including a visit to the goddess Devi. Their sacred friendship establishes life-long bonds that link not just the two wives, but also their entire (often unrelated) families.

Further, the Kurukh share their villages with other ethnic groups, including Hindus. Economically, the Kurukh depend on village norms that govern their interactions with other ethnic groups (castes), who engage in a variety of occupations that support the Kurukh's farming efforts. Disagreements among households are adjudicated and villages policies set by a pluralistic and largely democratic council of elders, which include members from across the village. There is also a higher council, with representatives from many villages, that addresses inter-village issues. These councils appear to be a retrofit version of the supra-clan institution that once organized the Kurukh's ancestors, modified to draw representatives from either diverse ethnic/caste groups or different communities (instead of different clans).

Kin-based institutions play important roles for both the Marri Baluch and the Kurukh, but among the Marri Baluch kinship norms weave political, social and economic interdependence around a tightly bound patrilineage. Meanwhile, though the Kurukh have clans and authoritarian fathers, their incest taboos compel marriages to socially distant Kurukh families, and other non-kin-based social norms govern economic and political interactions with both other Kurukh and non-Kurukh villagers. Obviously, neither our KII or F measures directly tap all of these complexities—such as the ritualized friendships among Kurukh wives—but when non-kin-based institutions exist to support interaction, they often replace or supplant elements or reproductive events that are captured by our measures. Kinship norms that encourage, for example, economic interdependence or physical security among kinfolk tend to foster kin-based community organizations, alliances fashioned by cousin marriage, or extended families in co-residence. When other non-kin-based institutions take up these functions, these kin-based practices generally deteriorate.

2.2.3 The English ($KII = -2.14$; $F = 0.007$)

The Marri Baluch's and the Kurukh's more intensive kin-based institutions contrast with those found in European and European-descent societies, such as the English, who are characterized by love-based marriages (but often with taboos on cousins) that form small, monogamous nuclear families in which new couples reside neither with the bride's or the groom's families but establish a new residence. Descent is not a source of identity and is traced roughly equally through both mothers and fathers. With such tiny, ephemeral families, individuals must necessarily build their own network of friends and partners and seek out voluntary groups for economic production, religious devotion, and political activity. By the High Middle Ages in parts of Europe, researchers have argued that the appearance of these forms of low intensity kinship fostered the proliferation of charter towns, guilds, universities and monastic orders as well as the expansion of impersonal trade and commerce (Henrich, 2020; Greif, 2006a,b; Greif and Tabellini, 2010, 2017).

3 Data

This section describes the data (other than our measures of kinship intensity) we used in our main analyses, including our measures of economic performance and our control variables. It also outlines how we matched data from various sources. Appendix A.1 provides a more detailed description of all the variables and data sources, and Appendix Tables B.2.1, B.3.1, B.4.1, B.5.1, and B.6.1 provide summary statistics for all the variables.

3.1 Measuring economic performance

In most of our analyses, we rely on nighttime luminosity data from satellites to measure economic prosperity at the micro-level (pixels). Because we are interested in *per capita* levels of economic development, we control for log population density in these analyses.[8]

---

[8] Specifically, we regress the logarithm of nighttime luminosity on the KII (or average F) and log population density (and other controls). If the coefficient on log population density were unity, this

To verify the validity of our results with the luminosity data, in some analyses we use regional GDP per capita as an alternative measure of economic performance.

3.1.1 Satellite nighttime luminosity data

Now commonly used in the literature, satellite nighttime luminosity data (also often referred to as light density data) has been shown to be a good proxy for economic development (Michalopoulos and Papaioannou, 2018; Donaldson and Storeygard, 2016). Importantly for our purposes, nighttime luminosity data is available at fine scales, which allows us to aggregate and match it to the geographic areas peopled by different ethnolinguistic groups for whom kinship intensity data is available.

Following Henderson et al. (2018), we use the 2010 Global Radiance Calibrated Nighttime Lights data.[9] The raw luminosity data is presented as a grid of pixels with dimensions of 30 arc-seconds ($\frac{1}{120}$ of a degree) by 30 arc-seconds, or approximately one square kilometer at the Equator.[10] We aggregate the original pixels (which we will henceforth refer to as "subpixels") to larger pixels of size 0.125 ($\frac{1}{8}$) degrees x 0.125 degrees ($\sim$ 191 square kilometers at the Equator) by taking the mean of the nighttime luminosity across the 225 original subpixels. This aggregation mitigates the problem of overglow of light in a pixel due to light emanated from nearby pixels and helps mitigate concerns about spatial correlation at finer scales. This gives us a sample of 926,864 pixels around the globe.

Next, we apply several filters to prepare the luminosity data for our analyses. First, we drop all pixels that do not cover land within the borders of a country, and thus all pixels in large bodies of water. Some of the remaining pixels after that step are partially covered by water or permanent ice. Luminosity is recorded as zero for subpixels that fall on water or permanent ice, so we normalize the pixels' mean luminosity by dividing it by the fraction of their component subpixels that are covered by land.[11]

Second, and relatedly, night lights may appear brighter than they are over water or ice-covered areas. We largely avoid this blurring issue by dropping the pixels for which more than 25 percent of the area is covered by water or permanent ice ($\sim$ 2.5% of the remaining pixels at this stage). We also drop pixels through which a coastline passes ($\sim$ 3.5% of the remaining pixels).

Third, around 70% of the remaining pixels emit too little light and are coded as zero. Following Henderson et al. (2018), we consider this a censoring problem since the lowest nonzero values are considered noise and generally recoded to zero in the initial data processing. Assuming that all pixels emit some positive amount of light, we assign

---

would be equivalent to regressing log nighttime luminosity *per capita* on the KII (and other controls). In fact, in nearly all the regressions we report below with luminosity as the dependent variable, the estimated coefficient on log population density is close to unity. We further discuss this below in Section 4.1.

[9] These data are available online at `https://ngdc.noaa.gov/eog/dmsp.html`. Other, uncalibrated versions of the data use a strong amplification to detect low levels of light, which can saturate measurement in the most brightly lit pixels such as those in large cities, leading to top coding of those pixels. The radiance-calibrated data we use combines a high-amplification regime for low-light pixels and a low-amplification regime for more brightly lit pixels, thus removing all top coding (Henderson et al., 2018).

[10] There are 3,600 arc-seconds in a degree. Pixel size varies with latitude: the higher the latitude, the larger the pixel. This will not be an issue in our analyses since the luminosity data measure the *density* of light (i.e., the light emitted from a pixel divided by the pixel's area).

[11] This is similar to the approach used by Henderson et al. (2018).

the lowest nonzero value (0.013) from across all the pixels to the pixels that are coded as zero. Since we will take the logarithm of nighttime luminosity, this step allows us to include these recoded pixels in our analyses.

Fourth, we drop pixels with zero population density ($\sim$ 9.7% of the remaining pixels) and pixels where one or more of the geographic control variables (which we include in our main specifications) are missing ($\sim$ 0.5%).

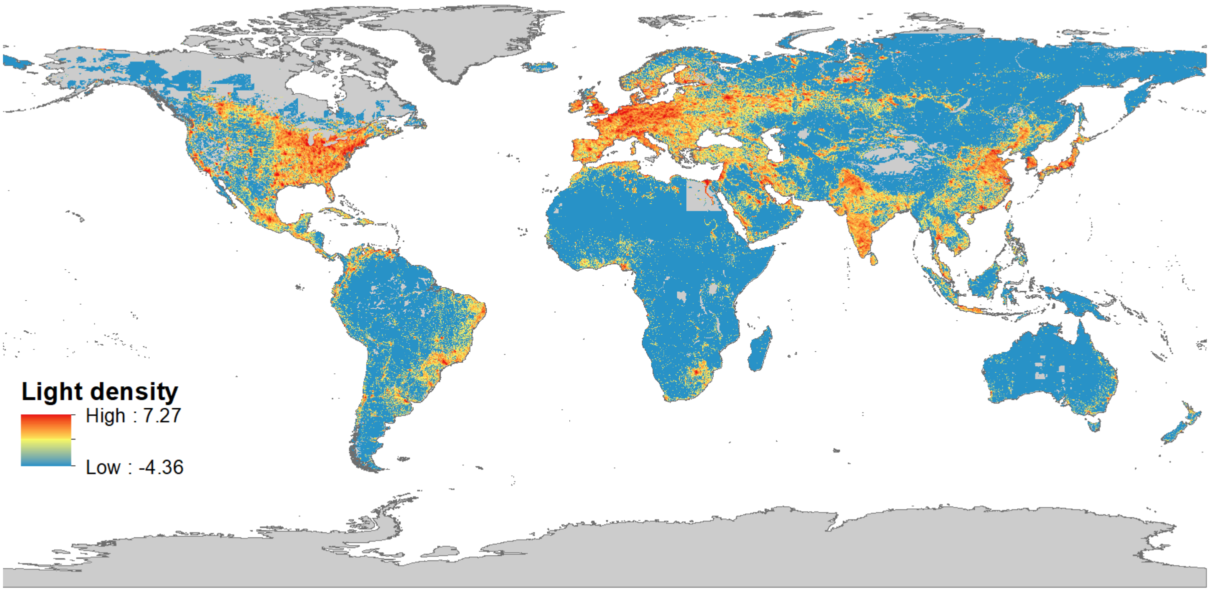


Figure 3: Distribution of log nighttime luminosity across the world. Unpopulated regions are in gray. In some areas or countries, such as Egypt, population data is available at a fine scale and unpopulated areas (like deserts) are coded as such; in other areas, population data is more heavily smoothed.

Applying these filters leaves 783,525 pixels in our sample. Figure 3 displays the variation in log nighttime luminosity across these pixels around the world. Appendix Figure B.1.1 shows histograms of nighttime luminosity, but without the pixels that were recoded with the lowest nonzero value ($\sim$ 0.013). As can be seen, the log transformation reduces skewness. Therefore, and following the literature (Michalopoulos and Papaioannou, 2018), in the rest of the paper we will use the natural logarithm of nighttime luminosity.

3.1.2 Population data

We use fine-grained population density data for the year 2010 from the Gridded Population of the World (GPW), adjusted to the 2015 Revision of the United Nation's World Population Prospects (Center for International Earth Science Information Network (CIESIN) at Columbia University, 2016). The GPW assumes uniform population density within statistical units. Therefore, population estimates are more heavily smoothed in countries with lower statistical capacity and in more sparsely populated regions. The GPW population density data is reported at a 30 arc-second resolution. We aggregate the data to 0.125-degree pixels by taking the mean across component subpixels, and then take the natural logarithm of population density for each pixel.

Although population density is itself often used as a proxy for economic development, our primary analyses all focus on the relationship between kinship intensity and nighttime luminosity while holding population density constant. As illustrated in Appendix Figure B.1.2, log population density has a convex relationship with log nighttime luminosity. The relationship is flat when population density is lower than 1 person per square kilometer (i.e., log(population density) < 0), but increasing and approximately linear at higher levels of population density. To avoid large variation in log(population density) driven by minor differences in population density across low-population-density pixels, we recoded to 1 the population density of all pixels whose original population density was less than 1 person per square kilometer.[12]

3.1.3 Regional GDP per capita data

To verify that our analyses of the association between kinship intensity and nighttime luminosity are robust to using a more direct measure of economic development, we also analyze data from Gennaioli et al. (2014) on regional per capita GDP (in Section 4.2). The data includes GDP per capita (in constant 2005 PPP dollars) for 1,528 regions in 83 countries between 1950 and 2010. We drop regions that are too small (consisting of less than five pixels) and end up with the sample of 1,452 regions. Appendix Figure B.3.1 shows the variation in per capita GDP across these regions.

3.1.4 Nighttime luminosity, population density, and regional GDP per capita

We conducted a simple exercise, following Henderson et al. (2018), to verify that global variation in nighttime luminosity reflects not just variation in population density, but also differences in income per capita. We aggregated our luminosity and population data to the level of subnational regions and, using Gennaioli et al's regional GDP per capita data for the year 2010 (the latest year available), regressed log nighttime luminosity on log population density and then log regional GDP per capita. Without country fixed effects, the R^2's from regressing log luminosity on log population density, then on GDP per capita alone, and finally on both, are 0.530, 0.338 and 0.839, respectively. When the data are demeaned by country, the corresponding R^2's are 0.819, 0.112, 0.856. These results highlight that much of the variation in nighttime luminosity across world regions is predicted by both population density and GDP per capita; within countries, much of the variation in luminosity is predicted by variation in population density, though some residual variation is also predicted by GDP per capita.

3.2 Control variables

Throughout our analyses, we include specifications with controls for geographic, ecological and cultural variables, most of which derive from a battery of potentially relevant factors based on prior research. Our baseline set of "geographic controls" includes temperature, precipitation, agricultural suitability, absolute latitude, elevation, ruggedness, distance to coast, and distance to the nearest river or lake. These variables are important for agriculture and trade and are associated with worldwide and within country variations in economic development and nighttime luminosity Henderson et al. (2018). The raw data for these variables come from various sources and are reported at different scales ranging

---

[12] For comparison: Mongolia is the least densely populated country in the world, with 2.1 people per square km.

from $\frac{1}{120}$- to 0.5-degree pixels (see Appendix Table A.1 for details). For the variables that are defined at a smaller scale than 0.125 degrees, we convert the data to 0.125-degree pixels by taking the mean across component subpixels. For variables that are defined at a larger scale than 0.125 degree, all 0.125-degree pixels that fall inside a larger pixel receive the same value.

We also verify the robustness of our results to controlling for a pixel-level measure of ecological suitability for malaria (Kiszewski et al., 2004). Malaria prevalence positively correlates with kinship intensity ($\hat{\rho} = 0.406$, $p < 0.001$, $n = 713$ country-ethnicities), and may foster the cultural evolution of higher levels of kinship intensity (Enke, 2019).

Culturally, we verify the robustness of our results by controlling for each ethnicity's subsistence activities and political hierarchy. To control for subsistence economic activities, we include a set of variables from the EA that measure the fraction of each ethnicity's economy that depends on gathering, hunting, fishing, animal husbandry, and agriculture (see Appendix A.1 for details). The EA includes five variables—one for each mode of subsistence—and we adjusted these so that they sum to 1 for each ethnicity and omitted the variable for gathering from the regressions. As discussed above, a population's mode of subsistence likely influences its KII. For instance, anthropologists have long argued that mobile hunter-gatherers culturally evolved extensive kinship as a mean of social insurance against environmental shocks with low spatial autocorrelation. "Political hierarchies" is a cultural variable from the EA that measures the number of levels of jurisdictional hierarchy for each ethnicity (Michalopoulos and Papaioannou, 2013). This variable correlates negatively with the KII ($\hat{\rho} = -0.190$, $p < 0.001$, $n = 678$ ethnicities) and has previously been linked to nighttime luminosity, so it could confound our results.

3.3 Matching data

To combine the various datasets whose data we analyzed, we first matched ethnographic variables from the EA (including the KII) and genetic variables from the HO dataset (including the inbreeding coefficient) to languages from the 23rd edition of the Ethnologue (Gordon, 2005; Lewis, 2009). We then matched the Ethnologue languages to geographic pixels and, for the analyses at the country-ethnicity (defined below), region, or country level, we then matched the pixels to the country-ethnicities, regions, or countries.

To match the EA variables, we followed the method developed by Giuliano and Nunn (2018) and matched the language spoken by each EA society to languages in the Ethnologue (each EA society is associated with a single language). The Ethnologue maps the geographic boundaries within which each contemporary language from around the world is spoken. We refer to the (ethnolinguistic) group that speaks a given language within a given country as an "country-ethnicity" and to the area where the country-ethnicity lives as its "country-homeland". (Thus, in our terminology, a language that is spoken in multiple countries is associated with multiple country-ethnicities.) The homelands of separate country-ethnicities often overlap, but we exclude overlapping areas from our analyses.

We employed two different matching methods. With the direct matching method, we matched the language spoken by each EA society to the exact same Ethnologue language (and dropped unmatched languages). With the language-tree matching method, we followed Bahrami-Rad et al. (2021) and, for each EA variable, we matched each Ethnologue language to the linguistically closest EA society that speaks a language within the same language family and with nonmissing data for the variable. (If no such society exists, that variable was coded as missing for that language.) We report the analyses using the direct matching method in the main text and those using the language-tree matching method in the Appendix.

For the analyses using the inbreeding coefficients (F), we matched each population from the Human Origins (HO) dataset to a single Ethnologue language (and to the single or multiple country-ethnicities associated with that language). This is similar to the direct-matching method described above for the EA data, but details of the two procedures differ because the HO data typically does not indicate the language spoken by each population. Appendix C.2 provides more details.

Next, we used a shapefile provided by the World Language Mapping System (World GeoDatasets)[13] to match pixels to Ethnologue country-ethnicities. We dropped pixels that were (1) matched to more than one country-ethnicity, (2) fell on the boundaries of a country-homeland, or (3) matched to a country-ethnicity with a very small country-homeland comprising less than five pixels. Appendix Tables B.2.1 and B.6.1 show summary statistics for the resulting samples of pixels.

Finally, to match pixel-level data to subnational regions or countries, we computed the population-weighted mean of each ethnographic variable as well as the simple mean of each geographic control across the pixels in each region or country. We then dropped regions or countries for which the pixels with nonmissing KII data accounted for less than 75% of the population. For the analyses at the country-ethnicity level, each country-ethnicity was assigned its matched EA variables and we took the simple mean of each geographic control across the pixels in each country-homeland.

Appendix A.2 provides additional details on the matching process.

4 Kinship intensity and economic development

We now examine the association between our primary measure of kinship intensity, the KII, and economic development. In Section 4.1), we first estimate a battery of OLS model specifications by regressing nighttime luminosity on the KII. In Section 4.2), we perform a parallel set of analyses by replacing our luminosity measure with regional GDP per capita measures. In Section 4.3, we then conduct a spatial regression discontinuity (RD) analysis of the association between luminosity and the KII.

4.1 Kinship intensity and nighttime luminosity

We begin by examining the relationship between nighttime luminosity and the KII, focusing on the sample that includes the country-ethnicities matched with the direct matching method. Figure 4 shows a negative relationship between the KII and the logarithm of mean nighttime luminosity (calculated across each country-ethnicity's pixels) after partialling out log population density ($\hat{\rho} = -0.202$, $p < 0.001$, $n = 2,352$).

To more thoroughly examine the association between luminosity and the KII, we adopted the following baseline specification, with geographic pixels as the unit of observation:

$$L_{i,e,c} = \alpha + \beta KII_e + \delta \log(P_{i,e,c}) + \gamma X_{i,e,c} + \lambda_c + \theta V_{(i),e} + \epsilon_{i,c}, \tag{1}$$

---

[13] The shapefile is available at worldgeodatasets.com/language

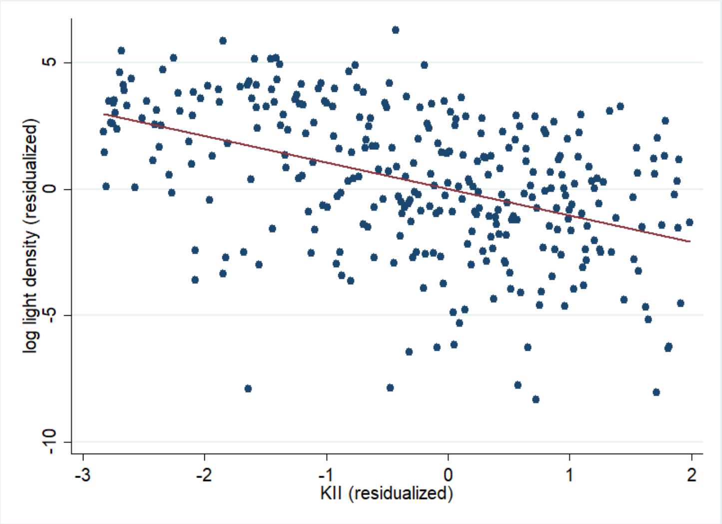


Figure 4: Log nighttime luminosity vs. the KII across country-ethnicities, both residualized on log population density.

where $L_{i,e,c}$ is the natural logarithm of the nighttime luminosity of pixel i in ethnicity e's homeland in country c. KII_e is the KII of country-ethnicity e, which takes the same value for all the pixels in e's homeland; $P_{i,e,c}$ is the pixel's population density; $X_{i,e,c}$ is the vector of geographic variables that includes temperature, precipitation, agricultural suitability, absolute latitude, elevation, ruggedness, distance to coast, and the distance to nearest river or lake; λ_c denotes continent or country fixed effects; and, $V_{(i),e}$ includes additional controls (defined at the pixel or ethnicity level, in some specifications).

As mentioned earlier, we control for log population density because we are interested in per capita levels of economic development. If we subtract $\log(P_{i,e,c})$ from both sides of equation (1) and let $\delta = 1$, the left-hand-side becomes the logarithm of luminosity per capita. We estimate δ as a free parameter to allow the luminosity-population elasticity to differ from unity (Michalopoulos and Papaioannou, 2018).[14] As it happens, our estimates of δ are close to 1 in nearly all our regressions with luminosity as the dependent variable (including those in other sections reporting other types of analyses), and so we can interpret the β coefficient on the KII in equation (1) as capturing the association between the KII and the logarithm of luminosity *per capita*.

Throughout, we account for spatial correlation in the data by clustering standard errors at the country level in most specifications. We also verify the robustness of our results to clustering at the level of the language families (based Ethnologue) and at both the country and language family level using two-way clustering (Cameron and Miller, 2015).

Table 1 reports the results of specifications that include various subsets of the covariates in equation (1). Column 1 reports a regression of log luminosity on the KII and log population density only. The KII's estimated coefficient, $\hat{\beta}$, is -0.512, indicating that a one-standard deviation increase in the KII is associated with a $\sim$ 40% $(= (1 - e^{-0.512}) \cdot 100\%)$ decrease in luminosity, as well as in luminosity per capita. (In all columns of the table, the estimated coefficient on log population density is close to and not statistically different from unity.) In column 2, when we add the geographic

---

[14] Even if the true luminosity-population elasticity were unity, the population density data is noisy, so the elasticity of luminosity with *measured* population density may be lower than unity.

Table 1: The KII and nighttime luminosity: OLS estimates

	Log nighttime luminosity								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
KII	-0.512***	-0.420***	-0.136***	-0.110***	-0.066	-0.085**	-0.101***	-0.110***	-0.110***
	(0.143)	(0.129)	(0.046)	(0.024)	(0.045)	(0.041)	(0.025)	(0.024)	(0.024)
Log population density	0.985***	0.992***	1.084***	1.067***	1.066***	1.066***	1.063***	1.067***	1.067***
	(0.064)	(0.051)	(0.071)	(0.061)	(0.063)	(0.063)	(0.061)	(0.104)	(0.106)
Subsistence variables					yes				
Political hierarchies						yes			
Malaria index							yes		
Log population density	yes	yes	yes	yes	yes	yes	yes	yes	yes
Geographic controls		yes	yes	yes	yes	yes	yes	yes	yes
Continent FE			yes						
Country FE				yes	yes	yes	yes	yes	yes
Observations	377,656	377,656	377,656	377,656	377,656	373,070	377,656	377,656	377,656
R-squared	0.488	0.537	0.582	0.660	0.660	0.661	0.660	0.660	0.660
Number of clusters	138	138	138	138	138	138	138	62	96 & 162

Notes: Each observation is a pixel in the homeland of an ethnicity matched with the direct matching method. The geographic controls include temperature, precipitation, agricultural suitability, absolute latitude, elevation, ruggedness, distance to coast, and distance to nearest river or lake. The subsistence variables measure the fraction of an ethnicity's subsistence economy that depends on hunting, fishing, animal husbandry, and agriculture (with gathering as the omitted category). Standard errors in parentheses are clustered at the country level in all regressions, except in column 8, where they are clustered at the language-family level, and in column 9, where they are clustered two-way at both the country and language-family levels. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

controls, β decreases in magnitude to -0.42. Adding continent and country fixed effects, in columns 3 and 4, reduces the magnitude of the estimates to -0.136 and -0.110, but also decreases the standard errors, so the estimates of β remain significant. These estimates imply that a one-standard-deviation increase in the KII is associated with a $\sim 12\%$ decrease in luminosity.

In columns 5 to 9, we control for potentially endogenous correlates of the KII that may confound its association with luminosity. In columns 5 and 6, respectively, we control for each ethnicity's subsistence economic activities and their degree of political hierarchy; and in column 7, we control for the pixel-level malaria index. Finally, in column 8, we cluster standard errors at the language-family level, and in column 9, we use two-way clustering at both the country and language-family levels. Our results are robust to these alternative specifications.

To confirm that these results do not hinge on our use of the direct matching method, we re-ran the analyses reported in Table 1 using only matches from the language-tree matching method. Appendix Table B.2.2 shows similar, if somewhat smaller (in magnitude), estimates for β.

We also examined the robustness of these results to analyzing the data at the country-ethnicity level instead of at the pixel level. Appendix Table B.2.3 reports the results of regressions that parallel those reported in Table 1 and Appendix Table B.2.2. With the direct-matching method, the estimates of β are consistent throughout and similar (but somewhat larger in magnitude). Using the matches from the language-tree method, the results are only significant without continent or country fixed effects.

To examine whether these strong and consistent associations between luminosity and the KII are driven by one or a subset of the five KII sub-indicators, we estimated the specification in columns 2 (with the geographic controls) and 4 (with country fixed effects as well) of Table 1 again, but separately for each of the five sub-indicators instead of the KII as the explanatory variable. Panel B of Appendix Table B.2.4 shows the results obtained with the direct matching method. The associations between luminosity and the sub-indicators are all negative (except that in column 2, which is nearly 0), and most are significant at the 5% level. The results with the language-tree matching method (in Panel A) are similar. Thus, most of the five dimensions of kinship intensity measured by our sub-indicators are associated with lower levels of economic development, and none is associated with higher levels. This gives us confidence in the theoretical soundness of our KII construct and in the appropriateness of aggregating the five sub-indicators into the KII to study its association with economic development.

4.2 Kinship intensity and regional GDP per capita

We now verify that the association between kinship intensity and economic development is robust to using a more direct, but less fine-grained, measure of economic development: log regional GDP per capita (mapped in Appendix Figure B.3.1). To remain consistent with the above analyses of the KII and nighttime luminosity, we estimated similar specifications, except that we included year, year-continent, or year-country fixed effects in the regressions (because of the panel nature of the data), and did not include population density (since the dependent variable is in per capita terms). As above, we clustered standard errors at the country level. We used the same set of control variables, constructed from the same data sources, and estimated one additional specification that also controls for a dummy equal to 1 if the national capital is located in the region as well as for each region's cumulative oil, gas, and liquid natural gas production from the time production began to 2000.

Table 2: The KII and regional GDP per capita: OLS estimates

	Log regional GDP per capita							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
KII	-0.445***	-0.459***	-0.455***	-0.085**	-0.168***	-0.129***	-0.081**	-0.081**
	(0.105)	(0.092)	(0.094)	(0.034)	(0.039)	(0.046)	(0.034)	(0.038)
Subsistence variables					yes			
Political hierarchies						yes		
Malaria index							yes	
Oil and Gas production								yes
Capital is in Region								yes
Geographic controls		yes	yes	yes	yes	yes	yes	yes
Year FE	yes	yes						
Year × Continent FE			yes					
Year × Country FE				yes	yes	yes	yes	yes
Observations	5,514	5,514	5,514	5,514	5,514	5,514	5,514	5,514
R-squared	0.313	0.511	0.610	0.889	0.890	0.889	0.890	0.902
Number of clusters	61	61	61	61	61	61	61	61

Notes: Each observation is a region-year in the analysis sample obtained with the direct matching method. The geographic controls include temperature, precipitation, agricultural suitability, absolute latitude, elevation, ruggedness, distance to coast, and distance to nearest river or lake. The subsistence variables measure the weighted fraction of a region's ethnicities' subsistence economies that depend on hunting, fishing, animal husbandry, and agriculture (with gathering as the omitted category). Standard errors in parentheses are clustered at the country level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 2 reports the results for the sample obtained with the direct matching method. In column 1, we only control for year fixed effects. The estimated coefficient on the KII is -0.445, implying that a one-standard-deviation increase in the KII is associated with a $\sim 36\%$ decrease in GDP per capita. Adding the geographical controls (in column

2) and year-continent (in column 3) or year-country (in column 4) fixed effects reduces the magnitude of the estimated coefficient on the KII, but also increases precision; as a result, the coefficient remains significant at the 1% or 5% level. The estimated coefficient in column 4, with country-year fixed effects, is -0.085, implying that a one-standard-deviation increase in the KII is associated with a $\sim$ 8% decrease in GDP per capita within country. These estimate are strikingly similar to those in Table 1, which imply that a one-standard deviation increase in the KII is associated a $\sim$ 40% decrease in luminosity per capita, and a $\sim$ 10% decrease within countries.

Columns 5 to 7 parallel the specifications in Table 1 with additional controls for the subsistence variables, political hierarchy, and the malaria index. In column 8, we control for cumulative oil and gas production and for the dummy indicating whether the national capital is located in the region. Appendix Table B.3.2 shows the results for the sample obtained with the language-tree matching method. The results remain robust across all of these alternative specifications.

4.3 Spatial regression discontinuity analysis

Our estimates of the association between the KII and economic development are robust to the inclusion of a rich set of control variables, including geographic, ecological and cultural variables as well as country fixed effects. Despite this, it is possible that unobservable factors may confound our results. These unobservable factors could include, for example, geographic characteristics such as suitability for certain subsistence practices that co-determine kinship intensity and economic prosperity. They could also include contemporary unobservable factors associated with remoteness such as the limited reach of central governments or a lack of infrastructure that may simultaneously increase kinship intensity while hampering economic activity.

To address these concerns, we conducted a spatial regression discontinuity (RD) analysis following Michalopoulos and Papaioannou (2013) and Moscona et al. (2020). We compared the luminosity of pixels that are geographically close to one another but belong to the homelands of neighboring country-ethnicities that vary in their kinship intensity. This estimation strategy helps account for unobservable factors—including those tied to geography, history, or infrastructure—that vary smoothly across space.

For our baseline RD analysis, we restrict the sample to pixels that belong to pairs of contiguous country-ethnicities and that are within 200 km of the ethnic boundary that separates the two groups. To address concerns that country-level unobservables such as national institutions may bias the estimates, we further restrict the sample to pairs of ethnicities located in the same country. Below, we report results based on the sample obtained with the direct matching method (see Section 3.3), but we verified that the results are robust to using the language-tree matching method (Appendix Table B.4.2).

Figure 5 illustrates this setup for a pair of contiguous country-ethnicities in Zimbabwe: the Ndebele and the Venda. All pixels within 200 km of the boundary are included, with the exception of pixels that fall directly onto an ethnic boundary or a country border (since those pixels cannot readily be attributed to only one ethnicity or country). The Venda, whose KII is 1.948, have higher kinship intensity than the Ndebele, whose KII is -0.159.

We begin by graphically examining the relationship between luminosity and the distance to the boundary separating the ethnicities in the pairs, for the set of pixels used in our baseline analysis. Figure 6 shows a binned scatterplot of log luminosity residualized

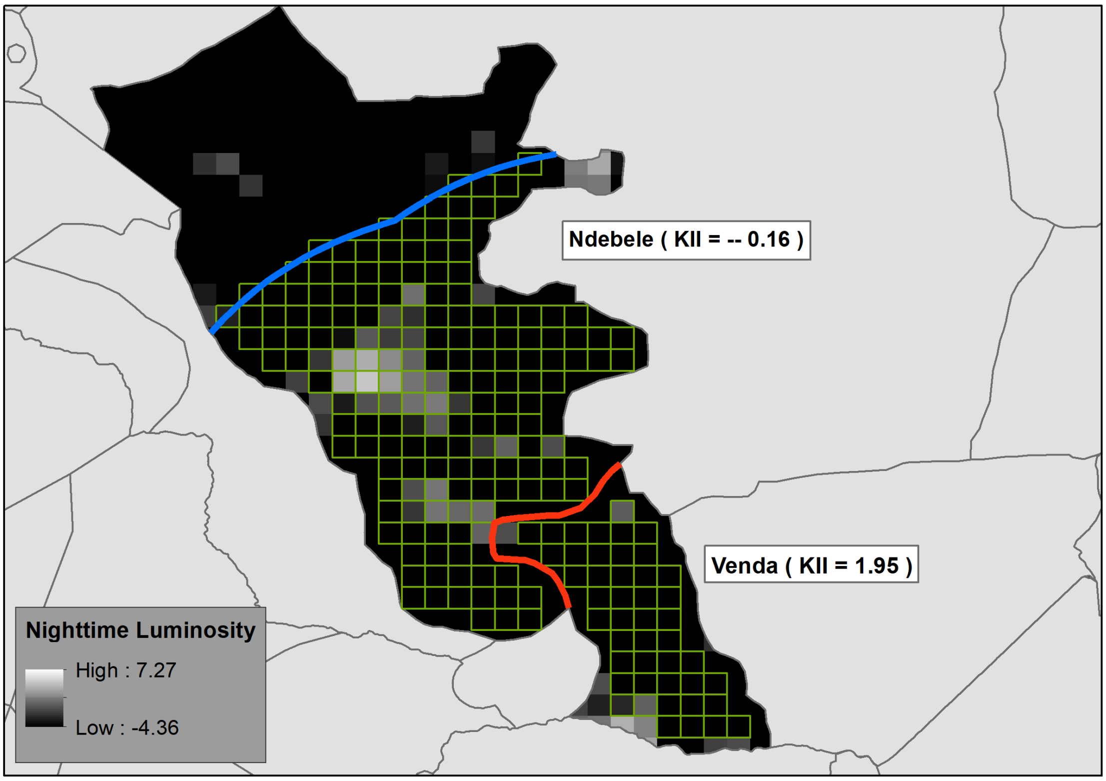


Figure 5: The map shows the nighttime luminosity of the pixels (marked by green squares) for a pair of contiguous ethnicities in Zimbabwe which are included in our analysis: the Ndebele (to the North) and the Venga (to the South). The red line is the ethnic boundary that separates the two ethnicities, and the blue line is located 200 km away from it. We drop pixels whose centroid is not located within 200 km of the ethnic boundary. The whole territory of Venda is within 200 km of the ethnic boundary, therefore, is included in the analysis. Our analysis only includes pixels that are not crossed by any other ethnic boundaries (the red line) or national borders (Gray borders). The Venda have a higher kinship intensity than the Ndebele.

on log population density (on the y-axis) vs. distance to the ethnic boundary (on the x-axis). Distance to the boundary is negative for pixels in the homeland of an ethnicity that has the lower KII in a pair, and positive for the ethnicity with the higher KII. The figure clearly reveals the discontinuity at the boundary. While luminosity is rather uniformly distributed to the left and to the right of the boundary, it sharply drops at the ethnic boundary as we move from the relatively low to the relatively high KII ethnicities.

Next, to implement our formal spatial RD analysis, we adopt the following baseline specification:

$$L_{i,e(e'),c} = \beta KII_e + \delta \log(P_{i,e,c}) + \gamma X_{i,e,c} + \theta V_{(i),e} + \lambda_{ee',c} + f(D_{i,e(e'),c}) + \epsilon_{i,e(e'),c}. \qquad (2)$$

The dependent variable $L_{i,e(e'),c}$ is the natural logarithm of the nighttime luminosity of pixel i in the homeland of ethnicity e that is adjacent to the homeland of ethnicity e', with both e and e' in the same country c.[15] We include ethnicity-pair fixed effects ($\lambda_{ee',c}$), which account for all (unobserved) factors unique to each ethnicity-pair. Following Moscona et al. (2020) and Gelman and Imbens (2019), we also include a local linear polynomial in the geodesic distance of each pixel's centroid from the boundary between the two adjacent

---

[15] Note that a pixel can enter the regression multiple times as separate observations if the country-ethnicity in whose homeland it falls is in pairs with multiple contiguous other country-ethnicities.

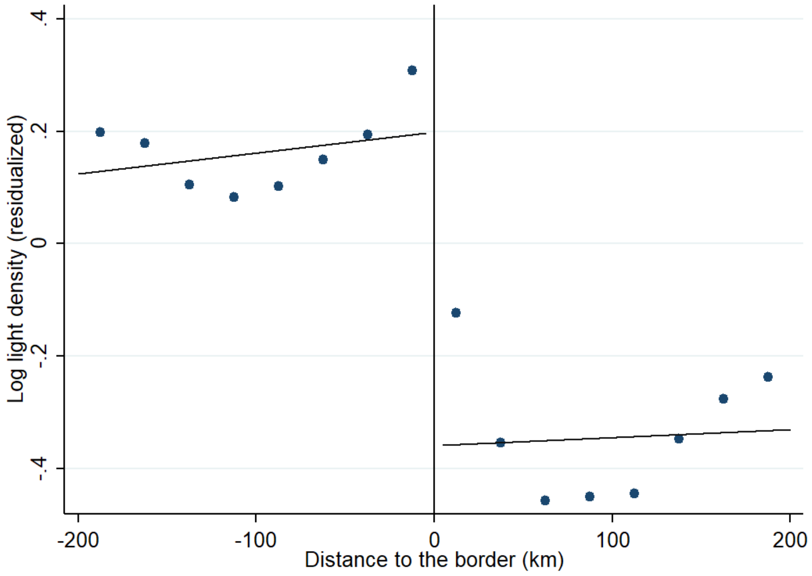


Figure 6: Binned scatterplot of the relationship between distance to the boundary and luminosity, across pixels located in the homelands of pairs of adjacent ethnicities in the same country. The y-axis shows log luminosity residualized on log population density. The x-axis shows the distance to the boundary in kilometers, with negative distances for pixels associated with the ethnicity with the lower KII in a pair, and positive distances for the ethnicity with the higher KII. The average KII across the pixels that belong to ethnicities with the relatively low KII is -1.54; for the ethnicities with the relatively high KII, it is 0.09. (We created 16 bins according to the distance to the border (25km each); this implies that bins vary in the number of the underlying data points.)

ethnicities, $f(D_{i,e(e'),c})$, with different coefficients on the distance term for the relatively high and the relatively low KII ethnicities in the pairs. As in equation (1), KII_e is the KII for ethnicity e, $P_{i,e,c}$ is the pixel's population density, $X_{i,e,c}$ is the vector of geographic controls, and $V_{(i),e}$ denotes the additional controls used in some specifications.

Table 3 reports the results for this baseline specification. In column 1, without the geographic controls, the estimated coefficient on the KII is $\hat{\beta} = -0.06$, which implies that a one-standard deviation increase in the KII is associated with a $\sim 6\%$ decrease in luminosity. Adding the baseline geographic controls in column 2 increases β only slightly in magnitude, to -0.072.

In columns 3–5, $\hat{\beta}$ remains fairly stable when we add controls for the subsistence variables, political hierarchy, and the malaria index. Similarly, clustering at the language-family level or two-way clustering at both the country and language-family levels (columns 6 & 7) does not appreciably impact the standard errors. As can be seen in Appendix Table B.4.2, these results also hold in the sample obtained using the language-tree-matching method, though estimates of β become slightly smaller in magnitude.

Table 4 shows the results of several robustness checks. In columns 1 and 2, we exclude ethnicity-pairs for which the difference in KII is less than 0.1 or 1 standard deviations, respectively. In column 3, to address concerns that some pixels that are part of a large number of ethnicity-pairs receive too much weight in the regressions, we drop pixels that are part of more than 20 different ethnicity pairs. In column 4, we include pairs of neighboring ethnicities that are located in different countries and include country fixed effects in the regression. Finally, in columns 5 and 6, we use alternative approaches to control for the pixels' distances to the ethnic boundaries. Column 5 interacts the distance-to-the-boundary polynomial with ethnicity-pair fixed effects, and so includes a

Table 3: The KII and nighttime luminosity: spatial RD analysis

	Log nighttime luminosity						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
KII	-0.060***	-0.072***	-0.085**	-0.078**	-0.071***	-0.072***	-0.072***
	(0.019)	(0.017)	(0.039)	(0.038)	(0.017)	(0.018)	(0.018)
Log population density	1.088***	1.059***	1.056***	1.061***	1.056***	1.059***	1.059***
	(0.084)	(0.077)	(0.079)	(0.076)	(0.077)	(0.112)	(0.115)
Subsistence variables			yes				
Political hierarchies				yes			
Malaria index					yes		
Log population density	yes	yes	yes	yes	yes	yes	yes
Geographic controls		yes	yes	yes	yes	yes	yes
Distance-to-the-boundary polynomial	yes	yes	yes	yes	yes	yes	yes
Ethnicity pair FE	yes	yes	yes	yes	yes	yes	yes
Observations	290,669	290,669	290,669	289,740	290,669	290,669	290,669
R-squared	0.600	0.613	0.613	0.614	0.614	0.613	0.613
Number of clusters	70	70	70	70	70	58	58 & 70

Notes: Each observation is a pixel that belongs to an ethnicity in a pair of contiguous ethnicities in the same country, and that falls within 200 km of the boundary between the two ethnicities. The geographic controls include temperature, precipitation, agricultural suitability, absolute latitude, elevation, ruggedness, distance to coast, and distance to nearest river or lake. "Distance-to-the-boundary polynomial" is the geodesic distance of each pixel's centroid from the boundary between the two adjacent ethnicities; we allow the coefficient on the distance term to differ between the relatively high and the relatively low KII ethnicities in the pairs. The subsistence variables measure the fraction of an ethnicity's subsistence economy that depends on hunting, fishing, animal husbandry, and agriculture (with gathering as the omitted category). Political hierarchy measures the number of levels of jurisdictional hierarchy for an ethnicity, while the Malaria Index captures the ecological suitability for malaria. Standard errors in parentheses are clustered at the country level (in columns 1–5), language family (in column 6), and two-way clustered at both the country and language-family levels (in column 7). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

different coefficient on the distance term for each ethnicity in each of 572 ethnicity pairs. This is a demanding specification, but it allows us to control for distance patterns specific to each ethnicity in each pair. In column 6, rather than controlling for the distance to the boundary, we interact the longitude and latitude of each pixel with ethnicity-pair fixed effects; this specification controls more directly for features that vary in two-dimensional space in the homelands of each ethnicity pair. The results are robust to these alternative specifications.

For several reasons, our analysis so far may underestimate the true magnitude of β. First, our estimates may capture the spillover effects of light across ethnic boundaries—i.e., some pixels may appear brighter than they actually are due to overglow from bright nearby pixels in the neighboring ethnicity's homeland. Second, there may be spillovers in economic activity across the boundaries. And third, the fraction of the population belonging to each ethnicity may not change discontinuously at the ethnic boundary, for instance due to intermarriage, trade, or the presence of multi-ethnic urban agglomerations.

To address these concerns, table 5 presents specifications using subsets of pixels that fall within different ranges of distances to the ethnic boundary. In columns 1 to 3, we still start at the boundary, but reduce the maximal distance to the boundary from 200 km (our baseline specification, also reported in column 2 of Table 3) to 150 km and then 100 km. In columns 4–6 and 7–9, to mitigate potential bias due to light and economic spillovers and the smooth variation in population composition across boundaries, we drop pixels that are within 25 km and 50 km of the boundary, respectively.

The estimates are remarkably robust to using these alternative subsets of pixels. Consistent with our intuition that light and economic spillovers and the smooth variation in population composition at the boundary may bias our estimates downward, the esti-

Table 4: Spatial RD analysis: additional robustness checks

	Log nighttime luminosity					
	(1) KII diff. ≥ 0.1	(2) KII diff. ≥ 1	(3) Only in $\leq$20 pairs	(4) Cross-country sample	(5)	(6)
KII	-0.069***	-0.037*	-0.072***	-0.070***	-0.057**	-0.051***
	(0.018)	(0.022)	(0.017)	(0.017)	(0.024)	(0.012)
Log population density	yes	yes	yes	yes	yes	yes
Ethnicity pair FE	yes	yes	yes	yes	yes	yes
Log population density	yes	yes	yes	yes	yes	yes
Geographic controls yes	yes	yes	yes	yes	yes	yes
Country FE				yes		
Distance-to-the-boundary polynomial	yes	yes	yes	yes		
Dist. to the boundary poly. X ethnicity pair					yes	
Latitude & longitude X ethnicity pair						yes
Observations	277,946	164,191	287,624	370,771	290,669	290,669
R-squared	0.609	0.578	0.614	0.637	0.633	0.646
Number of clusters	70	46	70	104	70	70

Notes: Each observation is a pixel that belongs to an ethnicity in a pair of contiguous ethnicities in the same country (except in column 4), and within 200 km of the boundary between the two ethnicities. The geographic controls include temperature, precipitation, agricultural suitability, absolute latitude, elevation, ruggedness, distance to coast, and distance to nearest river or lake. In columns 1–4, "Distance-to-the-boundary polynomial" is the geodesic distance of each pixel's centroid from the boundary between the two adjacent ethnicities; we allow the coefficient on the distance term to differ between the relatively high and the relatively low KII ethnicities in the pairs. In columns 1 and 2, we drop ethnicity pairs whose KII differences is less than 0.1 and 1 standard deviations, respectively. In column 3, we drop pixels that appear in more than 20 ethnicity pairs. In column 4, we also include in the analysis ethnicity pairs with ethnicities in different countries, and include country fixed effects in the regression. Column 5 interacts the distance-to-the-boundary polynomial with ethnicity pair fixed effects, and column 6 interacts instead each pixel's latitude and longitude with ethnicity-pair fixed effects. Standard errors in parentheses are clustered at the country level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 5: Spatial RD analysis with pixels at various distances from the ethnic boundaries

	Log nighttime luminosity								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Distance to border (in km)	0-200	0-150	0-100	25-200	25-150	25-100	50-200	50-150	50-100
KII	-0.072***	-0.065***	-0.046*	-0.096***	-0.093***	-0.079**	-0.118***	-0.121***	-0.115**
	(0.017)	(0.021)	(0.028)	(0.022)	(0.025)	(0.034)	(0.028)	(0.030)	(0.045)
Log population density	yes	yes	yes	yes	yes	yes	yes	yes	yes
Geographic controls	yes	yes	yes	yes	yes	yes	yes	yes	yes
Distance-to-the-boundary polynomial	yes	yes	yes	yes	yes	yes	yes	yes	yes
Ethnicity pair FE	yes	yes	yes	yes	yes	yes	yes	yes	yes
Observations	290,669	219,874	146,438	268,628	197,833	124,397	227,620	156,825	83,389
R-squared	0.613	0.619	0.623	0.615	0.621	0.626	0.616	0.624	0.630
Number of clusters	70	70	70	70	70	70	70	70	70

Notes: Each observation is a pixel that belongs to an ethnicity in a pair of contiguous ethnicities in the same country. The various specifications are identical, but the analysis samples include pixels located at different ranges of distances from the boundary separating the two ethnicities in a pair. The geographic controls include temperature, precipitation, agricultural suitability, absolute latitude, elevation, ruggedness, distance to coast, and distance to nearest river or lake. "Distance-to-the-boundary polynomial" is the geodesic distance of each pixel's centroid from the boundary between the two adjacent ethnicities; we allow the coefficient on the distance term to differ between the relatively high and the relatively low KII ethnicities in the pairs. Standard errors in parentheses are clustered at the country level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

mates become larger in magnitude as we drop pixels that are within 25 and then 50 km of the ethnic boundary. When the pixels within 50 km of the boundary are dropped, our estimates imply that a one-standard-deviation increase in the KII is associated with a decrease of $\sim 11\%$ ($= (1 - e^{-0.12}) \cdot 100\%$) in luminosity.

Since all the ethnicity pairs in our baseline RD analysis sample are in the same country, we can compare our RD estimates to those of the earlier OLS regressions with country fixed effects. With log luminosity and log regional GDP per capita as the dependent variables, the earlier corresponding estimates, reported in column 4 of Table 1 and of Table 2, were -0.110 and -0.085, similar to our spatial RD estimates when pixels within 25 or 50 km from the ethnic boundaries are dropped. This reveals a noteworthy consistency across the results of the three main empirical approaches presented thus far.

Our spatial RD analysis helps accounts for potential bias from unobservables that vary smoothly across ethnic boundaries. While it is not possible to test whether there are unobservables that vary discontinuously at the boundary, we can conduct a "placebo" RD analysis and examine whether observable variables show any signs of discontinuities at the boundaries. To do so, we estimated equation (2) again in our baseline analysis sample of pixels, but dropped the geographic controls and swapped the dependent variable, log luminosity, for each of our geographic controls in turn. The results, reported in Appendix Table B.4.3, show that the estimated coefficients on the KII are not significant at the 5% level for any of the geographic controls.[16] Overall, we find no evidence that discontinuities in observable geographic characteristics are present at ethnic boundaries.

5 The inbreeding coefficient, kinship intensity, and economic development

To address potential shortcomings of using the KII as a measure of kinship intensity, we used genetic data to compute the inbreeding coefficient for individuals in the Human Origins (HO) dataset (Reich Lab, 2020). The inbreeding coefficient, denoted F, is a core variable in population genetics that measures the relatedness of ones' parents. HO contains genetic data on 9,460 present-day individuals and 3,723 ancient individuals [17] from populations around the world, compiled from previously published genetics studies (including Jeong et al., 2019; Lazaridis et al., 2014, 2016; Lipson et al., 2018; Nakatsuka et al., 2017; and Pickrell et al., 2012). We matched the present-day HO individuals to Ethnologue ethnicities and used the ethnicities' average inbreeding coefficients as our alternative measure of kinship intensity.

We begin this section by briefly describing the inbreeding coefficient as well as the methodology we used to estimate it, and by defining several additional variables we computed with the HO genetics data. Appendix C provides further detail. Next, we examine the inbreeding coefficient's associations with kinship intensity, and then with economic development.

5.1 The inbreeding coefficient and other genetic variables

Inbreeding occurs when two related individuals mate and produce offspring. Related individuals share DNA from one or more common ancestors, and thus inbreeding increases the probability that at a given location on a chromosome, their offspring inherits identical DNA segments on the maternal and paternal copies of the chromosome (Ceballos et al.,

---

[16] The estimated coefficient is significant at the 10% level when "distance to coast" is the dependent variable, but this may be a false positive due to multiple hypothesis testing.

[17] The ancient individuals lived between $\sim 88,000$ BCE and ~ 1900 CE; we did not analyze their data.

2018). (Recall that humans have 23 pairs of chromosomes: they inherit one of the chromosomes in each pair from their mother and the other from their father.) First cousins, for instance, share both a grandmother and a grandfather, so some DNA segments they inherited from their shared grandparents will be identical across their two genomes. As a result, if two first cousins mate and produce a child, that child will have locations in their genome where they will have inherited identical DNA segments from their two cousin parents, coming originally from the same grandmother or grandfather. The maternal and paternal variants at such genomic location are said to be homozygous—since they are identical—and identical-by-descent (IBD)—since they can also be traced back to one common ancestor.

A common measure of the intensity of inbreeding is called the coefficient of inbreeding, denoted F. F measures the probability that the maternal and paternal variants at a location in the genome are IBD, and this is also equal to the expected fraction of the genome that is IBD for a given individual. An individual's F coefficient is typically equal to the coefficient of kinship, or one-half the coefficient of relationship, between their two parents (Cavalli-Sforza and Bodmer, 1999; Falconer and Mackay, 1996). The latter is a measure of the relatedness between the two parents; for example, the expected F coefficient of the offspring of two second-degree cousins is 0.015625 ($= \frac{1}{2} \cdot \frac{1}{62}$), whereas it is 0.0625 ($= \frac{1}{2} \cdot \frac{1}{8}$) for the offspring of first-degree cousins and 0.25 ($= \frac{1}{2} \cdot \frac{1}{2}$) for the offspring of two siblings. Thus, F measures the degree to which an individuals' two parents are related, and a population's average F should in principle correlate positively with its KII, and with its rate of cousin marriage in particular.

Various methods have been developed to estimate F using an individual's genetic data. Our main measure of F, denoted F_{ROH}, is the fraction of an individual's autosomal genome that contains runs of homozygosity (ROHs) of at least 1.5 Mb in length. The autosomal genome comprises the 22 pairs of chromosomes other than the sex (X and Y) chromosomes, and ROHs are continuous segments of homozygous variants. Because the length of the autosomal genome is estimated at 3,000 Mb[18] (Clark et al., 2019), an individual's F_{ROH} can be calculated as

$$F_{ROH} = \sum_i \frac{l_i}{3,000}, \tag{3}$$

where the sum is over the individual's ROHs that are at least 1.5 Mb in length and l_i is the length of ROH i in Mb.

Longer ROHs (e.g., ROHs at least 1.5 Mb in length) typically result from inbreeding (Ceballos et al., 2018) and have been shown to correlate positively with pedigree-based estimates of inbreeding (Kang et al., 2017; McQuillan et al., 2008; Pemberton and Rosenberg, 2014) and with population-level cousin marriage preferences (Sahoo et al., 2021). Below in Section 5.2, we too find that F_{ROH} (or, equivalently, longer ROHs) correlates positively with cousin marriage preferences; we also find that F_{ROH} correlates positively with the KII.

We estimated F_{ROH} for 4,756 present-day individuals in the Human Origins (HO) data set whose genetic data were collected using the Affymetrix Human Origin genotyping array and who passed a number of quality control filters. As we describe in more detail in Appendix C, we used the ROHgen2 software pipeline developed by the ROHgen

---

[18] A megabase (Mb) is a unit of physical distance in the genome

consortium (Clark et al., 2019)[19] to estimate F_{ROH}, and conducted diagnostic checks to verify the reliability of our F_{ROH} estimates. We then matched the HO individuals to 551 Ethologue ethnicities using a procedure that resembles the direct matching method described in Section 3.3 that we used to match the EA data. We computed the average F_{ROH}, which we denote $\overline{F_{ROH}}$, across the HO individuals matched to each ethnicity. We dropped ethnicities with fewer than eight individuals, leaving 416 ethnicities in the sample. For our main analysis with $\overline{F_{ROH}}$—of its relationship with luminosity at the pixel level—we further dropped pixels that did not pass the filters described in Sections 3.1 and 3.3, thus leaving 281,177 pixels that fall in the country-homelands of 245 country-ethnicities.

We note that while F is the fraction of the genome that is IBD, F_{ROH} is the fraction that is in ROHs longer than 1.5 Mb. Because some ROHs that are shorter than 1.5 Mb are IBD and some longer ROHs are not IBD, F_{ROH} is not a perfect measure of F. ROHs can arise in individuals for a variety of reasons unrelated to marital practices and kinship systems. Using only ROHs that are at least 1.5 Mb long to compute F_{ROH} helps mitigate, but does not eliminate, concerns that F_{ROH} may capture influences other than consanguineous marriage practices and intensive kinship.

For these reasons, we also used the HO genetic data to compute the following "genetic controls" that may correlate with F_{ROH} but are unrelated to kinship intensity and consanguineous marriage practices, and included them in our empirical analyses as controls:

- *Expected heterozygosity*: Expected heterozygosity is a measure of genetic diversity in a population and is defined as the probability that two randomly selected individuals from a population have different genetic variants at a randomly selected location on one chromosome in the genome. Expected heterozygosity has been shown to correlate negatively across populations with mean F estimated with genetics data (Pemberton and Rosenberg, 2014) and to correlate nonlinearly across countries with economic development (Ashraf and Galor, 2013). Expected heterozygosity decreases with *migratory distance from East Africa*, consistent with a migration model that predicts increased genetic drift and decreased heterozygosity with increased distance from Africa (Ramachandran et al., 2005). We also include migratory distance from East Africa among our "genetic controls".

- *The top 20 principal components of the genotypic data*: We computed the top 20 principal components (PCs) of the genotypic data for each individual and took the average value of each PC in each ethnicity. It has been shown that the top PCs are good proxies for individuals' geographic origins and that including them as controls in regressions helps mitigate bias related to population stratification (Novembre et al., 2008; Price et al., 2006). (Population stratification refers to systematic differences in genetic variants' frequencies that correlate with cultural or environmental differences (Hamer, 2000).)

- *Mean regional pairwise F_{ST}*: Both F and F_{ROH} tend to be inflated in populations that are genetically isolated (due to geographical or cultural distance or taboos prohibiting mating with individuals from other populations—see Pemberton and Rosenberg (2014)). In such populations, reduced mate choice can lead to increased cryptic inbreeding, in which two parents may have a recent common ancestor by

---

[19] The ROHgen consortium is a large consortium of research groups that seeks to estimate the effect of inbreeding on various traits (Clark et al., 2019; Joshi et al., 2015).

chance and not due to a deliberate cultural practice of mating with relatives. Following Pemberton and Rosenberg (2014), we estimate the degree of genetic isolation of a population using the average pairwise genetic distance "F_{ST}" between that population and other populations in the same geographic world region with sufficient sample sizes.

Appendix C provides additional details on these genetic controls and Appendix Table B.5.1 shows summary statistics for $\overline{F_{ROH}}$ and these genetic controls (except the PCs, whose scale is arbitrary). Across 416 Ethnologue ethnicities, $\overline{F_{ROH}}$ ranges from 0.003 to 0.146, with a mean of 0.020 and a standard deviation of 0.018. To put these figures in perspectives, two populations in which everyone were the offspring of second-cousin and uncle-niece unions would have mean F of 0.015625 and 0.125, respectively, close to the mean and maximum $\overline{F_{ROH}}$ in our sample. We estimated the mean $\overline{F_{ROH}}$'s of the English, Kurukh, and Marri Baloch (the three societies described in Section 2.2) to be 0.0065, 0.0157, and 0.0617, respectively.

5.2 The inbreeding coefficient and kinship intensity

As mentioned above, F_{ROH} has been shown to correlate with inbreeding and both consanguineous marriage preferences and practices (Ceballos et al., 2018; Kang et al., 2017; McQuillan et al., 2008; Pemberton and Rosenberg, 2014; Sahoo et al., 2021). In theory, endogamy and polygyny should also increase inbreeding, and so our community organization and polygamy KII sub-indicators should be positively associated with $\overline{F_{ROH}}$. Norms favoring endogamy mean that people marry and mate within a relatively small population, thus increasing the likelihood of consanguineous unions (without explicit prescribing them). The role of polygyny can be more difficult to understand. Polygyny, all else being equal, increases inbreeding in a population by decreasing the total number of fathers who provide genes each generation—meaning more people share the same fathers, grandfathers and great grandfathers, and are thus effectively paternal cousins.

To further validate our estimates of $\overline{F_{ROH}}$ as a proxy for kinship intensity, we examined its association with the cousin-marriage-preference KII sub-indicator, with the KII itself, as well as with the four other KII sub-indicators. Across the 398 ethnicities with non-missing data, the correlation between $\overline{F_{ROH}}$ and the cousin-marriage-preference sub-indicator is 0.28; after partialling out the effects of the genetic controls from $\overline{F_{ROH}}$, the correlation increases to 0.32. Unsurprisingly, given that the cousin-marriage-preference sub-indicator is one of the five components of the KII and that some other dimensions of the KII can impact consanguinity, $\overline{F_{ROH}}$ also correlates with the KII ($\hat{r} = 0.26$), though the correlation is lower if $\overline{F_{ROH}}$ is first residualized on the genetic controls ($\hat{r} = 0.14$).

Table 6 shows regression results of the cousin-marriage-preference sub-indicator and of the KII on $\overline{F_{ROH}}$. In all specifications, the coefficient on $\overline{F_{ROH}}$ is highly statistically significant and large in magnitude. For instance, in the regressions in columns 2 and 6 with the genetic controls, the coefficient estimates of 45.744 and 13.185 imply that the cousin-marriage preferences and KII of an ethnicity in which everyone is the offspring of second cousins ($\overline{F_{ROH}} = \frac{1}{64} = 0.015625$, assuming no influences other than inbreeding), are $\sim$ 0.7 and $\sim$ 0.2 standard deviations higher than those of an ethnicity in which everyone is the offspring of unrelated parents ($\overline{F_{ROH}} = 0$; by comparison, as shown in Appendix Table B.5.1, the standard deviation of $\overline{F_{ROH}}$ across country-ethnicities is 0.018). Further, $\overline{F_{ROH}}$ accounts for a nontrivial share of the variation in cousin marriage preferences: its incremental R^2, defined as the difference in R^2 between the regression on $\overline{F_{ROH}}$ and the controls and the same regression but on the controls only, is 0.095. As shown in Appendix Table B.5.2, these results are robust to alternative specifications.

Table 6: The inbreeding coefficient and kinship intensity

	Cousin marriage preference				KII			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$\overline{F_{ROH}}$	18.491***	45.744***	44.566***	37.557***	11.641***	13.185***	13.319***	14.549***
	(5.346)	(5.653)	(5.772)	(8.379)	(3.067)	(2.714)	(2.761)	(4.033)
Genetic controls		yes	yes	yes		yes	yes	yes
Continent FE			yes				yes	
Country FE				yes				yes
Observations	398	397	397	397	396	395	395	395
R-squared	0.077	0.433	0.453	0.683	0.066	0.717	0.720	0.843
$\Delta R^2(\overline{F_{ROH}})$	0.0775	0.0951	0.115	0.0419	0.0665	0.0171	0.0200	0.0136
Number of clusters	127	127	127	127	127	127	127	127

Notes: Each observation is a country-ethnicity from the Ethnologue. The genetic controls include expected heterozygosity, migratory distance from East Africa, the top 20 PCs, and mean regional pairwise F_{ST}. $\Delta R^2(\overline{F_{ROH}})$ is the incremental R^2 of $\overline{F_{ROH}}$, defined as the difference in R^2 between the regression on $\overline{F_{ROH}}$ and the controls and the regression on the controls only. Standard errors, clustered at the country-level, are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Appendix Table B.5.3 shows the results of regressions of the four other KII sub-indicators on $\overline{F_{ROH}}$. As expected, the polygamy, community organization, and lineage organization sub-indicators are positively associated with $\overline{F_{ROH}}$. Interestingly, the co-residence sub-indicator is negatively associated with $\overline{F_{ROH}}$. This may be due to a natural incest aversion that reduces sexual attraction between individuals who grow up in the same household.

5.3 The inbreeding coefficient and nighttime luminosity

Table 7 reports the results of regressions of luminosity on $\overline{F_{ROH}}$ across pixels. The analysis mirrors that of the association between luminosity and the KII in Section 4.1, but with $\overline{F_{ROH}}$ substituted for the KII as the explanatory variable of interest and with the genetic controls also included in the regressions (except in column 1).

Consistent with our estimates of a negative relationship between luminosity and the KII, the association between luminosity and $\overline{F_{ROH}}$ is consistently negative and accurately estimated in all specification ($p < 0.5$). When population density along with both the genetic and geographic controls are included (in column 3), the coefficient estimate of -25.841 implies that log luminosity is 0.40 units lower ($= 25.841\frac{1}{64}$), and luminosity $\sim 33\%$ lower ($= (1-e^{(-25.841\frac{1}{64})})\cdot 100\%$), for an ethnicity in which everyone is the offspring of second cousins compared to an ethnicity in which everyone is the offspring of unrelated parents. These results are robust to controlling for continent and country fixed effects (columns 4 and 5); for the subsistence variables, political hierarchy, and the malaria index (columns 6–8); as well as to clustering standard errors at the language-family level and two-way at both the country and language-family levels (columns 9–10).

Appendix Table B.6.2 reports additional regressions we conducted to further assess the robustness of the results. Column 1 shows that our results hold when the genetic controls are dropped from the regression. As mentioned above, for our baseline analyses, we dropped ethnicities to which fewer than 8 HO individuals could be matched, so columns

Table 7: The inbreeding coefficient and nighttime luminosity

	Log nighttime luminosity									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
$\overline{F_{ROH}}$	-41.297**	-27.265**	-25.841**	-27.094**	-21.026***	-18.235**	-15.943**	-20.923***	-21.026***	-21.026***
	(18.654)	(11.294)	(11.267)	(12.315)	(7.633)	(8.357)	(6.766)	(7.460)	(5.584)	(7.995)
Subsistence variables						yes				
Political hierarchies							yes			
Malaria index								yes		
Log population density	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Genetic controls		yes	yes	yes	yes	yes	yes	yes	yes	yes
Geographic controls			yes	yes	yes	yes	yes	yes	yes	yes
Continent FE				yes						
Country FE					yes	yes	yes	yes	yes	yes
Observations	281,177	281,177	281,177	281,177	281,177	281,177	281,100	281,177	281,177	281,177
R-squared	0.503	0.633	0.652	0.652	0.674	0.674	0.674	0.674	0.674	0.674
Number of clusters	98	98	98	98	98	98	98	98	30	95 & 157

Notes: Each observation is a pixel. The genetic controls include expected heterozygosity, migratory distance from East Africa, the top 20 PCs, and mean regional pairwise F_{ST}. The geographic controls include temperature, precipitation, agricultural suitability, absolute latitude, elevation, ruggedness, distance to coast, and distance to nearest river or lake. The subsistence variables measure the fraction of an ethnicity's subsistence economy that depends on hunting, fishing, animal husbandry, and agriculture (with gathering as the omitted category). Standard errors in parentheses are clustered at the country level in all regressions, except in column 9, where they are clustered at the language-family level, and in column 10, where they are clustered two-way at both the country and language-family levels. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

2–4 explore the sensitivity of the results to using alternative cutoffs of 5, 10, and 15 individuals. The association between luminosity and $\overline{F_{ROH}}$ is robust to using a cutoff of 5 individuals; however, with larger cutoffs of 10 and 15 individuals, the estimated coefficient on $\overline{F_{ROH}}$ shrinks slightly in magnitude (but remains negative) while the standard error increases, resulting in uncertainty that exceeds conventional cutoffs ($p > 0.05$). Given that relatively few individuals were sampled for each ethnicity (see Appendix Table B.5.1) and that they were sampled by different teams of researchers with different protocols, this lack of robustness is not particularly surprising.

6 Causal pathways

Now that we have empirically established a robust and economically relevant negative association between kinship intensity and economic development, we turn our attention to the causal pathways that may account for that association.

We begin, in Section 6.1, by verifying that the association between kinship intensity and economic development is robust to accounting for two important potential confounders: Christianity and European ancestry. We find that, while the association is generally less significant and robust when adjusting for these factors, it cannot be entirely driven by these factors. Next, in Section 6.2, we discuss the potential role of reverse causality—i.e., the potential causal impact of economic development on kinship intensity. We conclude that, while economic development likely does erode kin-based institutions, it is unlikely to account for much of the kinship intensity-economic development association we document.

We thus interpret our findings as pointing to a likely negative causal effect of kinship intensity on economic development. In Section 6.3, we discuss mechanisms that may account for such a negative causal effect. To do so, we briefly review the relevant literature and supplement that review with our own exploratory analyses. We consider the potential roles of (1) the division of labor and comparative advantage; (2) cultural psychology, including impersonal trust, impersonal cooperation, impartiality, and individualism, and conformity; (3) institutions; (4) innovation; and (5) inbreeding depression. We find supportive evidence for a role for the first four channels, but argue that inbreeding depression is unlikely to play an important role.

6.1 The Church, European ancestry, kinship intensity, and economic development

In Schulz et al. (2019), we argued that the Western Church—the branch of Christianity that evolved into the Roman Catholic Church—had a profound impact on European kin-based institutions during the Middle Ages through its "Marriage and Family Program", which involved a set of policies that forbade cousin marriage and promoted neolocal residence and weak family ties. Relatedly, Christian missions often explicitly preached against polygyny, cousin marriage, arranged marriages and other elements of intensive kinship (Fenske, 2015; Bergeron, 2020) while actively working to spread Christian supernatural beliefs. The Western Church, or Christianity, may also have directly impacted economic prosperity by encouraging schooling and literacy, as has been argued for Protestantism (Becker and Woessmann, 2009). Empirically, Christianity and historical exposure to the Western Church are negatively correlated with kinship intensity (Schulz et al., 2019) and positively correlated with economic development, as is European ancestry. It is thus possible that the association we have documented between kinship intensity and economic development is in fact driven by an omitted variable related to Christianity or European ancestry.

To test for that, we reran all our baseline analyses (1) excluding pixels or regions located in Europe, the Americas, Australia, or New Zealand, and (2) controlling for deep Christianization, a variable that indicates if an EA society had been Christianized for at least 500 years prior to the ethnographic present (Korotayev, 2003). Appendix Table B.7.1 shows the results. The results are not robust in the pixel-level regressions of nighttime luminosity on the KII (Panel A). However, despite the much reduced sample sizes (for (1)) and the high degree of collinearity between the KII and deep Christianization, the results are generally robust at the country-ethnicity level (Panel B), with regional GDP per capita as the dependent variable (Panel C), and with $\overline{F_{ROH}}$ as the explanatory variable (Panel E). The spatial RD coefficient estimates (in Panel D) are similar to those in Table 3, but are estimated less precisely and so mostly lose statistical significance.

We conclude that Christianity and European ancestry cannot fully account for the kinship intensity-economic development association, though accounting for these factors clearly makes the association less strong and robust. Our preferred interpretation, consistent with our prior work (Schulz et al., 2019), is that Christianity had a negative causal impact on European kinship intensity, which in turn had a positive causal impact on European economic development.

6.2 Reverse causation: from kinship intensity to economic development

The negative association between kinship intensity and economic development could in principle be driven by a causal effect of economic development on kin-based institutions. As economies grow, rising urbanization and migration may make it difficult to maintain the tight bonds and obligations that characterize intensive kinship networks. Contacts and exchanges with non-kin may increase. The Kurukh, whom we encountered earlier, were traditionally organized in endogamous patri-clans, but became exogamous when they transitioned from pastoralism to settled agriculture and had more sustained contact with Hindu communities. Further, rising wealth and education may directly impact preferences for maintaining kin-related traditions: Akyol and Mocan (2020) find that a Turkish reform that increased compulsory schooling by three years increased women's preferences for personal autonomy and decreased their tendency to marry cousins or blood relatives. And economic development may also allow governments to establish social safety nets, which may alter incentives to maintain some kin-based practices. For instance, Bau (2021) finds that new pension policies in Ghana and Indonesia reduced the practice of matrilocality and patrilocality.

Though economic development likely decreases kinship intensity, it is unlikely that our findings are driven primarily by reverse causality. First, the EA data based on which the KII is constructed comes from ethnographic observations that capture, or aimed to capture, ancestral conditions prior to European colonization and industrialization; our two measures of economic prosperity, by contrast, are from 2010 (nighttime luminosity) and from 1950 to 2010 (regional GDP per capita).

Second, while it is possible that past economic prosperity impacted both kinship intensity and contemporary prosperity, our spatial RD analysis addresses concerns that any factors that vary smoothly across ethnic boundaries drive our results. This means that geographic variables such as soil quality—which could influence past economic conditions—and remoteness (e.g., from the national capital, government agencies, infrastructure, and trading routes) are unlikely to drive our results. Thus, for past prosperity to confound our results, that past prosperity would have had to stem from factors other than these.

Third, we verified the robustness of our results to controlling for a proxy for country-ethnicities' past economic prosperity. We reran the regressions in columns 1-4 of Table 1 (with direct matches) and Appendix Table B.2.2 (with language-tree matches) with an additional control for settlement complexity, a variable from the EA. The results remain significant and the estimated coefficient on the KII does not change appreciably (untabulated results available upon request).

Finally, we examined whether urbanization, which typically accompanies economic growth, could drive our result. We reran the baseline regressions in columns 2 and 4 of Table 1, but using only pixels with population density (1) equal to 1, (2) between 1 and 10, and (3) higher than 10. The idea is that high-population density reflects urbanization or modernization. Appendix Table B.7.2 reports the results. Our results hold in each sub-sample of pixels (but lose significance in the sample of pixels with population density higher than 10 and with country fixed effects). In particular, our results hold in the sample of low-population-density pixels, thus indicating the urbanization is unlikely to account for the association we have documented between kinship intensity and economic development.

6.3 From kinship intensity to economic development: potential mechanisms

We have shown that the kinship intensity-economic development association is unlikely to be entirely driven by an omitted factor related to Christianity or European ancestry or by reverse causality. We now discuss four mechanisms that likely contribute to a causal effect of kinship intensity on economic development, as well as a fifth one (inbreeding depression) which we argue is unlikely to play a large role.

6.3.1 The division of labor and comparative advantage

At the heart of economics lie the concepts of the division of labor and comparative advantage, which can raise productivity and increase prosperity (Smith, 1776). We postulate that high kinship intensity makes it difficult for a society to fully exploit opportunities for specialization and trade. For example, ethnographically, the high kinship intensity Marri Baluch reported not trading in towns where they lacked life-long and often inherited "trading friends" (*bradirs*) and being reluctant to hire herders who were non-relatives. Overall, they tend to organize labor and production primarily within kinship lines. As we discuss below, intensive kinship is associated with low levels of trust towards strangers, which in turn is necessary for trade to flourish.

To explore the link between kinship intensity and specialization, we used data from the EA to construct an index of labor specialization across 11 different tasks: metal working, weaving, leather working, pottery, boat building, house building, gathering, hunting, fishing, animal husbandry, and agriculture. Task specialization can take one of three values: no or only age specialization (0), craft specialization (1), or industrial specialization (2). The specialization index for a society is obtained by taking the sum across all tasks and then standardizing the resulting sum.[20] To explore the link between kinship intensity and trade, we analyzed a measure of market participation and one of intercommunity trade in food from the Standard Cross-Cultural Sample (SCCS; (Kirby et al., 2016)). Market participation is a binary variable that takes the value of 1 if production is more oriented toward exchange than for consumption. The degree of intercommunity trade in food can take one of four values: no trade or food imports absent (0), less than 10 percent (1), less than 50 percent (2), or more than 50 percent (3).

Table 8 reports results of OLS regressions of these three variables on the KII and control variables. Consistent with our expectations, columns 1-3 show a negative association between the KII and the specialization index: a one-standard-deviation increase in the KII is associated with a ~ 0.15 standard-deviation decrease in the specialization index. This association holds while controlling for the geographic controls (in all three columns) and is robust to controlling for the subsistence variables (in columns 2 & 3) and country fixed effects (in column 3).[21] To address concerns about the non-independence of error terms due to shared cultural histories, we conservatively clustered standard errors at the

---

[20] In our regression analysis, we included all societies that had data entries for at least 8 tasks. For each ethnicity, tasks with missing data or that were coded as absent for the society were coded as 0; we included in our regressions 11 dummy variables indicating whether a task had missing data and 11 other dummy variables indicating if a task was coded as absent for each society.

[21] We report a regression with country fixed effects for consistency with our main analyses above, even though both the KII and the specialization index capture ancestral characteristics measured at a time when many contemporary countries did not exist.

language-family level. [22]

Focusing on trade, columns 4 and 5 regress market participation (in a linear probability model) and intercommunity trade in food on the KII, the geographic controls, and the subsistence dummies.[23] We find the expected negative relationship between kinship intensity and market participation (in column 5), but find no relationship between kinship intensity and intercommunity trade in food. That null result may reflect that among intensive kinship societies intercommunity trade in food may occur among established exchange partners with enduring personal relationships, and thus may be an inappropriate measure for the overall degree of market exchange.

Moving from the level of ancestral societies to modern countries, column 1 of Table 9 shows the results of a cross-country regression of openness to trade—defined as the sum of imports and exports as a share of a country's GDP—on the KII and the geographic controls. As expected, countries with higher levels of kinship intensity trade less, though the association is not precisely estimated.

Table 8: Ethnicity-level KII, specialization, and market exchange

	(1) Specialization index	(2) Specialization index	(3) Specialization index	(4) Intercommunity food trade	(5) Market participation
KII	-0.141*	-0.144**	-0.150**	-0.003	-0.080**
	(0.075)	(0.068)	(0.069)	(0.050)	(0.031)
Geographic controls	yes	yes	yes	yes	yes
Subsistence dummies		yes	yes	yes	yes
Country FEs			yes		
Observations	651	651	639	170	93
R-squared	0.557	0.576	0.738	0.257	0.272
Number of clusters	125	125	125	74	49

Notes: Ethnicity-level regressions of the specialization index (columns 1–3), intercommunity food trade (column 4), and market production (column 5) on the KII. The geographic controls include temperature, precipitation, agricultural suitability, absolute latitude, elevation, ruggedness, distance to coast, and distance to nearest river or lake. The subsistence variables measure the weighted fraction of a region's ethnicities' subsistence economies that depend on hunting, fishing, animal husbandry, and agriculture (with gathering as the omitted category). In columns 1 to 3, we include dummy variables indicating whether a task had missing data and dummy variables indicating whether a task was coded as absent for each society. Standard errors clustered at the language-family level are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

6.3.2 Cultural psychology

A growing body of evidence associates greater kinship intensity and stronger family ties with less impersonal or outgroup trust (Enke, 2019; Schulz et al., 2019; Alesina and Giuliano, 2015). Trust, in turn, has been shown to contribute to economic growth (e.g., Algan and Cahuc (2013)). As Arrow (1972) noted, "Virtually every commercial transaction has within itself an element of trust... It can be plausibly argued that much of the economic backwardness in the world can be explained by the lack of mutual confidence."

---

[22] Clustering at the country level decreases the standard errors such that all coefficients for the specialization index become significant at the 1% level.

[23] We do not include country fixed effects because the SCCS was designed to sample mostly independent societies from around the world, and few of these are in the same country.

Column 2 of Table 9 shows the results of a cross-country regression of trust—defined as the percentage of respondents who believe that most people can generally be trusted—on the KII and the geographic controls. As expected, a kinship intensity is negatively associated with less trust.

We also regressed the frequency of anonymous blood donations on the KII (and the geographic controls). Anonymous blood donations are a classic public good and are a measure of impersonal cooperation. Consistent with prior work, including similar cross-country regressions and within-country analyses in Italy (Schulz et al., 2019), kinship intensity is associated with fewer blood donations across countries.

Trust and impersonal cooperation are only two aspects of psychology that have been linked to kin-based institutions. In Schulz et al. (2019), we document that lower levels of kinship intensity are associated with cultural-psychological bundles that also include higher levels of impartiality, higher levels of individualism and independence, and lower levels of conformity and obedience; (Enke, 2019) and **?** document similar associations. The impartiality bundle includes lower levels of nepotism and less fierce loyalty to friends and family over impersonal institutions and strangers. It is thus a likely contributor to better formal economic and political institutions which, in turn, contribute to economic development, as we discuss below. Further, as Gorodnichenko and Roland (2016) argue, more individualism and less conformity and obedience may foster economic growth in part through their impact on innovation: individualistic cultures create incentive for innovation by awarding greater social status to personal accomplishments such as making important discoveries. Further, cultures with looser social norms and higher tolerance for deviance also tend to exhibit more creativity (e.g., Jackson et al. (2019)).

Table 9: Cross-country evidence on mechanisms

	(1) Openness to trade	(2) Trust	(3) Blood donations	(4) Expropriation Risk	(5) Institutional quality	(6) CPI	(7) Articles per m.	(8) Patents per m.
KII	-0.040*	-0.026**	-2.399***	-0.424***	-0.116***	-32.508***	-94.045***	-14.847
	(0.022)	(0.012)	(0.895)	(0.105)	(0.036)	(9.389)	(28.716)	(12.554)
Geographic controls	yes	yes	yes	yes	yes	yes	yes	yes
Observations	151	70	142	86	79	154	159	75
R-squared	0.261	0.383	0.560	0.634	0.272	0.506	0.559	0.177

Notes: Country-level regression of measures of openness to trade, cultural psychology, institutions, and innovation on the KII. The eight outcome variables are defined in Appendix Table A.1. The geographic controls include temperature, precipitation, agricultural suitability, absolute latitude, elevation, ruggedness, distance to coast, and distance to nearest river or lake. "CPI" stands for "Corruption Perceptions Index". Robust standard errors are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

6.3.3 Institutions

There is a sizeable literature linking economic prosperity to well-functioning formal political and economic institutions (Coase, 1960; North et al., 2009; Acemoglu and Robinson, 2012). But what accounts for the fact that similar formal institutions operate very differently in different populations? Part of the answer may be that the strong loyalty demands towards kin in intensive kinship societies can undermine large-scale formal institutions (Greif, 2006a; Greif and Tabellini, 2017; Greif, 2006b; Henrich, 2020), as these require high levels of trust and impartiality and low levels of corruption to properly function.

Existing work has already shown that across the world, among countries, regions, and ancestral ethnicities, kinship intensity is associated with more autocratic rule and less participation in the political process (Schulz, 2022). Akbari et al. (2019) and Akbari et al. (2020) find that cousin marriage is associated with more corruption. We add to this body of evidence by separately regressing expropriation risk, institution quality[24], and the Corruption Perceptions Index (CPI) on the KII (and the geographic controls) at the cross-country level. (Appendix Table A.1 provides more details on these variables.) The results, in columns 4-6 of Table 9 suggest that greater kinship intensity is associated with higher expropriation risk, lower institutional quality, and greater perceptions of corruption.

6.3.4 Innovation

To test more directly for a link between kinship intensity and innovation, we regressed two measures of innovation on the KII (and the geographic controls): the number of scientific and technical journal articles per capita and the number of patents per capita. (See Appendix Table A.1 for details on these variables.) As expected, columns 7-8 of Table 9 reveal that kinship intensity is negatively associated with these two measures of innovation across countries, though the coefficient for patents per capita is imprecisely estimated.

6.3.5 Inbreeding depression (ID)

Inbreeding depression (ID) is the reduction in biological fitness and related traits (such as health, cognitive ability, and height) observed in the offspring of related individuals.[25] Higher levels of ID are to be expected in intensive-kinship societies, since these societies tend to have more marriages and mating among relatives. Indeed, our second main measure of kinship intensity is the inbreeding coefficient, which is a direct measure of inbreeding intensity.

To assess the potential impact of ID on economic development, we compared our estimate of the association between the inbreeding coefficient and economic development across ethnicities to estimates from previous research of the effect of ID on educational attainment. To our knowledge, the largest study to date of the effects of ID is Clark et al. (2019), whose software pipeline we used to compute F_{ROH}. Clark et al. (2019) estimated the effects of ID by regressing various individual-level outcomes on F_{ROH}. Their main estimates imply that the educational attainment (EA) of the offspring of second cousins is ~ 0.11 year lower (and their cognitive performance ~ 0.7 IQ points lower) than that of the offspring of unrelated individuals.[26] This implies that the direct effect of ID would reduce the average educational attainment of an ethnicity in which everyone were the offspring of second cousins by ~ 0.11 year.

We do not have comparable educational attainment data at the level of the Ethnologue ethnicities, but to benchmark that figure, we used data from Gennaioli et al. (2014); across

---

[24] This variable is constructed based on the World Bank's Enterprise Survey and Doing Business Report and captures institutional features conducive for business.

[25] In economics, Ottinger and Voigtländer (2021) demonstrate, using genealogy data of royal lineages, that the inbreeding coefficient is a strong predictor of past European monarchs' intellectual abilities, which ultimately had a sizeable effect on state performance during their rule.

[26] The inbreeding coefficient of the offspring of second cousins (0.15625) is a little smaller than the standard deviation of $\overline{F_{ROH}}$ across country-ethnicities (0.018).

1,190 subnational regions with nonmissing data for the year 2000, the mean average EA is 8.19 years and the standard deviation is 2.69 years. Thus, the direct effects of ID in the offspring of second cousins represents only $\sim \frac{1}{25}$ of the standard deviation in average EA across subnational regions.

By contrast, our estimate of -25.841 in column 3 of Table 7 implies that log luminosity for an ethnicity in which everyone is the offspring of second cousins is 0.40 units lower, conditional on log population density (and the other control variables). This represents nearly one fourth of the standard deviation of log luminosity residualized on log population density (1.76) across the 245 Ethnologue ethnicities in the analysis sample for that regression, and nearly half of the standard deviation of log luminosity residualized on log population density (0.89) across 1,209 subnational regions with nonmissing data for the year 2000 in the Gennaioli et al. (2014) data.

This exercise is necessarily imperfect due to the lack of data allowing for a direct comparison of the effect of ID to our estimates, but it illustrates that ID is unlikely to have sufficiently large effects to account for much of the association we document between kinship intensity and economic development.

Overall, a cluster of findings in the extant literature, along with our own exploratory analyses, suggest that kinship intensity may impact economic development indirectly, via its effects on the division of labor and comparative advantage, cultural psychology, institutions, and innovation, but not principally via ID. Future work should focus on nailing down these pathways.

7 Conclusion

We have established a robust and economically significant negative association between kinship intensity and economic prosperity. Our primary analyses used a pixel-level measure of economic prosperity based on satellite nighttime luminosity and a measure of kinship intensity based on pre-industrialization ethnographic data captured at the country-ethnicity level (the KII). In a battery of regressions, we compared pixels within the same countries and controlled for a wide range of geographic, epidemiological, and cultural variables. We conducted a spatial RD analysis at the borders between neighboring pairs of country-ethnicities, to mitigate concerns that unobservable confounders that tend to vary smoothly at the borders (such as geographic factors) may drive our results. We also replicated our main analyses with an alternative dependent variable—regional GDP per capita—and an alternative explanatory variable—a society's average inbreeding coefficient, proxied by $\overline{F_{ROH}}$ and estimated using contemporary genetic data. Lastly, we conducted most of these analyses both in a sample where ethnographic variables were matched to pixels using direct language matches and in a sample where the matches were based on language phylogenies. Throughout, the estimated association between kinship intensity and economic prosperity remained robust, sizeable, and stable in magnitude.

Turning to the possible causal pathways, we showed that confounders related to Christianity and European ancestry and reverse causality are unlikely to fully account for the association. We discussed existing and novel evidence consistent with kinship intensity negatively impacting prosperity via its negative effects on the division of labor and trade; cultural psychological variables such as trust, individualism, and impartiality; the quality of formal institutions; and innovation.

In the absence of a natural experiment that quasi-randomly assigns kin-based institutions across societies, establishing an indisputable causal link between kinship intensity and economic development remains elusive. Taken together, however, the evidence presented in this paper strongly suggests that kinship intensity has a negative causal effect on economic development.

Despite this, we caution against concluding that intensive kin-based institutions are less desirable and that policy should seek to dismantle them: in many places, intensive kin-based institutions play a critical role in providing a safety net and maintaining social order. Tight family network may also foster happiness and life satisfaction (Alesina Giuliano 2012). Thus, in addition to more precisely elucidating the mechanisms through which kin-based institutions impact economic prosperity, future research should seek ways to mitigate their negative effect on prosperity without undermining the benefits people derive can derive from them.

References

Abbasi-shavazi, M. J. and Mcdonald, P. Modernization or cultural inheritance: the practice of consanguineous marriage in Iran. (March):911–933, 2008.

Acemoglu, D. and Robinson, J. *Why nations fail: the origins of power, prosperity, and poverty*. Crown Publishers, New York, 2012.

Acemoglu, D., Robinson, J. a., and Johnson, S. Reversal of Fortune: Geography and Institutions in the Making of the Modern World Income Distribution. *The Quarterly Journal of Economics*, 117(4):1231–1294, 2002.

Acemoglu, D., Naidu, S., Restrepo, P., and Robinson, J. A. Democracy does cause growth. *Journal of political economy*, 127(1):47–100, 2019.

Akbari, M., Bahrami-Rad, D., and Kimbrough, E. O. Kinship, fractionalization and corruption. *Journal of Economic Behavior & Organization*, 166:493–528, 2019.

Akbari, M., Bahrami-Rad, D., Kimbrough, E. O., Romero, P. P., and Alhosseini, S. An experimental study of kin and ethnic favoritism. *Economic Inquiry*, 58(4):1795–1812, 2020.

Akyol, Ş. P. and Mocan, N. H. Education and consanguineous marriage. Technical report, National Bureau of Economic Research, 2020.

Alesina, A. and Giuliano, P. The power of the family. *Journal of Economic Growth*, 15 (2750):93–125, 2010.

Alesina, A. and Giuliano, P. Family ties and political participation. *Journal of the European Economic Association*, 9(5):817–839, 2011.

Alesina, A. and Giuliano, P. Family ties. In Aghion, P. and Durlauf, S. N., editors, *Handbook of Economic Growth*, volume 2, chapter 4, pages 177–215. Elsevier, 2014.

Alesina, A., Giuliano, P., and Nunn, N. On the origins of gender roles: women and the plough. *Quarterly Journal of Economics*, 128(2):469–530, 2013.

Alesina, A., Algan, Y., Cahuc, P., and Giuliano, P. Family Values and the Regulation of Labor. *Journal of the European Economic Association*, 13(4):599–630, 2015.

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

Algan, Y. and Cahuc, P. Trust and growth. *Annual Review of Economics*, 5(1):521–549, 2013.

Arrow, K. J. Gifts and exchanges. *Philosophy & Public Affairs*, 1(4):343–362, 1972.

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

Bahrami-Rad, D., Becker, A., and Henrich, J. Tabulated nonsense? Testing the validity of the Ethnographic Atlas. *Economics Letters*, 204:109880, 2021.

Bai, Y. and Kung, J. K. S. Diffusing knowledge while spreading god's message: Protestantism and economic prosperity in china, 1840-1920. *Journal of the European Economic Association*, 13(4):669–698, 2015.

Bau, N. Can policy change culture? Government pension plans and traditional kinship practices. *American Economic Review*, 111(6):1880–1917, 2021.

Bau, N. and Fernández, R. The family as a social institution. National Bureau of Economic Research Working Paper No. 28918, 2021.

Becker, S. O. and Woessmann, L. Was Weber wrong? A human capital theory of Protestant economic history. *Quarterly Journal of Economics*, 124(2):531–596, 2009.

Bergeron, A. Religion and the Scope of Morality: Evidence from Exposure to Missions in the D.R. Congo. Working Paper, 2020.

Bittles, A. H. and Black, M. L. Consanguinity, human evolution, and complex diseases. *Proceedings of the National Academy of Sciences*, 107(suppl_1):1779–1786, 2010.

Bittles, R. H. A. H. The prevalence and demographic characteristics of consanguineous marriages in Pakistan. (1998):261–275, 2022.

Calvi, R. and Mantovanelli, F. G. Literacy and Economic Development: Evidence from Indian Districts. 2016.

Cameron, A. C. and Miller, D. L. A Practitioner's Guide to Cluster-Robust Inference. *Journal of Human Resources*, 50(2):317–372, 2015.

Carmichael, S. and Rijpma, A. Blood is thicker than water: Geography and the dispersal of family characteristics across the globe. *Cross-Cultural Research*, 51(2):142–171, 2017.

Cavalli-Sforza, L. and Bodmer, W. *The genetics of human populations.* Dover Publications, Inc., Mineola, New York, 1999.

Ceballos, F. C., Joshi, P. K., Clark, D. W., Ramsay, M., and Wilson, J. F. Runs of homozygosity: Windows into population history and trait architecture. *Nature Reviews Genetics*, 19(4):220–234, 2018.

Center for International Earth Science Information Network (CIESIN) at Columbia University. Gridded Population of the World, Version 4 (GPWv4): Population Count Adjusted to Match 2015 Revision of UN WPP Country Totals. NASA Socioeconomic Data and Applications Center (SEDAC), 2016.

Clark, D. W., Okada, Y., Moore, K. H., Mason, D., Pirastu, N., et al. Associations of autozygosity with a broad range of human phenotypes. *Nature communications*, 10 (4957):1–17, 2019.

Clark, G. *A Farewell to Alms.* Princeton University Press, Princeton, 2007.

Coase, R. H. The problem of social cost. *The Journal of Law and Economics*, 3:1–44, 1960.

Comin, D., Easterly, W., and Gong, E. Was the Wealth of Nations Determined in 1000 B. C.? *American Economic Journal: Macroeconomics*, 2(3):65–97, 2010.

Dalton, J. T. and Leung, T. C. Why is polygyny more prevalent in western africa?: An african slave trade perspective. *Economic Development and Cultural Change*, 62(4): 599–632, 2014.

Daynes, K. M. *More wives than one: transformation of the Mormon marriage system, 1840-1910.* University of Illinois Press, Urbana, 2001.

de Moor, T. and van Zanden, J. L. Girl power: the European marriage pattern and labour markets in the North Sea region in the late medieval and early modern period. *The Economic History Review*, 63(1):1–33, 2010.

Dell, B. M., Jones, B. F., and Olken, B. A. Temperature Shocks and Economic Growth: Evidence from the Last Half Century. 4(3):66–95, 2012.

Dell, M. The Persistent Effects of Peru's Mining Mita. *Econometrica*, 78(6):1863–1903, 2010.

Diamond, J. M. *Guns, germs, and steel: the fates of human societies.* W.W. Norton & Co, 1997.

Dohmen, T., Enke, B., Falk, A., Huffman, D., and Sunde, U. Patience and the wealth of nations. *Working paper*, pages https://www.iame.uni–bonn.de/people/thomas–dohmen/, 2015.

Donaldson, D. and Storeygard, A. The view from above: Applications of satellite data in economics. *Journal of Economic Perspectives*, 30(4):171–198, 2016.

Easterly, W. and Levine, R. The European origins of economic development. *Journal of Economic Growth*, 21(3):225–257, 2016.

Edlund, L. Cousin Marriage Is Not Choice: Muslim Marriage and Underdevelopment. *AEA Papers and Proceedings*, 108:353–57, 2018.

Ember, M. and Ember, C. R. The conditions favoring matrilocal versus patrilocal residence. *American Anthropologist*, 73(3):571–594, 1971.

Enke, B. Kinship, cooperation, and the evolution of moral systems. *The Quarterly Journal of Economics*, 134(2):953–1019, 2019.

Fafchamps, M. Development, social norms, and assignment to task. *Proceedings of the National Academy of Sciences of the United States of America*, 108(Supplement 4): 21308–21315, 2011.

Falconer, D. and Mackay, T. F. *Introduction into quantitative genetics.* Prentice Hall, Essex, 4th edition, 1996.

Fenske, J. African polygamy: Past and present. *Journal of Development Economics*, 117: 58–73, 2015.

Galor, O. and Özak, Ö. The agricultural origins of time preference. *American Economic Review*, 10:3064–3103, 2016.

Gelman, A. and Imbens, G. Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs. *Journal of Business and Economic Statistics*, 37(3): 447–456, 2019.

Gennaioli, N., La Porta, R., De Silanes, F. L., and Shleifer, A. Growth in regions. *Journal of Economic growth*, 19(3):259–309, 2014.

Ghosh, A., Hwang, S., and Squires, M. Economic consequences of kinship: Evidence from US bans on cousin marriage. Working Paper, 2021.

Giuliano, P. and Nunn, N. Ancestral characteristics of modern populations. *Economic History of Developing Regions*, 33(1):1–17, 2018.

Glaeser, E. L., La Porta, R., Lopez de Silanes, F., and Shleifer, A. Do Institutions Cause Growth? *Journal of Economic Growth*, 9(3):271–303, 2004.

Goody, J. *The development of the family and marriage in Europe: Past and present publications.* Cambridge University Press, Cambridge; New York, 1983.

Gordon, R. G., Jr., editor. *Ethnologue: languages of the world.* SIL International, 15th edition, 2005.

Gorodnichenko, Y. and Roland, G. Culture, institutions, and the wealth of nations. *The Review of Economics and Statistics*, 99(3):402–416, 2016.

Greif, A. Family Structure, Institutions, and Growth: The Origins and Implications of Western Corporations. *American Economic Review*, 96(2):308–312, 2006a.

Greif, A. *Institutions and the path to the modern economy: Lessons from medieval trade.* Cambridge University Press, Cambridge, 2006b.

Greif, A. and Tabellini, G. Cultural and institutional bifurcation: China and Europe compared. *American Economic Review*, 100(2):135–140, 2010.

Greif, A. and Tabellini, G. The clan and the corporation: Sustaining cooperation in China and Europe. *Journal of Comparative Economics*, 45(1):1–35, 2017.

Guglielmino, C. R., Viganotti, C., Hewlett, B., and Cavallisforza, L. L. Cultural Variation in Africa - Role of Mechanisms of Transmission and Adaptation. *Proceedings of the National Academy of Sciences of the United States of America*, 92(16):7585–7589, 1995.

Hajnal, J. Two kinds of preindustiral household formation system. *Population and Development Review*, 8(3):449–494, 1982.

Hamer, D. H. Beware the chopsticks gene. *Molecular psychiatry*, 5(1):11–13, 2000.

Hammarström, H., Bank, S., Forkel, R., and Haspelmath, M. Glottolog 2.7. *Max Planck Institute for the Science of Human History*, 2016.

Henderson, J. V., Squires, T., Storeygard, A., and Weil, D. The global distribution of economic activity: Nature, history, and the role of trade. *The Quarterly Journal of Economics*, 133(1):357–406, 2018.

Henrich, J. *The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter.* Princeton University Press, Princeton, 2016.

Henrich, J. *The WEIRDest People in the World: How the West became psychologically peculiar and particularly prosperous.* Farrar, Straus and Giroux, 2020.

Hibbs, D. A. and Olsson, O. Geography, biogeography, and why some countries are rich and others are poor. *Proceedings of the National Academy of Sciences of the United States of America*, 101(10):3715–3720, 2004.

Hoff, K. and Sen, A. The Kin System as a Poverty Trap? In Bowles, S., Durlauf, S. N., and Hoff, K., editors, *Poverty Traps*, chapter 4, pages 95–115. Princeton University Press, 2011.

Holden, C. J. and Mace, R. Spread of cattle led to the loss of matrilineal descent in Africa: a coevolutionary analysis. *Proceedings of the Royal Society B: Biological Sciences*, 270 (1532):2425–2433, 2003.

Islam, M. M. The practice of consanguineous marriage in Oman: prevalence, trends, and determinants. (2012):571–594, 2022.

Jackson, J. C., Gelfand, M., De, S., and Fox, A. The loosening of american culture over 200 years is associated with a creativity–order trade-off. *Nature human behaviour*, 3 (3):244–250, 2019.

Jeong et al., C. The genetic history of admixture across inner Eurasia. *Nature Ecology and Evolution*, 3(6):966–976, 2019.

Jones, D. The matrilocal tribe: An organization of demic expansion. *Human Nature*, 22 (1–2):177–200, 2011.

Jordan, F. M., Gray, R. D., Greenhill, S. J., and Mace, R. Matrilocal residence is ancestral in austronesian societies. *Proceedings of the Royal Society B-Biological Sciences*, 276 (1664):1957–1964, 2009.

Joshi et al., P. K. Directional dominance on stature and cognition in diverse human populations. *Nature*, 523(7561):459–462, 2015.

Jurdi, R. and Saxena, P. C. The prevalence and correlates of consanguineous marriages in Yemen: similarities and contrast with other Arab countries. pages 1–13, 2003.

Kang, J. T., Goldberg, A., Edge, M. D., Behar, D. M., and Rosenberg, N. A. Consanguinity Rates Predict Long Runs of Homozygosity in Jewish Populations. *Human Heredity*, 82(3-4):87–102, 2017.

Khaldûn, I. *The Muqaddimah: An introduction to history*, 1377.

Kirby, K., Gray, R. D., Greenhill, S. J., Jordan, F. M., Gomes-Ng, S., et al. D-PLACE: A global database of cultural, linguistic and environmental diversity. *PLoS One*, 11 (7):e0158391, 2016.

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

Korotayev, A. Parallel-cousin (FBD) marriage, Islamization and Arabization. 39(4): 395–407, 2015.

Korotayev, A. V. Unilineal descent organization and deep Christianization: a cross-cultural comparison. *Cross-Cultural Research*, 37(1):133–157, 2003.

Landes, D. Culture Makes Almost All the Difference. In Harrison, L. E. and Huntington, S. P., editors, *Culture Matters: How values shape human progress*, chapter 1, pages 2–13. Basic Books, New York, 2000.

Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., et al. Ancient human genomes suggest three ancestral populations for present-day europeans. *Nature*, 513: 409–413, 2014.

Lazaridis, I., Nadel, D., Rollefson, G., Merrett, D. C., Rohland, N., et al. Genomic insights into the origin of farming in the ancient near east. *Nature*, 536:419–424, 2016.

Leach, E. Reply to raoul naroll's "on ethnic unit classification". *Current Anthropology*, 5(4):283–312, 1964.

Leutenegger, A.-L., Sahbatou, M., Gazal, S., Cann, H., and Génin, E. Consanguinity around the world: what do the genomic data of the HGDP-CEPH diversity panel tell us? *European journal of human genetics : EJHG*, 19(5):583–587, 2011.

Lewis, B. *The Muslim discovery of Europe*. W. W. Norton & Company, New York, kindle edition, 2001.

Lewis, M. P., editor. *Ethnologue: Languages of the World*. SIL International, Dallas, Texas, 16 edition, 2009.

Lipson, M., Cheronet, O., Mallick, S., Rohland, N., Oxenham, M., et al. Ancient genomes document multiple waves of migration in Southeast Asian prehistory. *Science*, 361 (6397):92–95, 2018.

Lowes, S. Matrilineal Kinship and Spousal Cooperation: Evidence from the Matrilineal Belt. *Mimeo.*, (September), 2018.

Machiavelli, N. *Discourses on Livy*, 1531.

Mauro, P. Corruption and growth. *The quarterly journal of economics*, 110(3):681–712, 1995.

McQuillan, R., Leutenegger, A.-L., Abdel-Rahman, R., Franklin, C. S., Pericic, M., et al. Runs of Homozygosity in European Populations. *American Journal of Human Genetics*, 83(3):359–372, 2008.

Michalopoulos, S. and Papaioannou, E. Pre-colonial Ethnic Institutions and Contemporary African Development. *Econometrica*, 81(1):113–152, 2013.

Michalopoulos, S. and Papaioannou, E. Spatial patterns of development: A meso approach. *Annual Review of Economics*, 10:383–410, 2018.

Mitterauer, M. *Why Europe?: the medieval origins of its special path.* University of Chicago Press, Chicago, 2010.

Mokyr, J. *A culture of growth: the origins of the modern economy.* Princeton University Press, Princeton, 2016.

Moscona, J., Nunn, N., and Robinson, J. A. Keeping it in the Family: lineage organization and the scope of trust in Sub-saharan Africa, 2017.

Moscona, J., Nunn, N., and Robinson, J. A. Segmentary Lineage Organization and Conflict in Sub-Saharan Africa. *Econometrica*, 88(5):1999–2036, 2020.

Murdock, G. P. *Social structure.* Free Press, New York, 1949.

Nakatsuka, N., Moorjani, P., Rai, N., Sarkar, B., Tandon, A., et al. The promise of discovering population-specific disease-associated genes in South Asia. *Nature Genetics*, 49:1403–1407, 2017.

Noblit, G. *The Origin and Evolution of Chinese Lineages.* 2021.

North, D. C., Wallis, J. J., and Weingast, B. R. *Violence and Social Orders: A Conceptual Framework for Interpreting Recorded Human History.* Cambridge University Press, New York, 2009.

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., et al. Genes mirror geography within Europe. *Nature*, 456(7218):98–101, 2008.

Nunn, N. The long-term effects of Africa's slave trades. *Quarterly Journal of Economics*, 123(1):139–176, 2008.

Nunn, N. Religious conversion in Colonial Africa. *American Economic Review*, 100(2): 147–152, 2010.

Nunn, N. Culture and the historical process. *Economic History of Developing Regions*, 27(Supplement 1):108–126, 2012.

Ottinger, S. and Voigtländer, N. History's masters: The effect of european monarchs on state performance. Technical report, National Bureau of Economic Research, 2021.

Parkin, D. *Kinship: An Introduction to the Basic Concepts.* Wiley, 1997.

Pehrson, R. N. *The Social Organization of the Marri Baluch.* Viking Fund publications in anthropology. Indus Publications, Karachi, 1977.

Pemberton, T. J. and Rosenberg, N. A. Population-Genetic Influences on Genomic Estimates of the Inbreeding Coefficient: A Global Perspective. *Human Heredity*, 77(1–4): 37–48, 2014.

Pickrell, J. K., Patterson, N., Barbieri, C., Berthold, F., Gerlach, L., et al. The genetic prehistory of southern Africa. *Nature Communications*, 3(1143):1–6, 2012.

Platteau, J. P. *Institutions, social norms, and economic development.* Harwood Academic Publishers, Amsterdam, The Netherlands, 2000.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, 2006.

Ramachandran, S., Deshpande, O., Roseman, C. C., Rosenberg, N. A., Feldman, M. W., et al. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *PNAS*, 102(44):15942–15947, 2005.

Reich Lab. Human origins dataset, 2020.

Rodrik, D., Subramanian, A., and Trebbi, F. Institutions rule: The primacy of institutions over geography and integration in economic development. *Journal of Economic Growth*, 9(2):131–165, 2004.

Romeo, G. and Bittles, A. H. Consanguinity and Genomics. *Human Heredity*, 77(1-4): 5–5, 2014.

Roy, S. C. *Oraons of Chota Nagpur : Their history, economic life, and social organization.* The Brahmo Mission Press, Calcutta, 1915.

Sahlins, M. D. The Segmentary Lineage: An Organization of Predatory Expansion. *American Anthropologist*, 63(2):322–345, 1961.

Sahoo, S. A., Zaidi, A. A., Anagol, S., and Mathieson, I. Long runs of homozygosity are correlated with marriage preferences across global population samples. Working Paper. Available at bioRxiv, 2021.

Sarma, N., Patouillard, E., Cibulskis, R. E., and Arcand, J. L. The economic burden of Malaria: Revisiting the evidence. *American Journal of Tropical Medicine and Hygiene*, 101(6):1405–1415, 2019.

Schulz, J., Bahrami-Rad, D., Beauchamp, J. P., and Henrich, J. The origins of WEIRD psychology. 2018.

Schulz, J. F. Kin networks and Institutional Development. *Economic Journal*, 2022.

Schulz, J. F., Bahrami-Rad, D., Beauchamp, J. P., and Henrich, J. The Church, intensive kinship, and global psychological variation. *Science*, 366(6466), 2019.

Smith, A. *The Wealth of Nations*, 1776.

Spolaore, E. and Wacziarg, R. How deep are the roots of economic development? *Journal of Economic Literature*, 51(2):325–69, 2013.

Tabellini, G. Culture and Institutions: Economic Development in the Regions of Europe. *Journal of the European Economic Association*, 8(4):677–716, 2010.

Tène, E. On the Historical Roots of Gender Norms : Evidence from Matrilineal Societies in Sub-Saharan Africa. 2021.

van Zanden, J. L., De Moor, T., and Carmichael, S. *Capital Women: The European Marriage Pattern, Female Empowerment and Economic Development in Western Europe 1300-1800.* Oxford University Press, 2019.

Walker, R. S. and Hill, K. R. Causes, consequences, and kin bias of human group fissions. *Human Nature*, 25(4):465–475, 2014.

Walker, R. S., Beckerman, S., Flinn, M. V., Gurven, M., von Rueden, C. R., et al. Living with kin in lowland horticultural societies. *Current Anthropology*, 54(1):96–103, 2013.

World GeoDatasets. World Language Mapping System (WLMS) Version 19. SIL International.

Kinship Systems and Economic Development

ONLINE APPENDIX

Duman Bahrami-Rad[*] Jonathan Beauchamp[†] Jonathan Schulz[‡]
Joseph Henrich[§]

April 11, 2022

---

[*] Department of Economics, St. Olaf College.

[†] Interdisciplinary Center for Economic Science and Department of Economics, George Mason University.

[‡] Department of Economics, George Mason University.

[§] Department of Human Evolutionary Biology, Harvard University.

A Data Appendix

A.1 Variables construction

Pixel-level variables:

Variable name	**Description**
Nighttime luminosity	Nightlight luminosity (light emitted from a pixel divided by pixel area, also commonly referred to as light density) measured by satellites in the year 2010-2011. The Global Radiance Calibrated Nighttime Lights data is gridded with an output resolution of 30 arc-seconds. We aggregate the data to 0.125 × 0.125 degree pixels by taking the mean of 225 constituent pixels from raw data, and then divide it by fraction of component pixels covered by land. Source: Elvidge et al. (1999).
Population density	Population density (number of persons per square kilometer) in year 2010. The Gridded Population of the World (GPWv4) collected population input data at the most detailed spatial resolution available from the results of the 2010 round of Population and Housing Censuses, which occurred between 2005 and 2014. The input data are extrapolated to produce population estimates for different years. The data is adjusted to Match 2015 Revision of UN WPP Country Totals. GPWv4 is gridded with an output resolution of 30 arc-seconds. We aggregate the data to 0.125 × 0.125 degree pixels by taking the mean of 225 constituent pixels from raw data. Source: (Center for International Earth Science Information Network (CIESIN) at Columbia University, 2016).
Land and water area	Surface areas of land and water in square kilometers per pixel. The Land Area raster provides estimates of the land area, excluding permanent ice and water, within each pixel, and was used to calculate the population density rasters. The Water Area raster provides estimates of the water area (permanent ice and water) within each pixel. The sum of land area and water area of a pixel equals the total surface area of that pixel. The data files were produced as global rasters at 30 arc-second resolution. Therefore, we aggregate the data to 0.125 × 0.125 degree pixels by taking the sum of 225 constituent pixels from raw data. Source: (Center for International Earth Science Information Network (CIESIN) at Columbia University, 2016).
Temperature/Precipitation	Mean monthly temperature/precipitation over the period 1901-2017. The gridded Climatic Research Unit (CRU) Time-series (TS) data version 4.02 data are month-by-month variations in climate over the period January 1901 - December 2017, based on monthly observational data calculated from daily or sub-daily data by National Meteorological Services and other external agents, provided on 0.5 × 0.5 degree cells. Therefore, the 16 constituent 0.125 × 0.125 degree pixels of each cell are assigned with the same value. Source: University of East Anglia Climatic Research Unit; University of East Anglia Climatic Research Unit et al. (2017): CRU TS4.02.

Elevation	Mean elevation in 100 meters. The data is provided in a 30 arc-second resulution map by the Global 30 Arc-Second Elevation data set. We aggregate the data to 0.125 × 0.125 degree pixels by taking the mean of 225 constituent pixels from raw data. Source: U.S. Geological Survey, Global Digital Elevation Model (GTOPO30).
Ruggedness	Mean terrain ruggedness. As a measure of the local variance in elevation, terrain ruggedness at a point (an individual cell) is defined as the square root of the sum of the squared differences in elevation between the central point and the eight adjacent points. Formally, it is defined as follows. Let $e_{r,c}$ denote elevation at the point located in row r and column c of a grid of elevation points. Then the Terrain Ruggedness Index for this point is calculated as $\sqrt{\Sigma_{j=c-1}^{i=c+1}\Sigma_{i=r-1}^{i=r+1}(e_{i,j} - e_{r,c})^2}$. The data is calculated at the level of 30 arc-second cells on a regular geographic grid covering the Earth. The data is provided with the resolution of 30 arc-seconds. We aggregate the data to 0.125 degree pixels by taking the mean of 225 constituent pixels from raw data. Source: Nunn and Puga (2012).
Suitability for agriculture	The fraction of each grid cell that is suitable to be used for agriculture. It is based on the temperature and soil conditions of each grid cell. The data is constructed based on the global map (0.5 × 0.5 degree cells) obtained from Suitability for Agriculture. Therefore, the 16 constituent 0.125 degree pixels of each cell are assigned with the same value. Source: Ramankutty et al. (2002), Atlas of the Biosphere.
Distance to the nearest coast	Distances in kilometers from 0.125 degree pixel centroids to the nearest coast. Source: Natural Earth, Coastline.
Distance to the nearest river or lake	Distances in kilometers from 0.125 degree pixel centroids to the nearest river or lake centerline. Source: Natural Earth, River and Lake Centerline.
Malaria index	An index representing the contribution of regionally dominant biological characteristics of vector mosquitoes to the force of malaria transmission. The data is constructed based on the world map of 0.5 × 0.5 degree cells. Source: (Kiszewski et al., 2004).

Ethnicity-level variables: **Variable name**	**Description**
Cousin marriage preference	The variable is created based on entry EA026, then linked to contemporary populations and standardized. The variable takes value 0 if cousin marriage is not preferred, value 1 if second- (but not first-) cousin marriage is preferred, value 2 if cross-cousin marriage is preferred, value 3 if parallel-cousin marriage is preferred. To enlarge the sample, we predicted cousin marriage preference based on cousin terms in entry EA027 that contains data on cousin terms classified into eight categories (Crow, Descriptive, Eskimo, Hawaiian, Iroquois, Omaha, Sudanese, and Mixed). To predict cousin marriage preference, we first analyzed the sample of ethnicities with data on both cousin marriage preference and cousin terms. For each cousin term category, we calculated the mean of the cousin marriage sub-indicator. We used the resulting mean values to predict the expected cousin marriage preference sub-indicator for ethnicities where only data on cousin terms is available. Source: Ethnographic Atlas (Murdock, 1967; Kirby et al., 2016).
Polygamy	The variable is created based on entry EA009, then linked to contemporary populations and standardized. The variable takes value 0 for monogamy, value 1 for occasional or limited polygyny, and value 2 for common polygyny and polyandary. Source: Ethnographic Atlas (Murdock, 1967; Kirby et al., 2016).
Co-residence of extended families	The variable is created based on entries EA008 and EA012, then linked to contemporary populations and standardized. We created a variable for domestic organization based on EA008 that takes value 0 for nuclear and polygamous families, 1 for minimal extended or stem families, 2 for small extended families and 3 for large extended families. We also created a second variable, marital residence based on EA012 that takes the value 0 for neolocality (spouses move to a new location), 1 for ambilocality (individuals can choose which side of the family they relocate to), and 2 for the remaining cases (e.g., patrilocality or matrilocality). We take the average of the domestic organization and marital residence variables to create the sub-indicator for the co-residency of extended families. Source: Ethnographic Atlas (Murdock, 1967; Kirby et al., 2016).
Lineage organization	The variable is created based on entry EA043, then linked to contemporary populations and standardized. The variable takes value 0 for bilateral descent and 1 otherwise (i.e., patrilineal, matrilineal, duolateral, quasi-lineage, and ambilineal descent). Source: Ethnographic Atlas (Murdock, 1967; Kirby et al., 2016).
Community organization	The variable is created based on entry EA015, then linked to contemporary populations and standardized. The variable takes value 1 if localized clans are present and/or if community endogamy exists, and takes value 0 if both are absent (i.e., agamous communities without localized clans and exogamous communities without clan structure). Source: Ethnographic Atlas (Murdock, 1967; Kirby et al., 2016).

KII	Following Schulz et al. (2019), we standardize the five sub-indicators of Kinship Intensity Index: cousin marriage preference, Polygamy, Co-residence of extended families, Lineage organization, Community organization. Then we take the average of the five standardized sub-indicators.
Subsistence variables (gathering, hunting, fishing, animal husbandry, and agriculture)	The data is linked to contemporary populations using entries EA001-EA005 of the Ethnographic Atlas. The five variables measure the fraction of a population's economic subsistence that depends on gathering, hunting, fishing, animal husbandry, and agriculture, respectively. Each entry takes the values 0 to 9 respectively for 0–5%, 6–15%, 16–25%, 26–35%, 36–45%, 46–55%, 56–65%, 66–75%, 76–85%, 86–100% dependence on the activity. We normalized the variables separately for each population so that they sum to 1 for each population. Source: Ethnographic Atlas (Murdock, 1967; Kirby et al., 2016).
Settlement complexity	The data is linked to contemporary populations using entry EA030. The variable classifies "settlement patterns" in the following categories: 1- nomadic; 2- seminomadic; 3- semisedentary; 4- impermanent; 5- dispersed homesteads; 6- hamlets; 7- village/town; 8- complex permanent. Source: Ethnographic Atlas (Murdock, 1967; Kirby et al., 2016).
Political hierarchies	The data is linked to contemporary populations using entry EA033 of the Ethnographic Atlas. The variable classifies "jurisdictional hierarchy beyond local community" in the following categories: 1- acephalous (e.g., autonomous bands and villages); 2- one level (e.g., petty chiefdoms); 3- two levels (e.g., larger chiefdoms); 4- three levels (e.g., states); 5- four levels (e.g., large states). Source: Ethnographic Atlas (Murdock, 1967; Kirby et al., 2016).
Migratory distance from Addis Ababa	The potential minimum travel time (on land only, and measured in months of travel) from Addis Ababa to the centroid of an ethno-lingustic group. Centroids of ethno-linguistic groups are defined using polygons provided by World Language Mapping System (WLMS 19th) for all languages of the world as defined by the Ethnologue 16th (Lewis, 2009). The raw data for travel time on land (Human Mobility Index) is defined for 1 km × 1 km cells. Whenever two pieces of land are disconnected by waters, we connect them manually in the closest linear distance using values from adjacent cells. Source: Özak (2018), Human Mobility Index (HMI).
Average inbreeding coefficient ($\overline{F_{ROH}}$)	Average inbreeding coefficient computed across the Human Origins (HO) individuals matched to each ethnicity. Individual inbreeding coefficients were computed using the ROHgen2 software pipeline developed by the ROHgen consortium (Clark et al., 2019). Source: computed using the HO data for this paper; see Appendix C for details.

Mean regional pairwise F_{ST}	Mean regional pairwise F_{ST} was first computed for each Human Origins (HO) population with at least five individuals, as the average of the pairwise F_{ST} between that population and all other populations in the same region with at least eight individuals in the (QCed) data. Then, each HO individual was assigned their HO population's mean regional pairwise F_{ST} estimate, and each Ethnologue ethnicity's mean regional pairwise F_{ST} was computed by taking the average its matched individuals. Source: computed using the HO data for this paper; see Appendix C for details.
Expected heterozygosity	For each Human Origins (HO) population j with at least five individuals, expected heterozygosity was computed for each SNP i using Nei and Roychoudhury's formula (Nei and Roychoudhury, 1974); the HO population j's expected heterozygosity was then estimated by summing up the estimated heterozygosities across all SNPs. Then, each HO individual was assigned their HO population's expected heterozygosity estimate, and each Ethnologue ethnicity's expected heterozygosity was computed by taking the average across its matched individuals. Source: computed using the HO data for this paper; see Appendix C for details.
Top 20 PCs	The top 20 principal components (PCs) of the genotypic data (i.e., of the $N \times N$ matrix with the entry in row i and column j equal to the correlation between individuals i and j's SNP data) were computed for each Human Origins (HO) individual. For each Ethnologue ethnicity, the mean of each PC was then computed across the ethnicity's matched individuals. Source: computed using the HO data for this paper; see Appendix C for details.
Specialization index	The Ethnographic Atlas (EA) contains data on labor specialization across 11 different tasks: metal working, weaving, leather working, pottery, boat building, house building, gathering, hunting, fishing, animal husbandry, and agriculture. For each task and society the EA codes for non-missing observations whether (1) the task is absent, or unimportant in the particular society, (2) Junior age specialization, i.e., the activity is largely performed by boys and/or girls before the age of puberty (3) Senior age specialization, i.e., the activity is largely performed by men and/or women beyond the prime of life (4) Normally performed by many or most adult men, women, or both (5) Craft specialization, i.e., the activity is largely performed by a small minority of adult males or females who possess specialized skills. Occupational castes are treated as instances of craft specialization (6) Industrial specialization, i.e., the activity is largely removed from the domain of a division of labor by sex, age, or craft specialization and is performed industrialized mainly by techniques of production. We reclassified by assigning each task one of three values: task is absent of unimportant, no or only age specialization (0), craft specialization (1), or industrial specialization (2). In addition, we assigned missing observations the value of 0 (note that in our regression analysis we control for missing observations and whether the task is absent in a society). We then took the sum across all tasks and standardized the resulting sum.

Market participation	This variable is taken from the Standard Cross Cultural Sample (SCCS). It is a binary variable that codes the relationship between production for subsistence and production for market exchange. According to the SCCS, it takes the value of 0 if production for consumption is more important and the value of 1 if production for consumption and production for market exchange are of equal importance.
Intercommunity food trade	This indicator is based on the variable ”Intercommunity trade as a foodsource” of the SCCS. We coded the variable such that it takes the value 0 for the categories ’no trade’ or ’no food imports’, 1 for the category ’less than 10 percent of food’, 2 for ’less than 50 percent of food’ and 3 for ’more than 50 percent of food’.

Region-level variables: **Variable name**	**Description**
Regional GDP per capita	Gross domestic product per capita in the region (in constant 2005 PPP dollars). For each country, regional GDP per capita is scaled so that their population-weighted sum equals the value of Gross Domestic Product reported in Penn World Tables or, when unavailable, World Development Indicators. Similarly, for each country, the regional population values are adjusted so that their sum equals the country-level analog in World Development Indicators. Source: Gennaioli et al. (2014)
Capital is in region	Dummy equal to 1 if the region contains a national capital, 0 otherwise, using ESRI World Cities: `http://www.esri.com/data/data-maps`. Source: Gennaioli et al. (2014)
Years of education	The average years of schooling from primary school onward for the population aged 15 years or older. To make levels of educational attainment comparable across countries, educational statistics are translated into the International Standard Classification of Education (ISCED) standard and use UNESCO data on the duration of school levels in each country for the year for which we have educational attainment data. Eurostat aggregates data for ISCED levels 0-2 and we assign such observations an ISCED level 1. Following Barro and Lee (1993): (1) zero years of schooling are assigned to ISCED level 0 (i.e., pre-primary); (2) zero years of additional schooling are assigned to (a) ISCED level 4 (i.e., vocational), and (b) ISCED level 6 (i.e., post-graduate); and (3) 4 years of additional schooling are assigned to ISCED level 5 (i.e., graduate). Since regional data is not available for all countries, unlike Barro and Lee (1993), zero years of additional schooling are assigned: (a) to all incomplete levels; and (b) to ISCED level 2 (i.e., lower secondary). Thus, the average years of schooling in a region is calculated as: (1) the product of the fraction of people whose highest attainment level is ISCED 1 or 2 and the duration of ISCED 1; plus (2) the product of the fraction of people whose highest attainment level is ISCED 3 or 4 and the cumulative duration of ISCED 3; plus (3) the product of the fraction of people whose highest attainment level is ISCED 5 or 6 and the sum of the cumulative duration of ISCED 3 plus 4 years. See Gennaioli et al. (2014)'s online data appendix for sources and time periods. Source: Gennaioli et al. (2014)

Oil and gas production	Cumulative oil, gas and liquid natural gas production from the time production began to 2000. For onshore oil and gas, we calculated the production by region by allocating oil production to regions based on the fraction of the petroleum assessment areas within the region. Offshore assessment areas that are closer at all points to countries not included in the dataset than to countries included are removed. Offshore assessment areas that are in the area of countries within the dataset and countries not in the dataset are clipped at a distance of 100 km from the regions in the dataset. The assessment fields were then converted to a raster layer containing the cumulative production values in each cell, and the cells were allocated to regions based on the closest region. Oil and liquid natural gas were collected in millions of barrels. Gas was collected in billions of cubic feet and divided by 6 to convert to millions of barrels of oil equivalents. The datasets come from USGS World Petroleum Assessment Data. Source: Gennaioli et al. (2014)
Geographic variables	We aggregate the the pixel-level variables and data to region-level by taking the mean of constituent pixels. Source: Gennaioli et al. (2014)

Country-level variables:

Variable name	**Description**
Openness	Following Gennaioli et al. (2014), we calculated the sum of exports and imports as a share of GDP. Source: Penn World Tables.
Article per million	Following Gennaioli et al. (2014), we use number of scientific and technical journal articles and population from the World Bank to generate the variables. The variable is calculated as the average from 2000 to 2018. Source: The World Bank
Patents per million	The variable is produced by Economist Intelligence Unit (2007, 2009). Source: Gorodnichenko and Roland (2016)
Expropriation risk	The variable is provided by International Country Risk Guide. Risk of "outright confiscation and forced nationalization" of property. This variable ranges from 0 to 10 where higher values indicate a lower probability of expropriation. This variable is calculated as the average from 1982 through 1997. Source: Gennaioli et al. (2013)
Institutional quality	The variable is generated based on World Bank's Enterprise Surveys. Latent variable of: (1) (minus) Informal payments; (2) (minus) Ln(tax days); (3) (minus) Ln(days without electricity); (4) (minus) Security costs; (5) (minus) Access to land, (6) (minus) Access to finance; (7) Government predictability; and (8) (minus) Doing Business percentile rank. Higher values indicate better institutions. Source: Gennaioli et al. (2013)
CPI	Following Gennaioli et al. (2013), we use Corruption Perception Index as a measure of institutional quality. CPI ranges from 0 to 100 where lower values indicate a higher corruption. Source: Transparency International
Blood donations	Voluntary blood donations per 1,000 inhabitants is taken from the WHO Global Status on Blood Safety and Availability 2016. The report contains data for the years 2011, 2012, and 2013. Following Schulz et al. (2019), we use the averages across these years. When data for a year is missing for a country, the indicator is based on the non-missing years. In about a third of the countries, not all blood centers are covered. In these cases, we did the following: The WHO report states the estimated percentage of blood donations covered; we calculated the total amount of blood donated based on those estimates. In the non-fully covered countries, the coverage is still reasonably high: 60% of those countries have a coverage above 75%. The WHO report gives further information on the types of donations: (i) voluntary non-remunerated donations, (ii) family/replacement donations, (iii) paid donations, or (iv) other. Our main variable of interest is voluntary non-remunerated donations per 1,000 inhabitants. To calculate per capita values, we divided the number of voluntary non-remunerated donations by the World Bank's population estimates (per 1000) for the year 2012. Source: WHO (2016).

Trust	The percentage of respondents in the country who believe that most people can generally be trusted. Data comes from World Values Survey. The variable takes the most recent available year collected by Gennaioli et al. (2013), ranging from 1980 through 2005. Source: Gennaioli et al. (2013)

A.2 Matching data

This Appendix section complements Section 3.3 and provides more detail about the procedures we used to match the various datasets we analyzed. At a high level, we matched ethnographic variables from the EA and genetic variables from the HO dataset to languages from the Ethnologue; then, we matched the Ethnologue languages to geographic pixels; finally, for the analyses at the country-ethnicity, region, or country level, we collapsed the resulting pixel-level data to country-ethnicities, regions, or countries.

Matching ethnographic and genetic variables to the Ethnologue languages

As we describe in the main text, to match the KII and other EA-derived variables to the pixel-level data, we first matched each EA society to contemporary languages from the 23rd edition of the Ethnologue (Gordon, 2005; Lewis, 2009). We employed two different matching methods. The direct-matching method does one-to-one matches. With that method, the 911 EA societies with non-missing KII data are matched to 911 Ethnologue languages corresponding to 1,254 country-ethnicities (as some of the languages are spoken in multiple countries).

The language-tree matching method does many-to-one matches. For each EA variable, the vast majority of the 5,756 Ethnologue languages are matched to one of the EA societies with no missing data for the variable. Specifically, for each variable, each Ethnologue language is matched to the linguistically closest EA society that speaks a language within the same language family and with nonmissing data for the variable; when no such society exists for a language and variable, the variable is coded as missing for the language.

After matching the Ethnologue languages to geographic pixels and applying quality-control filters, as described below and in Section 3.1, we are left with 713 country-ethnicities with the direct matching method and 2,352 country-ethnicities with the language-tree matching method.

For the analyses using the inbreeding coefficients (F), we matched each population from the Human Origins (HO) dataset to a single Ethnologue language (and to the single or multiple country-ethnicities associated with that language). Appendix C.2 provides more details.

Matching to pixels

We used a shapefile provided by the World Language Mapping System (World GeoDatasets)[1] to match pixels to Ethnologue country-ethnicities. Of the 783,525 pixels that remain after applying the filters described in Section 3.1, 133,614 ($\sim$ 17%) could not be matched to any ethnicity (either because the pixel is unpopulated according to the Ethnologue or because Ethnologue data is missing).

For the pixel-level analyses, we applied three additional filters. We dropped (1) 52,782 pixels ($\sim$ 8%) that were matched to more than one country-ethnicity, as well as a further 104,965 ($\sim$ 18%) pixels that (2) fell on the boundaries of a country-homeland or (3) were matched to a country-ethnicity with a very small country-homeland comprising less than five pixels.

With the language-tree matching method, we dropped a further 3,606 pixels due to missing data for a KII indicator, thus leaving 488,558 pixels located in the country-homelands of 2,352 country-ethnicities (which account for $\sim$ 52% of the world population) around the world. The direct matching method left 377,656 pixels with nonmissing KII data and located in the country-homelands of 713 country-ethnicities (which account for $\sim$ 38% of the world population). And matching the HO data to pixels leaves 281,177 pixels located in the homelands of 245 country-ethnicities.

Matching to subnational regions, countries, and country-ethnicities

---

[1] The shapefile is available at worldgeodatasets.com/language

To match pixel-level data to country-ethnicities, subnational regions, or countries, we collapsed the pixel data—along with the ethnographic variables already matched to that data using either matching method—to the regions or countries. We did not apply here the three additional filters described just above for the pixel-level analyses, nor did we apply the filters listed in Section 3.1 to the pixel data prior to collapsing it, since these filters are only needed for the nighttime luminosity data. Specifically, to collapse the data for each ethnographic variable, we computed the population-weighted mean of the variable across the regions' or countries' pixels. For pixels matched to more than one Ethnologue language, we took the simple average of the EA variable across each pixel's languages. For variables defined at the pixel level, such as the geographic controls, we proceeded analogously, except that we took the simple (instead of the population-weighted) mean of each variable across each region's or county's pixels.

Finally, we dropped regions or countries for which the pixels with nonmissing KII data account for less than 75% of the population. With the language-tree matching method, this left 1452 regions in 83 countries for the region-level analyses with the Gennaioli et al. (2014) data; with the direct matching method, this left 873 regions in 61 countries.

Matching to country-ethnicities

For the analyses at the country-ethnicity level, each country-ethnicity was assigned its matched EA variables. The filters described above and in Section 3.1 for the pixel-level analyses were applied, and we took the simple mean of each geographic control across the pixels in each country-homeland and computed nighttime luminosity (i.e., light density) for each country-homeland.

B Additional graphs and tables

B.1 Additional graphs and tables for Section 3 (Data)

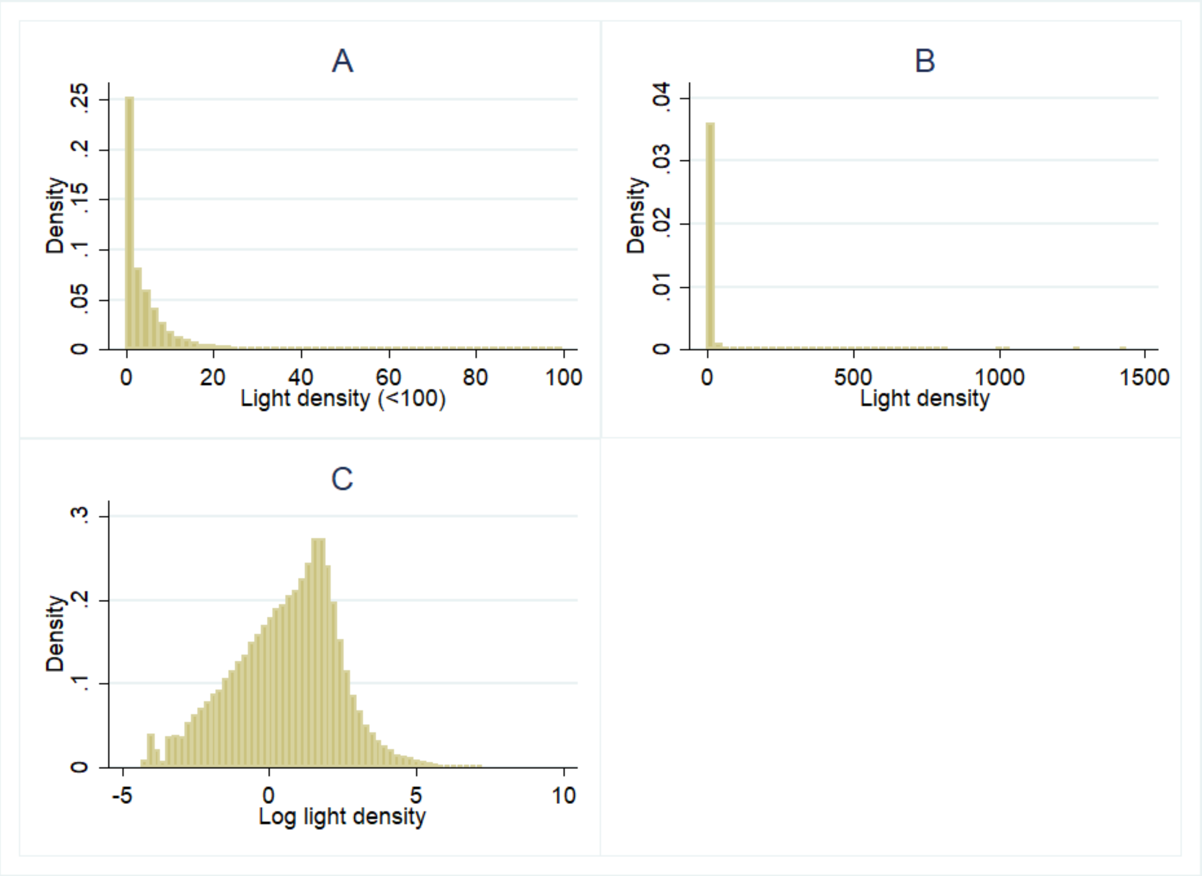


Figure B.1.1: Histograms of light and log nighttime luminosity for the 783,525 pixels in our sample. Panels A shows an histogram for the subset of these pixels with luminosity less than 100; Panel B shows the same histogram but for all these pixels; and Panel C shows an histogram of log luminosity, also for all these pixels.

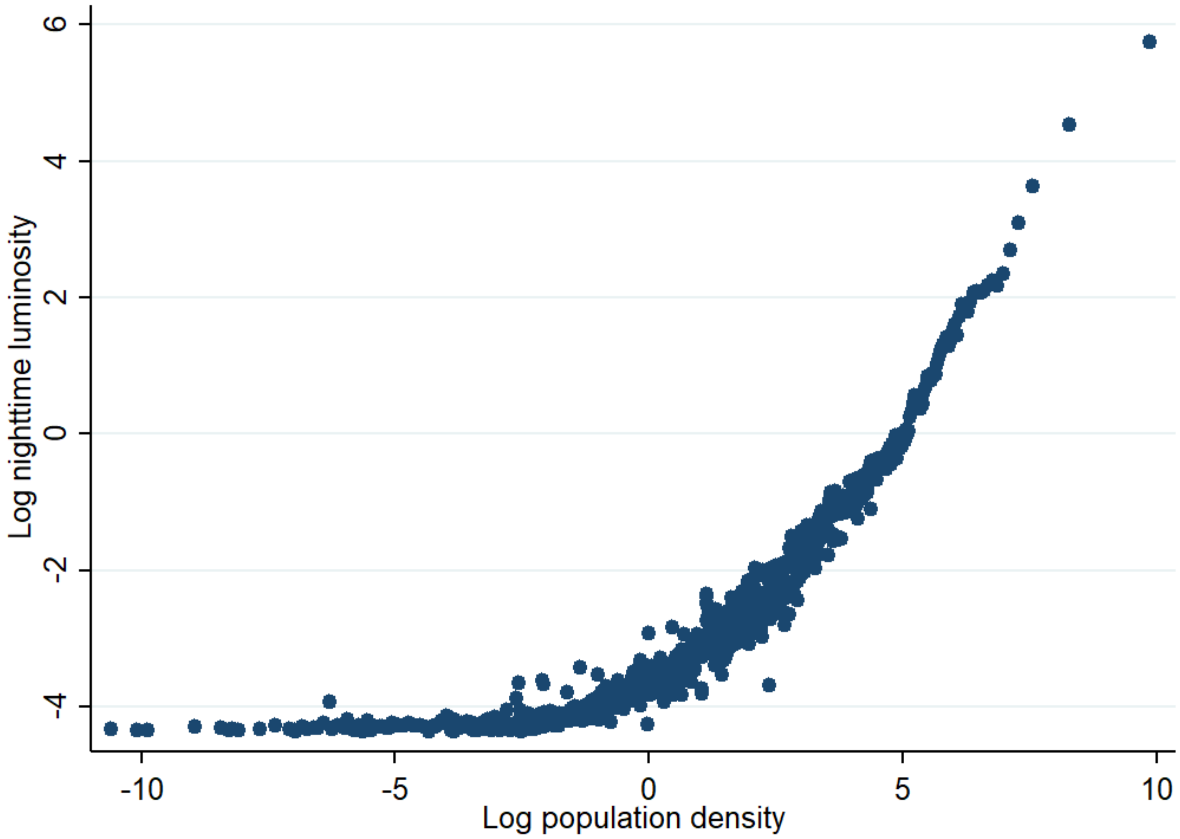


Figure B.1.2: Binned scatterplot (with one bin for every 1,000 observations) of the relationship between log nighttime luminosity and log population density across the 783,525 pixels in our sample. In the analyses reported in the paper, unlike for this figure, we recoded to 1 the population density of pixels whose original population density was less than 1 person per square kilometer.

	KII	Cousin marriage preference	Polygyny	Co-residence of extended family	Lineage organization organization	Community organization
KII	1					
Cousin marriage preference	0.4728^{***}	1				
Polygyny	0.5509^{***}	0.0261	1			
Co-residence of extended family	0.5565^{***}	0.0737^{*}	0.2041^{***}	1		
Lineage organization	0.7203^{***}	0.2135^{***}	0.2779^{***}	0.2408^{***}	1	
Community organization	0.5927^{***}	0.0375	0.1323^{***}	0.1536^{***}	0.4375^{***}	1

Table B.1.1: Pearson correlation between the KII and its five sub-indicators across the 911 EA societies with non-missing KII data. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

B.2 Additional graphs and tables for Section 4.1 (Kinship intensity and nighttime luminosity

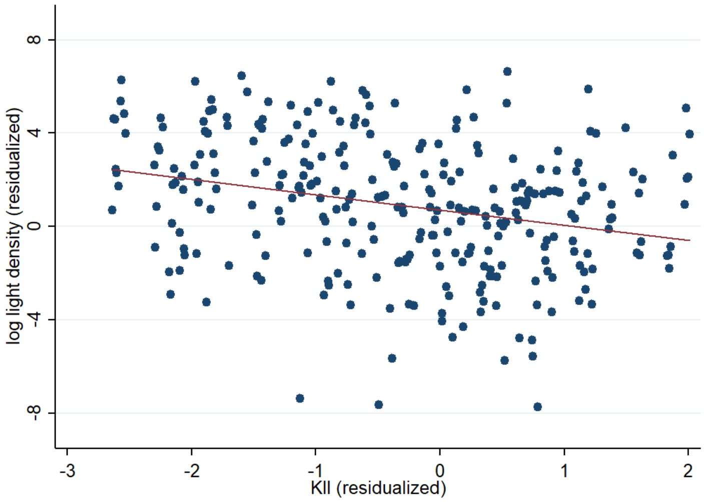


Figure B.2.1: Log nighttime luminosity vs. the KII across ethnicities, both residualized on log population density, for the sample of ethnicities matched based on the language-tree matching method.

Table B.2.1: The KII and nighttime luminosity: summary statistics

Panel A.
488,558 pixels (in the homelands of 2,352 ethnicities)

	mean	sd	min	max
Log nighttime luminosity	-2.418	2.674	-4.360	7.142
Kinship intensity index	-0.819	1.456	-2.510	2.104
Log population density	1.773	1.877	0.000	10.215
Monthly temperature (degree Celsius)	3.256	20.383	-54.400	33.800
Precipitation (cm)	5.479	8.148	0.000	76.650
Agricultural suitability	0.277	0.313	0.000	0.999
Absolute latitude	34.759	18.566	0.063	74.813
Elevation (km)	0.660	0.804	-0.137	6.184
Ruggedness (km)	0.092	0.135	0.000	1.518
Distance to coast (1000 km)	0.576	0.493	0.003	2.464
Distance to nearest river/lake (1000 km)	0.146	0.232	0.000	1.537
Gathering	0.339	0.846	0.000	8.000
Hunting	0.561	0.963	0.000	9.000
Fishing	0.811	0.966	0.000	8.000
Animal husbandry	3.401	2.080	0.000	9.000
Agriculture	4.887	2.339	0.000	9.000
Settlement complexity	5.580	2.354	1.000	8.000
Political hierarchies	3.493	1.273	1.000	5.000
Absence of private property	0.263	0.437	0.000	1.000
Malaria index	1.925	5.096	0.000	38.081

Panel B.
377,656 pixels (in the homelands of 713 ethnicities)

	mean	sd	min	max
Log nighttime luminosity	-2.381	2.672	-4.360	7.142
Kinship intensity index	-1.121	1.389	-2.510	2.104
Log population density	1.659	1.847	0.000	10.215
Monthly temperature (degree Celsius)	1.489	21.231	-54.400	33.800
Precipitation (cm)	5.593	8.348	0.000	76.650
Agricultural suitability	0.298	0.322	0.000	0.999
Absolute latitude	36.167	18.426	0.063	74.813
Elevation (km)	0.611	0.706	-0.089	6.121
Ruggedness (km)	0.085	0.124	0.000	1.317
Distance to coast (1000 km)	0.595	0.497	0.003	2.464
Distance to nearest river/lake (1000 km)	0.122	0.182	0.000	1.376
Gathering	0.273	0.564	0.000	8.000
Hunting	0.463	0.859	0.000	9.000
Fishing	0.816	0.907	0.000	7.500
Animal husbandry	3.486	1.986	0.000	9.000
Agriculture	4.962	2.279	0.000	9.000
Settlement complexity	5.775	2.288	1.000	8.000
Political hierarchies	3.611	1.215	1.000	5.000
Absence of private property	0.278	0.445	0.000	1.000
Malaria index	1.643	4.590	0.000	38.081

Notes: The table shows summary statistics for the main variables used in the analysis of the relationship between the KII and nighttime luminosity, for the pixels in the analysis samples obtained with the language-tree (Panel A) and direct (Panel B) matching methods.

Table B.2.2: The KII and nighttime luminosity with the language-tree matching method: OLS estimates

	Log nighttime luminosity								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
KII	-0.500***	-0.391***	-0.120***	-0.096***	-0.063**	-0.068**	-0.088***	-0.096***	-0.096***
	(0.123)	(0.114)	(0.037)	(0.020)	(0.031)	(0.029)	(0.019)	(0.017)	(0.021)
Log population density	0.930***	0.964***	1.035***	1.012***	1.012***	1.010***	1.008***	1.012***	1.012***
	(0.056)	(0.042)	(0.070)	(0.062)	(0.063)	(0.063)	(0.062)	(0.112)	(0.115)
Subsistence variables					yes				
Political hierarchies						yes			
Malaria index							yes		
Geographic controls		yes	yes	yes	yes	yes	yes	yes	yes
Continent FE			yes						
Country FE				yes	yes	yes	yes	yes	yes
Observations	488,558	488,558	488,558	488,558	488,558	486,936	488,558	488,558	488,558
R-squared	0.462	0.530	0.570	0.653	0.653	0.654	0.654	0.653	0.653
Number of clusters	162	162	162	162	162	162	162	74	96 & 162

Notes: Each observation is a pixel in the country-homeland of a country-ethnicity matched with the language-tree matching method. The geographic controls include temperature, precipitation, agricultural suitability, absolute latitude, elevation, ruggedness, distance to coast, and distance to nearest river or lake. The subsistence variables measure the fraction of an ethnicity's subsistence economy that depends on hunting, fishing, animal husbandry, and agriculture (with gathering as the omitted category). Standard errors in parentheses are clustered at the country level in all regressions, except in column 8, where they are clustered at the language-family level, and in column 9, where they are clustered two-way at both the country and language-family levels. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table B.2.3: The KII and nighttime luminosity at the country-ethnicity level: OLS estimates

	Log nighttime luminosity								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
KII	-0.442***	-0.286***	-0.029	0.006	0.001	0.017	0.010	0.006	0.006
	(0.125)	(0.076)	(0.053)	(0.040)	(0.041)	(0.044)	(0.038)	(0.047)	(0.042)
Log population density	0.712***	0.727***	0.859***	0.907***	0.915***	0.895***	0.903***	0.907***	0.907***
	(0.070)	(0.056)	(0.054)	(0.057)	(0.058)	(0.057)	(0.061)	(0.058)	(0.067)
Subsistence variables					yes				
Political hierarchies						yes			
Malaria index							yes		
Log population density	yes	yes	yes	yes	yes	yes	yes	yes	yes
Geographic controls		yes	yes	yes	yes	yes	yes	yes	yes
Continent FE			yes						
Country FE				yes	yes	yes	yes	yes	yes
Observations	2,143	2,143	2,143	2,143	2,143	2,137	2,143	2,143	2,143
R-squared	0.281	0.538	0.609	0.737	0.738	0.738	0.737	0.737	0.737
Number of clusters	161	161	161	161	161	161	161	72	88&166

	Log nighttime luminosity								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
X	-0.671***	-0.375***	-0.097*	-0.142***	-0.150***	-0.115**	-0.130**	-0.142***	-0.142***
	(0.098)	(0.061)	(0.053)	(0.053)	(0.052)	(0.058)	(0.050)	(0.045)	(0.048)
Log pop density	0.725***	0.785***	0.930***	0.939***	0.959***	0.903***	0.924***	0.939***	0.939***
	(0.088)	(0.068)	(0.054)	(0.081)	(0.086)	(0.082)	(0.087)	(0.076)	(0.087)
Subsistence variables					yes				
Political hierarchies						yes			
Malaria index							yes		
Log population density	yes	yes	yes	yes	yes	yes	yes	yes	yes
Geographic controls		yes	yes	yes	yes	yes	yes	yes	yes
Continent FE			yes						
Country FE				yes	yes	yes	yes	yes	yes
Observations	659	659	659	659	659	655	659	659	659
R-squared	0.348	0.646	0.715	0.831	0.833	0.833	0.832	0.831	0.831
Number of clusters	135	135	135	135	135	135	135	60	88&161

Notes: each observation is a country-ethnicity matched with the language-tree (Panel A) and direct (Panel B) matching methods. The geographic controls include temperature, precipitation, agricultural suitability, absolute latitude, elevation, ruggedness, distance to coast, and distance to nearest river or lake. The subsistence variables include four variables that measure the fraction of an ethnicity's subsistence economy that depends on hunting, fishing, animal husbandry, and agriculture; these four variables, together with a fifth variable for gathering, sum to 1, so we omitted the fifth variable for gathering. Standard errors in parentheses are clustered at the country level in all regressions. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table B.2.4: KII sub-indicators and nighttime luminosity: OLS estimates

Panel A.	Log nighttime luminosity									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Cousin marriage preference	-0.085	0.003								
	(0.084)	(0.044)								
Coresidence of extended families			-0.343***	-0.092***						
			(0.125)	(0.019)						
Polygamy					-0.464***	-0.063*				
					(0.120)	(0.033)				
Community organization							-0.120	-0.122***		
							(0.079)	(0.038)		
Lineage organization									-0.597***	-0.101***
									(0.162)	(0.024)
Log population density	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Geographic controls	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Country FE	no	yes	no	yes	no	yes	no	yes	no	yes
Observations	488,717	488,717	489,114	489,114	489,106	489,106	488,793	488,793	489,114	489,114
R-squared	0.500	0.652	0.526	0.653	0.530	0.653	0.501	0.653	0.541	0.653
Number of clusters	162	162	162	162	162	162	162	162	162	162

Panel B.	Log nighttime luminosity									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Cousin marriage preference	-0.057	0.037								
	(0.094)	(0.066)								
Coresidence of extended families			-0.409***	-0.121***						
			(0.140)	(0.027)						
Polygamy					-0.485***	-0.057				
					(0.135)	(0.044)				
Community organization							-0.138	-0.168***		
							(0.110)	(0.040)		
Lineage organization									-0.655***	-0.133***
									(0.188)	(0.033)
Log population density	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Geographic controls	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Country FE	no	yes	no	yes	no	yes	no	yes	no	yes
Observations	386,160	386,160	390,720	390,720	390,350	390,350	381,151	381,151	390,542	390,542
R-squared	0.496	0.652	0.527	0.652	0.525	0.652	0.502	0.659	0.539	0.652
Number of clusters	140	140	141	141	141	141	138	138	141	141

Notes: Each observation is a pixel in the country-homeland of a country-ethnicity matched with the language-tree (Panel A) and direct (Panel B) matching methods. The geographic controls include temperature, precipitation, agricultural suitability, absolute latitude, elevation, ruggedness, distance to coast, and distance to nearest river or lake. The subsistence variables include four variables that measure the fraction of an ethnicity's subsistence economy that depends on hunting, fishing, animal husbandry, and agriculture; these four variables, together with a fifth variable for gathering, sum to 1, so we omitted the fifth variable for gathering. Standard errors in parentheses are clustered at the country level in all regressions. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

B.3 Additional graphs and tables for Section 4.2 (Kinship intensity and regional GDP per capita)

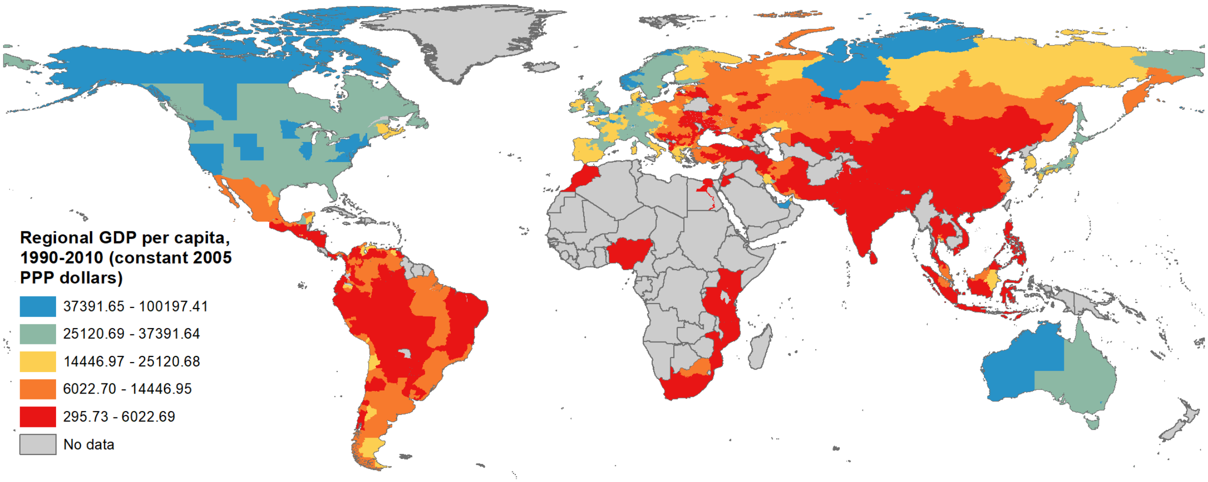


Figure B.3.1: GDP per capita (in constant 2005 PPP dollars) for 1,452 regions around the world. For this map, we take the average GDP per capita for each region for the period from 1990 to 2010 (since data is available for all regions during that period). Our regression analyses use panel data from 1950 to 2010.

Table B.3.1: The KII and regional GDP per capita: summary statistics

Panel A.
9,019 region-years (in 1,452 distinct subnational regions in 83 countries)

	mean	sd	min	max
Log GDP per capita	8.798	1.149	5.242	12.020
Kinship intensity index	-1.335	1.099	-2.510	2.045
Oil and Gas production	0.001	0.006	0.000	0.122
Capital is in Region	0.044	0.206	0.000	1.000
Monthly temperature (degree Celsius)	6.431	13.487	-42.781	30.749
Precipitation (cm)	5.998	6.193	0.000	53.476
Agricultural suitability	0.501	0.279	0.000	0.998
Absolute latitude	33.543	16.362	0.228	70.117
Elevation (km)	0.606	0.680	-0.008	4.801
Ruggedness (km)	0.142	0.129	0.000	0.818
Distance to coast (1000 km)	0.253	0.331	0.002	2.225
Distance to nearest river/lake (1000 km)	0.080	0.097	0.003	0.896

Panel B.
5,514 region-years (in 873 distinct subnational regions in 61 countries)

	mean	sd	min	max
Log GDP per capita	8.876	1.019	5.545	11.340
Kinship intensity index	-1.603	1.046	-2.576	1.714
Oil and Gas production	0.001	0.006	0.000	0.122
Capital is in Region	0.050	0.218	0.000	1.000
Monthly temperature (degree Celsius)	5.806	13.636	-42.781	28.272
Precipitation (cm)	5.875	5.922	0.000	53.475
Agricultural suitability	0.536	0.280	0.001	0.998
Absolute latitude	34.762	15.296	0.583	66.671
Elevation (km)	0.555	0.596	-0.008	3.308
Ruggedness (km)	0.125	0.110	0.000	0.818
Distance to coast (1000 km)	0.258	0.342	0.003	2.225
Distance to nearest river/lake (1000 km)	0.080	0.087	0.005	0.580

Notes: Summary statistics for the main variables used in the analysis of the relationship between the KII and regional GDP per capita, for the region-years in the analysis samples obtained with the language-tree (Panel A) and direct (Panel B) matching methods.

Table B.3.2: The KII and regional GDP per capita with the direct matching method: OLS estimates

	Log regional GDP per capita							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
KII	-0.493***	-0.456***	-0.334***	-0.090**	-0.121***	-0.088**	-0.094**	-0.084**
	(0.099)	(0.074)	(0.068)	(0.039)	(0.042)	(0.041)	(0.039)	(0.036)
Subsistence variables					yes			
Political hierarchies						yes		
Malaria index							yes	
Oil and Gas production								yes
Capital is in Region								yes
Geographic controls		yes	yes	yes	yes	yes	yes	yes
Year FE	yes	yes						
Year × Continent FE			yes					
Year × Country FE				yes	yes	yes	yes	yes
Observations	9,019	9,019	9,019	9,019	9,019	9,019	9,019	9,014
R-squared	0.329	0.597	0.673	0.896	0.897	0.896	0.898	0.906
Number of clusters	83	83	83	83	83	83	83	83

Notes: Each observation is a region-year in the analysis sample obtained with the language-tree matching method. The geographic controls include temperature, precipitation, agricultural suitability, absolute latitude, elevation, ruggedness, distance to coast, and distance to nearest river or lake. The subsistence variables measure the weighted fraction of a region's ethnicities' subsistence economies that depend on hunting, fishing, animal husbandry, and agriculture (with gathering as the omitted category). Standard errors in parentheses are clustered at the country level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

B.4 Additional graphs and tables for Section 4.3 (Spatial regression discontinuity (RD) analysis)

Table B.4.1: Spatial RD analysis: summary statistics

750,996 pixels (in the homelands of 2716 pairs of neighboring ethnicities)				
	mean	sd	min	max
Log nighttime luminosity	-2.758	2.441	-4.360	7.142
KII	-1.063	1.420	-2.576	2.087
Log population density	2.042	1.929	0.000	10.215
Agricultural suitability	0.278	0.298	0.000	0.999
Absolute latitude	28.306	19.251	0.063	73.938
Monthly temperature (degree Celsius)	7.583	19.852	-52.900	33.800
Precipitation (cm)	7.051	9.688	0.000	76.650
Elevation (km)	0.779	0.828	-0.137	6.184
Ruggedness (km)	0.103	0.147	0.000	1.518
Distance to coast (1000 km)	0.637	0.506	0.003	2.464
Distance to nearest river/lake (1000 km)	0.104	0.167	0.000	1.499
Malaria index	3.540	6.729	0.000	38.081

290,669 pixels (in the homelands of 572 pairs of neighboring ethnicities)				
	mean	sd	min	max
Log nighttime luminosity	-2.694	2.483	-4.360	7.142
KII	-1.014	1.414	-2.576	2.087
Distance to border	102.921	56.021	5.000	200.000
dist_border_HighKII	29.885	54.222	0.000	200.000
Log population density	1.642	1.834	0.000	10.215
Agricultural suitability	0.273	0.303	0.000	0.999
Absolute latitude	34.401	20.017	0.063	73.938
Monthly temperature (degree Celsius)	0.880	21.652	-52.900	32.800
Precipitation (cm)	6.362	9.321	0.000	76.650
Elevation (km)	0.732	0.707	-0.089	5.543
Ruggedness (km)	0.093	0.125	0.000	1.317
Distance to coast (1000 km)	0.677	0.513	0.003	2.464
Distance to nearest river/lake (1000 km)	0.082	0.107	0.000	1.090
Malaria Index	2.592	5.808	0.000	36.849

Notes: Summary statistics for the main variables used in the spatial RD analysis, for the pixels in the baseline analysis sample.

Table B.4.2: The KII and nighttime luminosity: spatial RD analysis with the language-tree matching method

	Log nighttime luminosity						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
KII	-0.045**	-0.054***	-0.047**	-0.036	-0.054***	-0.054***	-0.054***
	(0.018)	(0.018)	(0.020)	(0.022)	(0.017)	(0.018)	(0.020)
Subsistence variables			yes				
Political hierarchies				yes			
Malaria index					yes		
Log population density	yes	yes	yes	yes	yes	yes	yes
Geographic controls		yes	yes	yes	yes	yes	yes
Distance-to-the-boundary polynomial	yes	yes	yes	yes	yes	yes	yes
Ethnicity pair FE	yes	yes	yes	yes	yes	yes	yes
Observations	750,996	750,996	750,996	748,603	750,996	750,996	750,996
R-squared	0.614	0.623	0.623	0.624	0.623	0.623	0.623
Number of clusters	104	104	104	104	104	69	69 & 104

Notes: Each observation is a pixel that belongs to an ethnicity matched with the language-tree method, in a pair of contiguous ethnicities in the same country, and within 200 km of the boundary between the two ethnicities. The geographic controls include temperature, precipitation, agricultural suitability, absolute latitude, elevation, ruggedness, distance to coast, and distance to nearest river or lake. "Distance-to-the-boundary polynomial" is the geodesic distance of each pixel's centroid from the boundary between the two adjacent ethnicities; we allow the coefficient on the distance term to differ between the relatively high and the relatively low KII ethnicities in the pairs. The subsistence variables measure the fraction of an ethnicity's subsistence economy that depends on hunting, fishing, animal husbandry, and agriculture (with gathering as the omitted category). Standard errors in parentheses are clustered at the country level (in columns 1–5), language family (in column 6), and two-way clustered at both the country and language-family levels (in column 7). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table B.4.3: Spatial RD analysis for the geographic controls

	(1) Temperature	(2) Precipitation	(3) Agricultural suitability	(4) Elevation	(5) Ruggedness	(6) Distance to coast	(7) Distance to river/lake	(8) Malaria index
KII	-0.377	0.064	0.005	-0.006	-0.005	-0.020*	0.003	-0.061
	(0.297)	(0.093)	(0.007)	(0.026)	(0.005)	(0.011)	(0.003)	(0.095)
Distance-to-the-boundary polynomial	yes	yes	yes	yes	yes	yes	yes	yes
Ethnicity pair FE	yes	yes	yes	yes	yes	yes	yes	yes
Observations	290,669	290,669	290,669	290,669	290,669	290,669	290,669	290,669
R-squared	0.981	0.907	0.638	0.664	0.370	0.809	0.714	0.883
Number of clusters	70	70	70	70	70	70	70	70

Notes: Each observation is a pixel that belongs to an ethnicity in a pair of contiguous ethnicities in the same country, and within 200 km of the boundary between the two ethnicities. "Distance-to-the-boundary polynomial" is the geodesic distance of each pixel's centroid from the boundary between the two adjacent ethnicities; we allow the coefficient on the distance term to differ between the relatively high and the relatively low KII ethnicities in the pairs. Standard errors in parentheses are clustered at the country level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

B.5 Additional graphs and tables for Section 5.2 (The inbreeding coefficient and kinship intensity)

Table B.5.1: The inbreeding coefficient and kinship intensity: summary statistics

	No. ethnicities	mean	sd	min	max
HO sample size	416	29.0	42.7	8	203
KII	396	-0.242	0.782	-1.511	1.210
Cousin marriage preference	398	0.101	1.148	-0.546	2.808
Co-residence	410	-0.252	1.278	-2.542	1.408
Polygamy	410	-0.637	1.066	-1.801	0.937
Community organization	407	-0.053	0.968	-1.072	0.931
Lineage organization	410	-0.278	1.087	-1.595	0.626
$\overline{F_{ROH}}$	416	0.020	0.018	0.003	0.146
Expected heterozygosity	415	17,410	1,084	12,156	18,630
Migratory distance (months)	416	2.798	1.805	0.229	8.412
Mean regional pairwise F_{ST}	415	0.024	0.017	0.001	0.117

mean_f_roh 0.012 0.009 0.003 0.146

Notes: Summary statistics for the main variables used in the analysis of the relationship between kinship intensity and the inbreeding coefficient, for the Ethnologue country-ethnicities in the analysis sample. 'HO sample size' denotes the number of individuals in the HO dataset used to compute $\overline{F_{ROH}}$, expected heterozygosity, mean regional pairwise F_{ST}, and the top 20 PCs; summary statistics are not shown for the top 20 PCs because they have been normalized and thus have zero mean and unit variance.

Table B.5.2: The inbreeding coefficient and kinship intensity: robustness checks

	Cousin marriage preference				KII			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$\overline{F_{ROH}}$	43.636***	38.142***	52.054***	43.308***	10.783***	9.411***	15.391***	13.591***
	(6.228)	(4.828)	(5.307)	(6.101)	(2.923)	(2.474)	(2.712)	(3.505)
Genetic controls	yes	yes	yes	yes	yes	yes	yes	yes
Geographic controls	yes				yes			
Min. HO sample per ethnicity	8	5	10	15	8	5	10	15
Observations	364	488	324	210	362	486	322	208
R-squared	0.496	0.414	0.495	0.621	0.751	0.641	0.753	0.799
$\Delta R^2(\overline{F_{ROH}})$	0.0765	0.0922	0.113	0.0706	0.0103	0.0137	0.0210	0.0159
Number of clusters	117	137	123	92	117	137	123	92

Notes: Each observation is a country-ethnicity from the Ethnologue. The genetic controls include expected heterozygosity, migratory distance from East Africa, the top 20 PCs, and mean regional pairwise F_{ST}. The geographic controls include temperature, precipitation, agricultural suitability, absolute latitude, elevation, ruggedness, distance to coast, and distance to nearest river or lake. 'Min. HO sample per ethnicity' denotes the minimum number of matched HO individuals for an Ethnologue language (and its associated country-ethnicities) to be included in the regression (the baseline used for all the specifications in Table 6 is 8). $\Delta R^2(\overline{F_{ROH}})$ is the incremental R^2 of $\overline{F_{ROH}}$, defined as the difference in R^2 between the regression on $\overline{F_{ROH}}$ and the controls and the regression on the controls only. Standard errors, clustered at country-level, are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table B.5.3: The inbreeding coefficient and kinship intensity: other KII sub-indicators

	Co-residence		Polygamy		Community organization		Lineage organization	
VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$\overline{F_{ROH}}$	-14.666***	-13.799*	9.753*	15.722**	11.025**	16.013***	6.650*	9.818*
	(5.012)	(7.326)	(5.326)	(7.474)	(4.489)	(6.012)	(3.920)	(5.394)
Genetic controls	yes	yes	yes	yes	yes	yes	yes	yes
Country FE		yes		yes		yes		yes
Observations	409	409	409	409	406	406	409	409
R-squared	0.498	0.717	0.608	0.704	0.244	0.511	0.706	0.841
Adjusted R-squared	0.464	0.549	0.581	0.528	0.192	0.217	0.686	0.747
Number of clusters	127	127	127	127	127	127	127	127
$\Delta R^2(\overline{F_{ROH}})$	0.008	0.005	0.005	0.009	0.008	0.011	0.002	0.003

Notes: Each observation is a country-ethnicity from the Ethnologue. The genetic controls include expected heterozygosity, migratory distance from East Africa, the top 20 PCs, and mean regional pairwise F_{ST}. $\Delta R^2(\overline{F_{ROH}})$ is the incremental R^2 of $\overline{F_{ROH}}$, defined as the difference in R^2 between the regression on $\overline{F_{ROH}}$ and the controls and the regression on the controls only. Standard errors, clustered at the country-level, are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

B.6 Additional graphs and tables for Section 5.3 (The inbreeding coefficient and nighttime luminosity)

Table B.6.1: The inbreeding coefficient and nighttime luminosity: summary statistics

281,177 pixels (in the homelands of 245 ethnicities)				
	mean	sd	min	max
Log nighttime luminosity	-1.919	2.833	-4.360	7.142
$\overline{F_{ROH}}$	0.012	0.009	0.003	0.146
fst_pair_mean	0.017	0.016	0.001	0.095
exp_het	17,798.633	468.939	12,155.658	18,630.277
Log population density	1.758	1.913	0.000	10.152
Migratory distance (months)	3.441	1.917	0.229	8.412
Monthly temperature (degree Celsius)	-4.748	20.080	-54.400	33.800
Precipitation (cm)	3.660	5.115	0.000	61.930
Agricultural suitability	0.327	0.334	0.000	0.999
Absolute latitude	43.322	16.042	0.063	74.813
Elevation (km)	0.569	0.660	-0.137	5.547
Ruggedness (km)	0.095	0.134	0.000	1.459
Distance to coast (1000 km)	0.574	0.523	0.003	2.464
Distance to nearest river/lake (1000 km)	0.096	0.142	0.000	1.353

Notes: Summary statistics for the main variables used in the analysis of the relationship between the inbreeding coefficient and nighttime luminosity, for the pixels in the analysis sample.

Table B.6.2: The inbreeding coefficient and nighttime luminosity: robustness checks

	Log nighttime luminosity			
	(1)	(2)	(3)	(4)
$\overline{F_{ROH}}$	-15.186***	-19.712***	-15.246	-15.758
	(5.780)	(6.891)	(12.315)	(12.347)
Log population density	yes	yes	yes	yes
Genetic controls		yes	yes	yes
Geographic controls	yes	yes	yes	yes
Country FE	yes	yes	yes	yes
Min. HO sample per ethnicity	8	5	10	15
Observations	281,177	305,467	273,018	178,616
R-squared	0.671	0.668	0.675	0.614
Number of clusters	98	111	92	61

Notes: Each observation is a pixel. The genetic controls include expected heterozygosity, migratory distance from East Africa, the top 20 PCs, and mean regional pairwise F_{ST}. The geographic controls include temperature, precipitation, agricultural suitability, absolute latitude, elevation, ruggedness, distance to coast, and distance to nearest river or lake. 'Min. HO sample per ethnicity' denotes the minimum number of matched HO individuals for an Ethnologue ethnicity to be included in the regression (the baseline used for all the specifications in Table 7 is 8). Standard errors in parentheses are clustered at the country level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

B.7 Additional graphs and tables for Section 6 (Causal pathways)

Table B.7.1: Robustness to excluding European-ancestry countries and to controlling for deep Christianization

	Language-tree matches				Direct matches			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Panel A. Regressions of log nighttime luminosity at the pixel level							
KII	-0.141*	-0.016	0.002	-0.059	-0.097	0.035	0.062	-0.062
	(0.085)	(0.065)	(0.088)	(0.052)	(0.101)	(0.079)	(0.137)	(0.065)
Log population density	yes	yes	yes	yes	yes	yes	yes	yes
Observations	201,391	201,391	488,558	488,558	129,168	129,168	377,656	377,656
R-Squared	0.515	0.601	0.546	0.653	0.558	0.642	0.551	0.660
Number of clusters	96	96	162	162	85	85	138	138
	Panel B. Regressions of log nighttime luminosity at the country-ethnicity level							
KII	-0.124*	0.023	-0.271***	0.031	-0.209**	-0.175***	-0.392***	-0.098
	(0.074)	(0.050)	(0.081)	(0.043)	(0.082)	(0.066)	(0.082)	(0.067)
Log population density	yes	yes	yes	yes	yes	yes	yes	yes
Observations	1,711	1,711	2,143	2,143	482	482	659	659
R-Squared	0.543	0.695	0.538	0.737	0.650	0.802	0.646	0.831
Number of clusters	95	95	161	161	85	85	135	135
	Panel C. Regressions of log regional GDP per capita at the region level							
KII	-0.360***	-0.103	-0.375***	-0.111**	-0.358***	-0.113	-0.367***	-0.115**
	(0.078)	(0.074)	(0.073)	(0.046)	(0.077)	(0.074)	(0.072)	(0.045)
Year FEs	yes		yes		yes		yes	
Observations	3,182	3,182	9,019	9,019	3,182	3,182	9,019	9,019
R-Squared	0.645	0.856	0.608	0.896	0.646	0.856	0.607	0.896
Number of clusters	30	30	83	83	30	30	83	83
	Panel D. Regressions of log nighttime luminosity (spatial RD analysis)							
KII	-0.083*		-0.068		-0.113*		-0.120	
	(0.043)		(0.048)		(0.060)		(0.091)	
Log population density	yes		yes		yes		yes	
Distance-to-the-boundary polynomial	yes		yes		yes		yes	
Ethnicity pair FE	yes		yes		yes		yes	
Observations	378,764		750,996		115,660		290,669	
R-Squared	0.634		0.623		0.659		0.613	
Number of clusters	71		104		48		70	
	Panel E. Regressions of log nighttime luminosity at the pixel level with $\overline{F_{ROH}}$							
$\overline{F_{ROH}}$					-29.364**	-35.593***	-26.052**	-20.328***
					(11.368)	(11.866)	(11.094)	(7.475)
Log population density					yes	yes	yes	yes
Genetic controls					yes	yes	yes	yes
Observations					63,142	63,142	281,177	281,177
R-Squared					0.621	0.641	0.652	0.674
Number of clusters					51	51	98	98
Subsamble, no Europeans	yes	yes			yes	yes		
Deep Christianization			yes	yes			yes	yes
Geographic controls	yes	yes	yes	yes	yes	yes	yes	yes
Country FE (Year × Country FE for Panel C)		yes		yes		yes		yes

Notes: Each observation is a pixel (in Panels A, D, and E), a subnational region (in Panel B), or a country-ethnicity (in Panel B). Columns 1-4 and 5-8 show results in the samples obtained using the language-tree and direct matching methods, respectively. Columns 1-2 and 5-6 show results in the subsample that excludes Europe, the Americas, Australia, and New Zealand, for regressions without (columns 1 and 5) and with (columns 2 and 6) country fixed effects. Columns 3-4 and 7-8 show results of regressions (in the full sample) that include controls for deep Christianization without (columns 3 and 7) and with (columns 4 and 8) country fixed effects. The geographic controls include temperature, precipitation, agricultural suitability, absolute latitude, elevation, ruggedness, distance to coast, and distance to nearest river or lake. Standard errors in parentheses are clustered at the country level in all regressions. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table B.7.2: The KII and nighttime luminosity in subsamples defined by population density: OLS estimates

	(1) Pop. Density =1	(2) Pop. Density =1	(3) Pop. Density >1 & <=10	(4) Pop. Density >1 & <=10	(5) Pop. Density >10	(6) Pop. Density >10
KII	-0.108***	-0.069**	-0.273**	-0.067***	-0.641***	-0.074
	(0.036)	(0.032)	(0.121)	(0.018)	(0.144)	(0.054)
Log population density			0.722***	0.782***	1.052***	1.056***
			(0.095)	(0.115)	(0.059)	(0.063)
Geographic controls	yes	yes	yes	yes	yes	yes
Country FEs		yes		yes		yes
Observations	165,715	165,715	147,425	147,425	175,418	175,418
R-squared	0.052	0.118	0.206	0.410	0.543	0.701
Number of clusters	72	72	135	135	159	159

Notes: Each observation is a pixel in the homeland of a country-ethnicity matched with the language-tree matching method. Regressions were estimated in the subsamples of pixels with population density equal to 1 (columns 1 and 2; recall we recoded population density to 1 for pixels whose original population density was less than 1), between 1 and 10 (columns 3 and 4), and higher than 10 (columns 5 and 6). The geographic controls include temperature, precipitation, agricultural suitability, absolute latitude, elevation, ruggedness, distance to coast, and distance to nearest river or lake. Standard errors in parentheses are clustered at the country level in all regressions. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

C Computing the inbreeding coefficient and other variables in the Human Origins (HO) dataset

To obtain a contemporaneous, on-the-ground measure of kinship intensity, we analyzed genetic data from the Human Origins (HO) dataset from the David Reich Lab. HO contains genetic data on thousands of present-day and ancient individuals from populations around the world. We used the genetic data to compute the average inbreeding coefficient—which we use in our analyses as a proxy for cousin marriage rates—and other control variables for each population.

This appendix describes HO (C.1) and the matching procedure we employed to match the HO individuals and populations to Ethnologue ethnicities (C.2); provides some background on key genetics concepts (C.3); summarizes the quality control (QC) filters that were applied to the HO data before the data was analyzed (C.4); provides an overview of the variables we computed using the HO data (C.5); describes the methods we employed to compute the inbreeding coefficient (C.6) and other control variables (C.7, C.8, C.9) using the HO individuals genetic data; and reports summary statistics for the variables we computed (C.10).

C.1 The Human Origins dataset

We obtained a non-public version of the Human Origins (HO) dataset (Reich Lab, 2020) from the David Reich Lab on February 17, 2020. The HO version we obtained contains genetic data on 13,183 individuals, including 9,460 present-day individuals and 3,723 ancient individuals who lived between $\sim 88,000$ BCE and $\sim$ 1900 CE. It is a compiled dataset that includes individuals from many published studies and from many populations around the world (including Jeong et al. (2019); Lazaridis et al. (2016), Lazaridis et al. (2014); Lipson et al. (2018); Nakatsuka et al. (2017); Pickrell et al. (2012)).

We only include a subset of 4,756 present-day (i.e., non-ancient) individuals who have been genotyped on the Affymetrix Human Origin array in our analyses. These individuals were genotyped at a set of 597,573 genetic variants called single nucleotide polymorphisms (SNPs).

C.2 Matching the HO's populations to Ethnologue's ethnolinguistic groups (ethnicities)

As we further describe below, the subset of the HO we used in our analyses contains 4,756 individuals in 435 populations.[2] For each individual, HO has data on the individual's population, country, latitude, longitude, and in some cases on the individual's city or region. HO also lists the original publication from which each individual's data comes from, and some original publications include the language ISO code of the individual's ethnicity. We used this information, as well as various linguistic resources, Wikipedia entries, and an Ethnologue shapefile, to match each HO population to one of the Ethnologue's 6427 languages (which constitute 7,651 ethnicity-country combinations (what we call ethnicities).

As a first step, we conducted a many-to-one match to match each of the 435 HO populations to a single Ethnologue ethnicity in the HO population's country, and then dropped Ethnologue ethnicities that contain less than five individuals. This step resulted in 313 distinct Ethnologue

---

[2] We use the term "HO population" to describe the group identifier included for each individual in the HO dataset under the heading "Group Label", but the terms "ethno-linguistic group" or "ethnicity" could also apply in many cases.

ethnicities in our data. As a second step, we conducted a many-to-one match to match each of the 313 Ethnologue ethnicities to other ethnicities that speak the same Ethnologue language but are located in surrounding countries. This second step resulted in 551 distinct Ethnologue ethnicities in our data. The combination of steps 1 and 2 amounts to a many-to-many match from the 435 HO populations to 551 Ethnologue ethnicities.

Some of the variables we computed (the inbreeding coefficient and the top principal components (PCs)) were computed at the individual level and then averaged across the individuals in each Ethnologue ethnicity. Other variables (mean regional pairwise genetic distance and expected heterozygosity) were first computed for each HO population, following which each individual was assigned their HO population's value for each variable, which was then averaged across the individuals in each Ethnologue ethnicity.

Though we matched the HO data to 551 Ethnologue ethnicities, we note that the analyses with the inbreeding coefficient reported in the main text (in Section 5) only used subsets of these 545 ethnicities. The analysis of the relationship between the inbreeding coefficient and cousin marriage preferences (Table 6) only used the 416 Ethnologue ethnicities with eight or more HO individuals; our analysis of the relationship between the inbreeding coefficient and nighttime luminosity, only used 245 of these 416 ethnicities that could be matched to a pixel that was not dropped from the data because of the filters we applied to the pixels (and which we describe in the main text).

C.3 Genetics background

We provide here a brief overview of some of the core genetic concepts needed to follow the remainder of this appendix. Other specific concepts are introduced below as needed. For a more detailed overview, see Beauchamp et al. (2011).

Humans have 23 pairs of chromosomes: they inherit 23 copies of each chromosomes from their mother and another 23 from their father. The elementary building blocks of DNA are nucleotide pairs associated with the four bases A, C, T, and G.[3] The human genome consists of approximately three billion pairs of nucleotide pairs and the associated bases on the 23 pairs of chromosomes.

Any two humans share $\sim$ 99.9% of their genomes. At the remaining $\sim$ 0.1%, different genetic variants exist in the populations. Single nucleotide polymorphism (SNPs) are the most prevalent type of genetic variants and our analyses only rely on SNP data. There are typically exactly two possible variants – also called alleles – at a given SNP's location in the genome[4]. A SNP's minor allele frequency (MAF) is the fraction the variants in the population that are the rarer variant at that SNP's location.

Autosomes are chromosomes other than the sex (X or Y) chromosomes, and our analyses only use SNPs on the 22 pairs of autosomes. Non-autosomal variants are variants on the sex (X or Y) chromosomes and are not used in our analyses.

---

[3] To be precise, a nucleotide pair is associated with a pair of bases rather than with a single base. Here, we ignore these subtleties and will refer to a single base associated with a nucleotide pair. Due to the principle of complementarity, the base A is always paired with the base T, the base C is always paired with the base G, and vice-versa; for that reason, one can focus on only one of the two bases associated with a nucleotide pair.

[4] By definition, there must be at least two possible variants at a SNP's location (else this would not be a SNP). SNP with more than two variants do exist but are rare.

C.4 Description of baseline quality control (QC) filters applied

As is customary when working with genetic data, we applied a number of quality control (QC) filters to the HO data to ensure that the data we analyzed was of sufficiently high quality.[5]. To do so, we used PLINK v2.00a3 (Chang et al., 2015), a software tool that is commonly used to work with genetic data.[6] Our QC procedure is similar to the one employed by Ochoa and Storey (2019).

We applied the following QC filters, in this order:

- Step 1: removed all variants that are not single nucleotide polymorphisms (SNPs), all non-autosomal variants, and all monomorphic SNPs.[7]
- Step 2: removed all ancient individuals and individuals who were not genotyped on the Affymetrix Human Origins array, all individuals flagged by the David Reich Lab as being outliers, all individuals who could not be mapped to an ethnicity in the Ethnologue (including all individuals in admixed populations, since the Ethnologue does not contain admixed populations), and all individuals who were not mapped to at least one ethnicity in the Ethnologue to which at least four other HO individuals were also mapped (i.e., we only kept Ethnologue ethnicities with at least five matched Human Origins individuals). This left 4,771 individuals.
- Step 3: removed variants with missing call rates greater than 0.03 (108,180 SNPs removed).[8]
- Step 4: removed individuals with missing genotype rates greater than 0.03 (15 individuals removed).[9]

Before any QC, there were 13,776 ancient and present-day individuals and 597,573 variants in the HO data. After the baseline QC filters above were applied, the data contained information on 484,944 SNPs and 4,756 present-day individuals (1008 females, 3,679 males, 69 unknown) in 435 HO populations.[10]

(As mentioned above, these 435 HO populations were matched to 551 Ethnologue ethnicities, but for the analyses with the inbreeding coefficient reported in the main text, we only used subsets of these 545 ethnicities.)

---

[5] For a good overview of the different types of QC filters typically applied when working with genetic data, see Price et al. (2006)

[6] We also used PLINK v1.90b6.17 for some calculations because some PLINK commands had not been incorporated yet into PLINK v2.00a3.

[7] Monomorphic SNPs are SNPs that show no variation across all individuals; the variables corresponding to these SNPs are thus constants and can't be used in our analyses.

[8] The missing call rate for a given SNP is equal to N_{miss}/N, where N_{miss} is the number of individuals missing data for that SNP and N is the total number of individuals in the dataset.

[9] The missing genotype rate for an individual is equal to M_{miss}/N, where M_{miss} is the number of SNPs that are missing calls for that individual, and M is the total number of SNPs in the dataset.

[10] We do not apply minor allele frequency (MAF) filters in our baseline QC (other than removing monomorphic SNPs). The reason why is that, for some calculations, MAF filters are applied after calculating MAF at the language level, but for other calculations, MAF filters are applied after calculating MAF using all individuals in the HO data. These MAF filters are described below. We also do not apply a Heidy-Weinberg Equilibrium (HWE) filter because this needs to be applied separately for each population, and we choose not to do this due to small population sizes.

C.5 Variables computed with the HO genetic data: Overview

Our goal is to infer consanguineous marriage practices from inbreeding coefficients, and therefore factors that affect the inbreeding coefficient but are unrelated to parental relatedness are possible confounders in our analyses. In addition to computing the inbreeding coefficient, we therefore compute several quantities that are correlated with it but unrelated to parental relatedness, and we include these quantities in our main regressions as controls. Thus, for each population, we use the HO genetic data to compute the following quantities:

- *Inbreeding coefficient*: Our main variable of interest is each ethnicity's average inbreeding coefficient, denoted F, which is the probability that two alleles at any given location in an individual's genome are identical by descent (IBD) – i.e., that they were inherited from the same ancestor. We compute F at the individual level and then take the ethnicity average. F is a measure of the relatedness of ones' parents. Consider, for instance, the offspring of two first cousins, who share a common grandfather and a common grandmother. The probability that two alleles at a location in the offspring's genome comes from a common grandparent and thus are IBD can be calculated to be 0.0625; this is the offspring's inbreeding coefficient. By contrast, the inbreeding coefficient of the offspring of completely unrelated individuals is 0. We compute two measures of the inbreeding coefficient. The baseline measure we use in the empirical analyses in the paper is F_{ROH}, defined as the fraction of an individual's genome that is located in runs of homozygosity (ROH) that are longer than 1.5 Mb.[11] ROH are segments of the genome that are identical across the maternally and paternally inherited chromosomes and thus could be IBD, and ROH that are longer than 1.5 Mb in length are more likely to arise from inbreeding (Ceballos et al., 2018). We also compute an alternative measure of the inbreeding coefficient, F_{SNP}, which we use in the Appendix to verify that our F_{ROH} estimates are reasonable.
- *Mean regional pairwise F_{ST}*: Following Pemberton and Rosenberg (2014), we estimate the degree of genetic isolation of a population using the average pairwise F_{ST} (genetic distance) between that population and all other populations in the same world geographic region with a post-baseline-QC sample size of at least 8 individuals.
- *Expected heterozygosity*: Controlling for a population's expected heterozygosity is especially important in our setting, since expected heterozygosity correlates negatively with population mean F (Pemberton and Rosenberg, 2014) and has been hypothesized to be related to economic development (Ashraf and Galor, 2013). Expected heterozygosity decreases with migratory distance from East Africa, consistent with a migration model that predicts increased genetic drift and decreased heterozygosity with increased distance from Africa (Ramachandran et al., 2005). In our empirical analyses with our F_{ROH} estimates, we also control for migratory distance from East Africa.
- *The top 20 PCs of the genotype*: We compute the top 20 principal components (PCs) of the genotypic data[12] for each individual, and take the average value of each PC in each population. It has been shown that the top PCs are good proxies for the individuals' geographic origins and that including these as controls in regressions can help mitigate bias related to population stratification (Novembre et al., 2008; Price et al., 2006).

Below, we discuss these quantities in more detail and describe how we computed them.

---

[11] A megabase (Mb) is a unit of physical distance in the genome; the entire genome is $\sim 3,100$ Mb in length.

[12] Specifically, we compute the 20 PCs of the $N \times N$ matrix with the entry in row i and column j equal to the correlation between individuals i and j's SNP data.

C.6 The inbreeding coefficient: background and estimation

Inbreeding occurs when related individuals mate and produce offspring. Related individuals share DNA from common ancestors, and thus inbreeding increases the probability that at a given location on a chromosome, their offspring inherits identical DNA segments from both parents (Ceballos et al., 2018). For instance, first cousins share both a grandmother and a grandfather, and so some DNA segments which they inherited from the common grandmother or grandfather will be identical across their two genomes. As a result, if two first cousins mate and produce a child, then that child will have locations in their genome where they will have inherited identical DNA segments from their two cousin parents, coming originally from the same grandmother or grandfather. The two alleles at a such genomic location are said to be homozygous – since the two alleles at the location are identical – and identical-by-descent (IBD; or autozygous) – since they are identical and can be traced back to one common ancestor.

A common measure of one's level of inbreeding is the coefficient of inbreeding, denoted F. F measures the probability that two alleles at a location in the genome are IBD, and is also equal to the fraction of the genome that is IBD for a given individual. The expected value of an individual's F coefficient is equal to the coefficient of kinship, or one-half the coefficient of relationship, between his two parents. The latter is a measure of the relatedness between two individuals (the two parents, here). For example, the expected F coefficient of the offspring of two second-degree cousins is 0.015625, whereas it is 0.0625 for the offspring of first-degree cousins and 0.25 for the offspring of two siblings. Thus, F measures the degree to which an individuals' two parents were related.

Various methods have been developed to estimate F using an individual's molecular genetic data. Recent advances in genetic technology have led to the development of methods to identify continuous segments of homozygous alleles, known as runs of homozygosity (ROH), in individual genomes. Longer ROH are typically observed as a result of inbreeding (Ceballos et al., 2018), and have been shown to correlate well with pedigree-based estimates of inbreeding (Kang et al., 2017; McQuillan et al., 2008; Pemberton and Rosenberg, 2014) and, as Sahoo et al. (2021) found and as we report in Section 6 of the main text, with population-level cousin marriage preferences.

However, ROH can arise in individuals for a variety of reasons unrelated to marital practices and kinship systems. For example, ROH and inbreeding coefficients are inflated in genetically isolated populations, since low effective population sizes can increase homozygosity (Pemberton and Rosenberg, 2014). In addition, inbreeding coefficients are highly correlated with expected heterozygosity and migratory distance from Africa (Pemberton and Rosenberg, 2014). In the main empirical analyses in the main text, we control for these and other variables.

C.6.1 Estimating F_{ROH}

Our main measure of inbreeding, denoted F_{ROH}, is the fraction of an individual's genome in runs of homozygosity (ROH) of at least 1.5 Mb in length (Clark et al., 2019). This is the baseline measure used by the ROHgen consortium, a large consortium of research groups administering cohorts with genetic data that seeks to estimate the effect of inbreeding on various traits (Clark et al., 2019; Joshi, 2015). We estimate F_{ROH} for all individuals in the (post-baseline-QC) dataset using the ROHgen2 software pipeline developed by the ROHgen consortium (Clark et al., 2019).

Shorter ROH are more likely to arise from population structure or lack of genetic diversity whereas longer ROH (e.g., ROH 1.5 Mb in length or longer) tend to arise from inbreeding (Ceballos et al., 2018; Kang et al., 2017; Pemberton et al., 2012; Sahoo et al., 2021). Using ROH that are at least 1.5 Mb long to compute F_{ROH} mitigates, but does not eliminate, concerns that an individual's homozygous alleles have arisen due to factors other than parental relatedness, such as low effective population size or decreased genetic diversity.

Formally, suppose an individual has n ROH greater than 1.5 Mb across their genome. For each of these ROH $i \in \{1, \ldots, n\}$, define l_i as the length (in Mb) of ROH i. F_{ROH} is defined as:

$$F_{ROH} = \sum_i \frac{l_i}{3Gb} \tag{1}$$

where the length of the autosomal genome is estimated at 3 Gb base pairs (i.e., 3,000 Mb Clark et al., 2019).

Let $\overline{F_{ROH}}$ denote the average F_{ROH} across the individuals in a population or ethnicity. We obtained estimates of $\overline{F_{ROH}}$ both for each Ethnologue ethnicity (for the main analyses in our paper) and for each HO population (for Supplementary Figures C.6.1 and C.6.2 below in this Appendix).

Before calculating the inbreeding coefficient F_{ROH}, we applied one additional QC filter (in addition to the QC filters described above): following the ROHgen2 guidelines, we filtered out the SNPs with minor allele frequency (MAF) less than 0.05.[13] Filtering out SNPs with low MAF is important because including those SNPs may result in more ROH that are not due to marital practices. Ideally, we would calculate MAF for each population before filtering out low-MAF SNPs. However, due to data limitations and small population sizes for many populations, this was not possible. We thus calculated MAF using all (post-baseline-QC) individuals in the dataset before filtering out SNPs with MAF < 0.05. 369,741 autosomal SNPs were left after this step, and these were used to compute F_{ROH}.[14]

After we applied the MAF filter, we calculated F_{ROH} using the ROHgen2 pipeline (Clark et al., 2019). The ROHgen2 pipeline calculates F_{ROH} using PLINK v2.00a3 (Chang et al., 2015) with the following parameters: –homozyg-window-snp 50; –homozyg-snp 50; –homozyg-kb 1500; –homozyg-gap 1000; –homozyg-density 50; –homozyg-window-missing 5; –homozyg-window-het 1.These parameters have been previously shown to call ROH that mainly correspond to alleles that are identical-by-descent (IBD), mitigating concerns that measured ROH are due to chance arrangements of independent homozygous SNPs (Clark et al., 2019). In addition, these parameters have been previously shown to be robust to genotyping array choice (Joshi, 2015).

The following Appendix Table shows summary statistics for F_{ROH} calculated across the individuals in each geographic region of the world.[15] As can be seen, F_{ROH} tends to be higher

---

[13] There are typically two possible variants at a given SNP's location in the genome, and a SNP's MAF is the fraction the variants in the population that are the rarer variant at that SNP's location.

[14] Instead of using all individuals in the dataset to calculate MAF, we could have chosen to calculate MAF separately for each language family in the data. This is what we do for our calculations of F_{SNP} (see below). (The language families are higher-level language groupings from the Ethnologue; each Ethnologue ethnicity corresponds to a single language, which belongs to a language family. Populations in the same language family tend to be ethnically and thus genetically more similar to each other and thus should exhibit similar allele frequencies.) However, we chose not to use language families to calculate MAF prior to MAF filtering for our preferred estimate of F_{ROH} for the following two reasons. First, the correlation between F_{ROH} computed after MAF filtering using MAF calculated with all individuals and F_{ROH} computed after MAF filtering using MAF calculated separately for each language group is almost unity ($\hat{r} = 0.998$). The two methods of computing F_{ROH} thus yield nearly identical quantities. Second, computing allele frequencies using all individuals allows us to compute F_{ROH} for a few individuals that were not mapped to a language family.

[15] The geographic regions we use are from the United Nations' "M49 standard" (*Standard Country or Area Codes for Statistical Use (M49)*).

in regions that are far from East Africa (such as the Americas and Melanesia), and lower in regions that are close to Africa (such as Northern and Sub-Saharan Africa). The reason for this is that, as mentioned above, F ecreases with expected heterozygosity, which in turn decreases with migratory distance from East Africa (Pemberton and Rosenberg, 2014). In the paper, the empirical analyses with our F_{ROH} estimates include controls for both migratory distance from East Africa and expected heterozygosity.

Appendix Table. Summary statistics for F_{ROH} in each geographic region of the world.

Region	Mean	SD	Minimum	Maximum	IQR*
Latin America and the Caribbean	0.0598	0.0442	0.0080	0.2579	0.0537
Melanesia	0.0363	0.0264	0.0072	0.1815	0.0222
Northern America	0.0345	0.0214	0.0044	0.0953	0.0286
Southern Asia	0.0263	0.0318	0.0005	0.1911	0.0240
South−eastern Asia	0.0247	0.0242	0.0032	0.1172	0.0305
Western Asia	0.0239	0.0243	0.0011	0.1318	0.0259
Eastern Europe	0.0198	0.0198	0.0018	0.3090	0.0187
Northern Africa	0.0178	0.0200	0.0011	0.1079	0.0205
Central Asia	0.0175	0.0265	0.002	0.1511	0.0071
Eastern Asia	0.0151	0.014	0.0019	0.0845	0.0090
Sub-Saharan Africa	0.0142	0.0179	0.0000	0.1441	0.0143
Northern Europe	0.0106	0.0087	0.0017	0.0664	0.0049
Western Europe	0.0100	0.0076	0.0023	0.0494	0.006
Southern Europe	0.0099	0.0079	0.0012	0.0554	0.0057

*Interquartile range.

C.6.2 Estimating F_{SNP}

As a check to ensure that our calculation of F_{ROH} yielded reasonable estimates, we also computed estimates for another measure of inbreeding, F_{SNP}, and then compared the two sets of estimates. F_{SNP} calculates excessive homozygosity in an individual, where the excess is relative to the level of homozygosity one would expect in a population of randomly mating individuals. F_{SNP} can be strongly biased as an estimator of inbreeding due to factors other than inbreeding, such as population structure and small population size (Clark et al., 2019), and we thus do not use F_{SNP} as the main estimator of inbreeding in our analyses. However, following Clark et al. (2019), we calculate F_{SNP} and compare our F_{SNP} and F_{ROH} estimates.

Thus, for each individual in our data we calculated F_{SNP}, defined as

$$F_{SNP} = \frac{O(HOM) - E(HOM)}{M - E(HOM)}, \tag{2}$$

where $O(HOM)$ is the observed number of homozygous SNPs and $E(HOM)$ is the expected number of homozygous SNPs in the individual's autosomal genome, and M is the number of non-missing genotyped SNPs on the individual's autosomes (Clark et al., 2019). The expected number of homozygous SNPs is given by $E(HOM) = \sum_i \left[(p_i)^2 + (1 - p_i)^2\right]$, where p_i is the MAF of SNP i,[16] and is equal to $M\left(1 - \hat{H}\right)$, where $\left(1 - \hat{H}\right)$ is expected homozygosity and $\hat{H}$ is expected heterozygosity (defined below).

We define $\overline{F_{SNP}}$ to be the average F_{SNP} across individuals in a population (this quantity is often denoted as F_{IS}). Populations in which consanguineous marriages are the main drivers

---

[16] This follows from the fact that, if mating is random (i.e., under Hardy-Weinberg equilibrium), the event in which an individual has only one copy of the rarer allele (and one copy of the more common allele) – i.e., is *not* homozygous – at SNP i follows a binomial distribution with two trials (corresponding to the two locations on the maternal and paternal chromosomes) and probability p_i. From this, the probability that an individual in a randomly mating population is homozygous at SNP i is $1 - 2p_i(1 - p_i) = (p_i)^2 + (1 - p_i)^2$.

of ROH have $\overline{F_{SNP}} \approx \overline{F_{ROH}}$, since inbreeding is the main driver of excessive homozygosity in these populations (Clark et al., 2019).

We used a slightly different set of SNPs to calculate F_{SNP} compared to F_{ROH}. Prior to filtering out SNPs with MAF < 0.05, for F_{ROH} we calculated allele frequencies using all individuals in the data. In contrast, for F_{SNP} we calculated allele frequencies separately for each language family before filtering out SNPs with MAF < 0.05.

Though F_{SNP} is calculated for each individual, the formula uses expected homozygosity as an input. Ideally, we would have calculated expected homozygosity separately for each population in the data. However, due to the small sample sizes for many populations in HO, calculating expected homozygosity separately for each population leads to noisy estimates of F_{SNP}. We thus calculated expected homozygosity for each language family, because populations in the same language family tend to be ethnically and thus genetically more similar to each other and thus should exhibit similar expected homozygosities.We merged language families with fewer than 40 individuals together to eliminate small language families.

We calculated F_{SNP} using the ROHgen2 pipeline (Clark et al., 2019). The ROHgen2 pipeline calculates F_{SNP} using PLINK's '–het' command. Although the ROHgen2 pipeline does not include the 'small-sample' flag with the '—het' command, we included the 'small-sample' flag for our calculations.

C.6.3 Diagnostic plots and quality checks

In Figure C.6.1, we display several histograms that are produced by the ROHgen2 pipeline, using individual-level estimates of F_{ROH} and F_{SNP}. The figure compares the distribution of F_{ROH} and F_{SNP} to the distribution of these variables in a sample population from the UK.[17] As can be seen, the distributions of F_{ROH} and F_{SNP} (denoted 'f_roh' and 'f_het' in the two bottom panels of the figure) in the sample UK population (in light blue) lie in the left tail of the distributions in HO (in pink).

Figure C.6.2 plots the inbreeding coefficients, with HO-population-level estimates of $\overline{F_{ROH}}$ on the y-axis and $\overline{F_{SNP}}$ on the x-axis. (We note that the empirical analyses in the main paper use Ethnologue-ethnicity level estimates of $\overline{F_{ROH}}$ rather than HO-population level estimates.)

As Panel A shows, populations in language groups with large sample sizes have a strong tendency to lie close to the 45-degree line in the figure. For populations with data points near the 45-degree line, $\overline{F_{SNP}} \approx \overline{F_{ROH}}$, which signifies that consanguinity is a main driver of excessive homozygosity for these populations (Clark et al., 2019). However, populations in language groups with small samples sizes tend to lie to the left of the 45-degree line; the reason for that is that small sample sizes lead to a more sizeable downward bias in F_{SNP}, which skews data points to the left in the figure.[18] Populations to the right of the 45-degree line tend to be characterized by admixture, since admixture biases F_{SNP} upwards (Clark et al., 2019).

In Panel B, we label several populations that are to the right of the 45-degree line. Three populations in Botswana— "Taa_East:Botswana" (label 400), "Taa_North:Botswana" (label 401), and "Taa_West:Botswana" (label 402)—are located noticeably to the right of the 45-degree line. These three populations are pooled in the "Niger-Congo" language family, and sufficient divergence from other populations in this group may explain the inflation of F_{SNP}

---

[17] While the set of SNPs we used to compute F_{ROH} in our main analyses is different than the one we used to compute F_{SNP}, for Supplementary Figure A1, for comparability, for each language family we used the exact same set of SNPs to compute F_{ROH} as the one we used to compute F_{SNP}.

[18] The reason for the downward bias is that in small samples, MAF ($\hat{p}_i$) is estimated with less precision, which by Jensen's inequality leads to a upward bias in estimates of $(p_i)^2 + (1 - p_i)^2$ and thus upward bias in estimates of $E(HOM)$ and downwards bias in estimates of F_{SNP}.

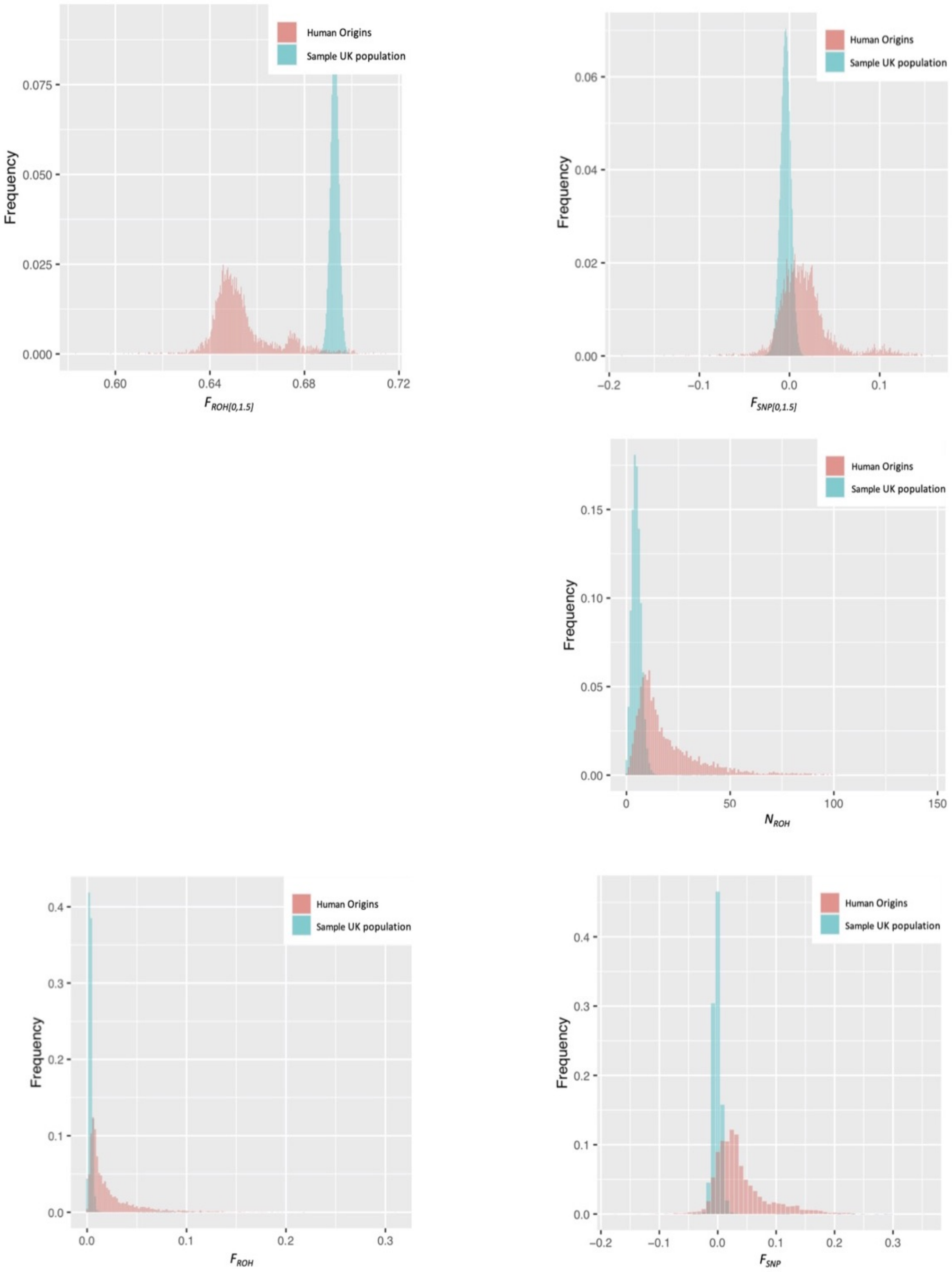
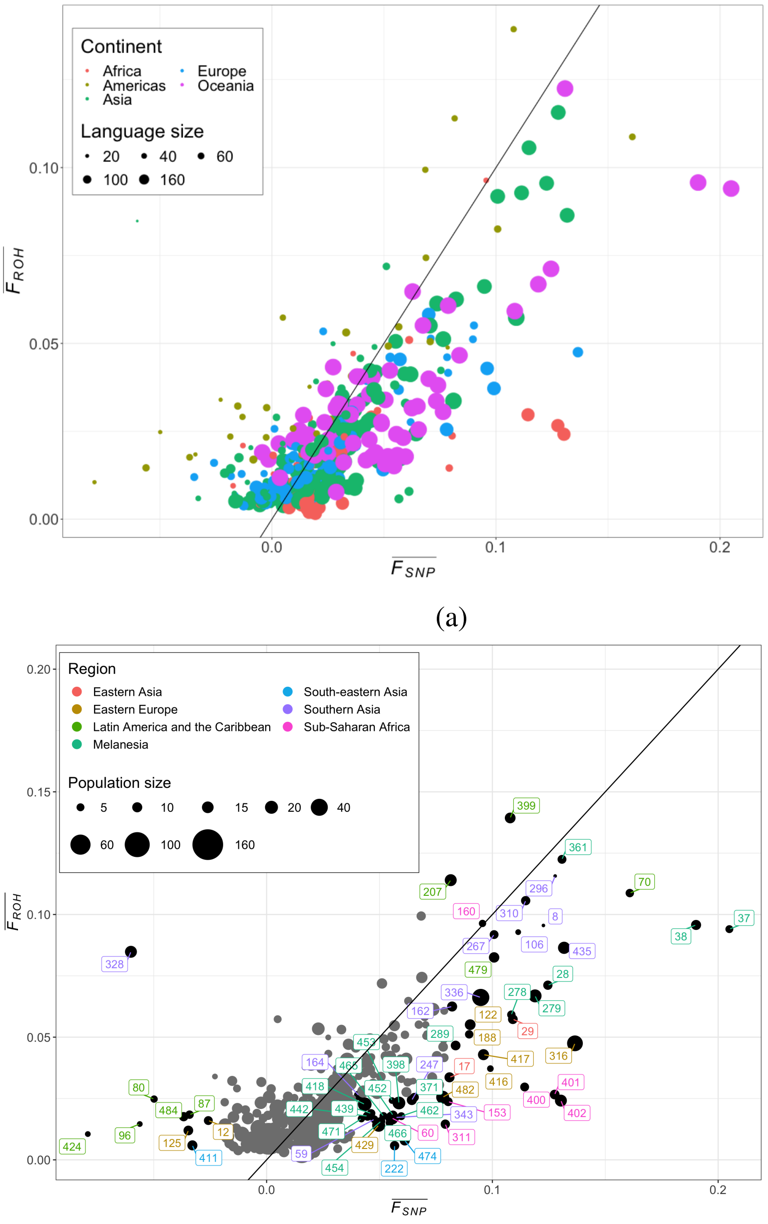


Figure C.6.1: This figure is produced by the ROHgen2 pipeline. It compares F_{ROH} and F_{SNP} in our data to those in a sample UK population. Note that F_{ROH} and F_{SNP} here are calculated using allele-frequency filters applied separately for each language language family. F_{ROH}, as described above, is the fraction of the autosomal genome in ROH longer than 1.5 Mb; $F_{ROH[0,1.5]}$ denotes the fraction of the autosome homozygous outside ROH. F_{SNP}, as described above, is calculated using all SNPs in the autosomal genome; $F_{SNP[0,1.5]}$ denotes F_{SNP} calculated in the autosomal genome outside ROH longer than 1.5 Mb; N_{ROH} denotes the count of ROH longer than 1.5 Mb.



(a)

(b)

Figure C.6.2: The two panels both display the mean F_{ROH} on the y-axis and mean F_{SNP} on the x-axis for each population. The solid black line in both panels is the 45-degree line. In Panel A, each circle is a population and the size of the circle corresponds to the number of individuals in each language family (after merging several language families with less than 40 individuals). In Panel B, the size of each circle corresponds to the number of individuals in each population. Several outliers are labeled in Panel B. Populations in lower right quadrant are labeled if ($\overline{F_{ROH}} \leq 0.0265$ and $\overline{F_{SNP}} \geq 0.04$) or ($\overline{F_{ROH}} > 0.0265$ and $\overline{F_{SNP}} \geq 0.08$). The list of populations corresponding to the numbers labeling the circles in Panel B is shown in **Appendix Table C.6.1**. '$\overline{F_{ROH}}$' denotes the population's mean F_{ROH}; '$\overline{F_{ROH}}$' denotes the population's mean F_{SNP}.

Appendix Table C.6.1. Labeled populations in Panel B of Supplementary Figure A2.

Label	HO population	Mean F_{ROH}	Mean F_{SNP}	Sample size	Continent
8	Agamudayar:India	0.0955	0.1226	1	Asia
12	Aleut:Russia	0.0160	-0.0258	7	Europe
17	Ami:Taiwan	0.0336	0.0810	10	Asia
28	Ata:Papua New Guinea	0.0712	0.1246	8	Oceania
29	Atayal:Taiwan	0.0573	0.1091	9	Asia
37	Baining_Malasait:Papua New Guinea	0.0940	0.2050	5	Oceania
38	Baining_Marabu:Papua New Guinea	0.0957	0.1902	10	Oceania
59	Bhunjiya:India	0.0165	0.0477	5	Asia
60	Biaka:Central African Republic	0.0170	0.0555	20	Africa
70	Cabecar:Costa Rica	0.1087	0.1609	6	Americas
80	Chilote:Chile	0.0248	-0.0498	4	Americas
87	Chulucanas:Peru	0.0184	-0.0341	6	Americas
96	Cusco2:Peru	0.0146	-0.0562	2	Americas
106	Dhokkali:India	0.0928	0.1114	2	Asia
122	Eskimo_Naukan:Russia	0.0551	0.0902	12	Europe
125	Even:Russia	0.0120	-0.0347	9	Europe
153	Gui:Botswana	0.0237	0.0804	7	Africa
160	Hadza1:Tanzania	0.0964	0.0956	4	Africa
162	Hakki_Pikki:India	0.0625	0.0822	10	Asia
164	Hallaki:India	0.0263	0.0408	9	Asia
188	Itelmen:Russia	0.0512	0.0897	6	Europe
207	Karitiana:Brazil	0.1140	0.0816	16	Americas
222	Kinh:Vietnam	0.0058	0.0567	8	Asia
247	Kusunda:Nepal	0.0244	0.0643	10	Asia
267	Malaikuarvar:India	0.0918	0.1008	7	Asia
278	Mamusi_Paleabu:Papua New Guinea	0.0592	0.1084	6	Oceania
279	Mamusi:Papua New Guinea	0.0668	0.1189	20	Oceania
289	Mengen:Papua New Guinea	0.0466	0.0837	9	Oceania
296	Mudaliar:India	0.1157	0.1278	1	Asia
310	Narikuruvar:India	0.1056	0.1147	8	Asia
311	Naro:Botswana	0.0145	0.0792	8	Africa
316	Nganasan:Russia	0.0475	0.1366	33	Europe
328	Onge:India	0.0848	-0.0601	16	Asia
336	Palliyar:India	0.0662	0.0948	41	Asia
343	Parhaiya:India	0.0173	0.0571	5	Asia
361	Rennell_and_Bellona:Solomon Islands	0.1225	0.1308	7	Oceania
371	Santa_Cruz:Solomon Islands	0.0255	0.0655	7	Oceania
398	Sulka:Papua New Guinea	0.0232	0.0586	19	Oceania
399	Surui:Brazil	0.1394	0.1079	12	Americas
400	Taa_East:Botswana	0.0297	0.1143	7	Africa
401	Taa_North:Botswana	0.0266	0.1276	9	Africa
402	Taa_West:Botswana	0.0241	0.1303	16	Africa
411	Thai:Thailand	0.0059	-0.0329	10	Asia
416	Todzin:Russia	0.0372	0.0991	3	Europe
417	Tofalar:Russia	0.0429	0.0961	13	Europe
418	Tolai:Papua New Guinea	0.0226	0.0435	23	Oceania
424	Tumbes:Peru	0.0105	-0.0792	2	Americas
429	Tuvinian:Russia	0.0141	0.0496	20	Europe
435	Ulladan:India	0.0864	0.1318	17	Asia
439	Vanuatu_Ambrym:Vanuatu	0.0190	0.0463	6	Oceania
442	Vanuatu_Efate_Mele:Vanuatu	0.0178	0.0447	8	Oceania
452	Vanuatu_Malekula:Vanuatu	0.0182	0.0560	8	Oceania
453	Vanuatu_Nguna:Vanuatu	0.0242	0.0556	3	Oceania
454	Vanuatu_Paama:Vanuatu	0.0157	0.0506	7	Oceania
462	Vanuatu_Santo_Mbauk:Vanuatu	0.0178	0.0597	5	Oceania
465	Vanuatu_Santo_Port_Olry:Vanuatu	0.0179	0.0518	5	Oceania
466	Vanuatu_Santo_Rango:Vanuatu	0.0151	0.0545	5	Oceania
471	Vanuatu_Tongoa:Vanuatu	0.0168	0.0420	4	Oceania
474	Vietnamese:Vietnam	0.0079	0.0612	10	Asia
479	Xavante:Brazil	0.0825	0.1008	11	Americas
482	Yakut:Russia	0.0255	0.0780	20	Europe
484	Yaquis:Peru	0.0176	-0.0367	8	Americas

Table C.6.1: The table lists the populations corresponding to the numbers labeling the circles in Panel B of Figure **C.6.2.**

relative to F_{ROH} in these populations. A similar explanation likely explains why the populations "Baining_Malasait:Papua New Guinea" (label 37) and "Baining_Marabu:Papua New Guina" (label 38) from the "Austronesian (Melanesia)" language family are far to the right of the 45-degree line.

(We also note that, in principle, populations that appear in the North-West quadrant of the figure, with high $\overline{F_{ROH}}$ and low $\overline{F_{SNP}}$, would tend to be genetically isolated (Clark et al., 2019). However, there is no population whose language family has a large sample size that is far to the left of the 45-degree line the figure.)

We conclude from this comparison exercise between our estimates of F_{ROH} and F_{SNP} that our estimates of F_{ROH} likely partly capture average levels of inbreeding across populations. This is consistent with the results of Sahoo et al. (2021) and with the analyses we report in Section 6 of the main text.

C.7 Mean regional pairwise F_{ST}: background and estimation

Genetic isolation occurs when a population has little genetic mixing with other populations. In humans, this can occur due to geographical or cultural distance or taboos prohibiting mating with individuals from other populations. Genetic isolation is associated with reduced mate choice and thus with increased cryptic inbreeding, in which two parents may have a recent common ancestor in common by chance (and not due to a deliberate cultural practice of mating relatives) (Pemberton and Rosenberg, 2014).

F_{ST} is commonly used to quantify genetic distance, or genetic divergence between populations. Two populations with a low genetic distance will have similar distributions of alleles, and two genetically distant populations will have relatively large between-population differences and relatively small within-population differences (Pemberton and Rosenberg, 2014).

Following Pemberton and Rosenberg (2014), we proxy for genetic isolation for a given population by first calculating the mean of the population's pairwise F_{ST} with other populations in the same geographic region of the world.[19] Pemberton and Rosenberg (2014) showed that mean regional pairwise F_{ST} correlates positively with ROH (Pemberton and Rosenberg, 2014).

Our estimator of F_{ST} is the Weir and Cockerham (1984) estimator. We only calculated pairwise F_{ST} for two HO populations in a region if both populations have at least five individuals. After excluding HO populations with less than five individuals, we obtained 371 populations that are mapped to 545 Ethnologue ethnicities.[20] For each of these 371 populations, we then calculated the mean regional pairwise F_{ST} as the average pairwise F_{ST} across all other populations in the same region that have at least eight individuals (Thus, while we computed the mean regional pairwise F_{ST} for all HO populations with at least five individuals, we only used other populations in the same region with at least eight individuals to compute each population's mean regional pairwise F_{ST}).

After calculating the mean regional pairwise F_{ST} for each of the 371 HO populations, we assigned each HO individual their HO population's mean regional pairwise F_{ST} estimate, and

---

[19] The world geographic regions we use are again those from the United Nations' "M49 standard" (*Standard Country or Area Codes for Statistical Use (M49)*). We merged Polynesia and Melanesia for this analysis because the Polynesia region only had 6 individuals.

[20] Note that we have only 371 HO populations here, while the main sample we obtained after applying our baseline QC filters as described in Section C.4) comprised 435 HO populations. These numbers differ because in Section C.4), we dropped individuals who belong to *Ethnologue ethnicities* with fewer than 5 individuals, whereas here when calculating F_{ST} we apply an additional filter and drop individuals who belong to *HO populations* with fewer than 5 individuals.

then obtained estimates of the mean regional pairwise F_{ST} for each of the 545 Ethnologue ethnicities by taking the average across the ethnicity's individuals (As mentioned above, for the analyses reported in the main text, of the relationships between F_{ROH} and cousin marriage preferences analyses and between nighttime luminosity and F_{ROH}, we only used subsets of the 545 ethnicities).

To compute pairwise F_{ST} for any two populations, we took the set of SNPs and individuals that passed our baseline QC filters and removed all SNPs with MAF < 0.05. That step removed 115,203 SNPs, resulting in 369,741 remaining SNPs in the data. We then used PLINK to LD prune the remaining SNPs. We used PLINK's '–indep-pairwise' command to LD prune the SNPs, with window-size 1E6 kb, step-size 50 kb, r^2 threshold of 0.1. This left a total of 50,739 quasi-independent SNPs for the F_{ST} calculations. (As mentioned below, we used that same set of SNPs to compute the top 20 Principal Components (PCs) in the HO data.) We calculated pairwise F_{ST} using PLINK's '–fst' and then used the reported weighted global mean as the pairwise F_{ST}.

C.8 Expected heterozygosity: background and estimation

Another correlate of F_{ROH} is expected heterozygosity (Pemberton and Rosenberg, 2014). Expected heterozygosity is a measure of genetic diversity in a population; it is defined as the probability that two randomly selected individuals from a population have different alleles at a randomly selected location on one chromosome in the genome. Equivalently, if people mate at random (i.e., under Hardy-Weinberg equilibrium), expected heterozygosity is the probability that the two alleles one has at a randomly selected location in the genome are identical across one's two chromosomes.

ROH and F_{ROH} are negatively correlated with expected heterozygosity across populations (Pemberton and Rosenberg, 2014). This is because, if there are relatively more homozygotes in the population, individuals will tend to exhibit more ROH, even in the absence of consanguineous cultural practices. Consider the following simple example, where there is no genetic variation in the population and all individuals are therefore fully homozygous. If we attempt to infer autozygosity by counting ROH in a sample of individuals from the population, we will incorrectly conclude that the population is fully inbred.

Expected heterozygosity has been shown to be highly correlated with migratory distance from Africa (Ramachandran et al., 2005). This relationship is consistent with a model in which, as human beings migrated outside the "cradle of humankind" in East Africa and spread across the globe, the individuals who migrated to new colonies only carried with them a subset of the genetic diversity of the parent colony (Ramachandran et al., 2005). This is known as the "serial founder effect."[21]

Expected heterozygosity has also been found to be (non-linearly) associated with economic development (Ashraf and Galor, 2013), though some have questioned the robustness of these results (Rosenberg and Kang, 2015). Thus, it is important to control for expected heterozygosity (and other confounders) in order to confidently use F_{ROH} as a proxy for consanguineous

---

[21] Another factor that may contribute to this relationship is ascertainment bias in genotyping microarrays (Patterson et al., 2012). Only a subset of SNPs are genotyped in microarrays, and SNPs with a higher MAF in a reference population are typically favored. For instance, if the reference population is European, then the inclusion of SNPs with high MAF among Europeans on the genotyping array may lead to an upward bias in estimates of expected heterozygosity and downward bias in estimates of ROH and F among Europeans vs. among other populations. (By definition, MAF are in the interval [0, 0.5], so higher MAF are closer to 0.5.) The Affymetrix Human Origins array used for the dataset we analyze, HO, has been specifically designed to mitigate ascertainment bias (Patterson et al., 2012), but ascertainment bias may nonetheless be a concern.

unions, especially in our analyses of the relationship between F_{ROH} and economic development.

As we did to estimate mean regional pairwise F_{ST}, we only computed expected heterozygosity for HO populations that have at least five individuals. After excluding populations with less than five individuals, we obtained 371 populations that are mapped to 545 Ethnologue ethnicities. After calculating expected heterozygosity (as described below) for each of the 371 HO populations, we assigned each HO individual their HO population's expected heterozygosity estimate, and then obtained estimates of the expected heterozygosity for each of the 545 Ethnologue ethnicities by taking the average across the ethnicity's individuals. (As mentioned above, for the analyses with $\overline{F_{ROH}}$ reported in the main text, we only used subsets of these 545 ethnicities.)

We calculated expected heterozygosity for each SNP i and each HO population j using Nei and Roychoudhury's formula for expected heterozygosity (Nei and Roychoudhury, 1974)[22]:

$$\hat{h}_{ij} = \frac{n_j}{n_j - 1}\left(1 - (p_{ij})^2 + \left(1 - p_{ij}\right)^2\right), \tag{3}$$

where p_{ij} is the minor allele frequency (MAF) of SNP i in population j and n_j is the number of alleles in the data for population j.[23] We calculated expected heterozygosity for population j by summing up the $\hat{h}_{ij}$ across all SNPs:

$$\hat{H}_j = \sum_i \hat{h}_{ij}. \tag{4}$$

Before calculating expected heterozygosity for each population, we took the SNPs that passed our baseline QC filters and removed all SNPs with MAF < 0.05 and missing call rates > 0.01. Note that we calculated MAF here using all individuals in the HO data that passed our baseline QC. We applied a more stringent missing call rate threshold of 0.01 (compared to 0.03, which is the threshold in our baseline QC) to reduce the number of SNPs that have very few individuals in a population with non-missing genotypes for those SNPs.

After applying those QC filters, we used PLINK to LD prune the remaining SNPs (i.e., to obtain a subset of SNPs that are no more than weakly correlated with one another in the genome). We used PLINK's '–indep-pairwise' to LD prune the SNPs, with window-size 1E6 kb, step-size 50 kb, r^2 threshold of 0.1. After LD pruning, we obtained 50,383 quasi-independent SNPs.

We then removed all SNPs that had at least one population with strictly less than four non-missing alleles for that SNP. This resulted in an additional four SNPs being removed. Therefore, we had 50,379 SNPs for our expected heterozygosity calculations.[24]

To obtain the quantities necessary to calculate expected heterozygosity, we used PLINK's "–freq" which reports the minor allele frequencies and number of non-missing alleles for each SNP.[25]

---

[22] Nei and Roychoudhury's method (1974) yields unbiased estimates of expected heterozygosity only in samples with no inbred or related individuals (DeGiorgio and Rosenberg, 2009). Here, our F_{ROH} estimates clearly suggest inbreeding is present in some populations (see Section 6) and some individuals are likely related, so our estimates of $\hat{H}_j$ are not unbiased. We expect this to generate no more than a small bias in our the empirical analyses we report in the paper.

[23] Thus, if there are N_j individuals in population j in the dataset and if there are no missing observations for that SNP, $n_j = 2N_j$.

[24] Note that we also calculated expected heterozygosity after removing all SNPs that had at least one population with strictly less than eight non-missing alleles (instead of less than four non-missing alleles). This removed 892 SNPs. The estimated correlation between the estimates of expected heterozygosity obtained using these two different sets of SNPs is unity.

[25] Note that, unlike for the estimation of expected homozygosity (which is equal to one minus expected het-

C.9 The top 20 Principal Components (PCs): background and estimation

In order to control for population structure, we also calculated the top 20 principal components (PCs) of the genotypic data for each individual. For each Ethnologue ethnicity, we then obtained the mean of each PC by taking the average across the ethnicity's individuals.

The top PCs have been shown to correlate strongly with ancestry (Novembre et al., 2008; Price et al., 2006). Controlling for the top PCs is standard in analyses with individual-level genetic data and can mitigate bias due to population stratification.[26] Population stratification could pose a threat to our empirical analyses since allele frequencies affect homozygosity and thus F_{ROH}.

To compute the top 20 PCs, we used PLINK's "–pca" command with the same set of 50,739 SNPs that we used to compute the pairwise F_{ST} between populations (see Section C.8).

C.10 Summary statistics

Table C6 reports the mean, standard deviation, minimum, and maximum for the core variables we computed with the genetic data: the inbreeding coefficient ($\overline{F_{ROH}}$), mean regional pairwise F_{ST}, and expected heterozygosity. These summary statistics were computed across the 416 Ethnologue ethnicities with eight or more HO individuals (and which we used in the analysis, reported in the main text, of the relationship between the inbreeding coefficient and cousin marriage preferences). The Table does not report these summary statistics for the top 20 PCs, which are all normalized to have a mean of zero and unit variance.

---

erozygosity) when computing F_{SNP}, which we computed at the language level using PLINK's "–het" command with the 'small-sample' flag, here we used Nei & Roychoudhury's formula to compute population-level expected heterozygosity. The main reason for the difference is that PLINK's "–het" command computes expected heterozygosity at the individual level using each individual's nonmissing SNPs, and here we wanted to use the same set of SNPs to compute expected heterozygosity for all populations.

[26] Population stratification refers to systematic differences in allele frequencies that correlate across populations with cultural or environmental differences that causally impact variables of interest (Hamer, 2000).

References

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

Beauchamp, J. P., Cesarini, D., Johannesson, M., van der Loos, M. J. H. M., Koellinger, P. D., et al. Molecular genetics and economics. *The Journal of Economic Perspectives*, 25(4): 57–82, 2011.

Ceballos, F. C., Joshi, P. K., Clark, D. W., Ramsay, M., and Wilson, J. F. Runs of homozygosity: Windows into population history and trait architecture. *Nature Reviews Genetics*, 19(4):220–234, 2018.

Center for International Earth Science Information Network (CIESIN) at Columbia University. Gridded Population of the World, Version 4 (GPWv4): Population Count Adjusted to Match 2015 Revision of UN WPP Country Totals. NASA Socioeconomic Data and Applications Center (SEDAC), 2016.

Chang, C. C., Chow, C. C., Tellier, L. C. A. M., Vattikuti, S., Purcell, S. M., et al. Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, 4(1), 2015.

Clark, D. W., Okada, Y., Moore, K. H., Mason, D., Pirastu, N., et al. Associations of autozygosity with a broad range of human phenotypes. *Nature communications*, 10(4957):1–17, 2019.

DeGiorgio, M. and Rosenberg, N. A. An unbiased estimator of gene diversity in samples containing related individuals. *Molecular Biology and Evolution*, 26(3):501–512, 2009.

Elvidge, C. D., Baugh, K. E., Dietz, J. B., Bland, T., Sutton, P. C., et al. Radiance calibration of DMSP-OLS low-light imaging data of human settlements. *Remote Sensing of Environment*, 68(1):77–88, 1999.

Gennaioli, N., La Porta, R., Lopez-de Silanes, F., and Shleifer, A. Human capital and regional development. *The Quarterly journal of economics*, 128(1):105–164, 2013.

Gennaioli, N., La Porta, R., De Silanes, F. L., and Shleifer, A. Growth in regions. *Journal of Economic growth*, 19(3):259–309, 2014.

Gordon, R. G., Jr., editor. *Ethnologue: languages of the world.* SIL International, 15th edition, 2005.

Gorodnichenko, Y. and Roland, G. Culture, institutions, and the wealth of nations. *The Review of Economics and Statistics*, 99(3):402–416, 2016.

Hamer, D. H. Beware the chopsticks gene. *Molecular psychiatry*, 5(1):11–13, 2000.

Jeong et al., C. The genetic history of admixture across inner Eurasia. *Nature Ecology and Evolution*, 3(6):966–976, 2019.

Joshi, P. K. e. a. Directional dominance on stature and cognition in diverse human populations. *Nature*, 523(7561):459–462, 2015.

Kang, J. T., Goldberg, A., Edge, M. D., Behar, D. M., and Rosenberg, N. A. Consanguinity Rates Predict Long Runs of Homozygosity in Jewish Populations. *Human Heredity*, 82(3-4): 87–102, 2017.

Kirby, K., Gray, R. D., Greenhill, S. J., Jordan, F. M., Gomes-Ng, S., et al. D-PLACE: A global database of cultural, linguistic and environmental diversity. *PLoS One*, 11(7):e0158391, 2016.

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

Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., et al. Ancient human genomes suggest three ancestral populations for present-day europeans. *Nature*, 513:409–413, 2014.

Lazaridis, I., Nadel, D., Rollefson, G., Merrett, D. C., Rohland, N., et al. Genomic insights into the origin of farming in the ancient Near East. *Nature*, 536(7617):419–424, 2016.

Lewis, M. P., editor. *Ethnologue: Languages of the World*. SIL International, Dallas, Texas, 16 edition, 2009.

Lipson, M., Cheronet, O., Mallick, S., Rohland, N., Oxenham, M., et al. Ancient genomes document multiple waves of migration in Southeast Asian prehistory. *Science*, 361(6397): 92–95, 2018.

McQuillan, R., Leutenegger, A.-L., Abdel-Rahman, R., Franklin, C. S., Pericic, M., et al. Runs of Homozygosity in European Populations. *American Journal of Human Genetics*, 83(3): 359–372, 2008.

Murdock, G. P. Ethnographic atlas: a summary. *Ethnology*, 6(2):109–236, 1967.

Nakatsuka, N., Moorjani, P., Rai, N., Sarkar, B., Tandon, A., et al. The promise of discovering population-specific disease-associated genes in South Asia. *Nature Genetics*, 49:1403–1407, 2017.

Nei, M. and Roychoudhury, A. K. Sampling variances of heterozygosity and genetic distance. *Genetics*, 76(2):379–390, 1974.

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., et al. Genes mirror geography within Europe. *Nature*, 456(7218):98–101, 2008.

Nunn, N. and Puga, D. Ruggedness: the blessing of bad geography in Africa. *The Review of Economics and Statistics*, 94(1):20–36, 2012.

Ochoa, A. and Storey, J. New kinship and F_{ST} estimates reveal higher levels of differentiation in the global human population. Working Paper. Available at bioRxiv, 2019.

Özak, Ö. Distance to the pre-industrial technological frontier and economic development. *Journal of Economic Growth*, 23(2):175–221, 2018.

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., et al. Ancient Admixture in Human History. *Genetics*, 192(3):1065–1093, 2012.

Pemberton, T. J. and Rosenberg, N. A. Population-Genetic Influences on Genomic Estimates of the Inbreeding Coefficient: A Global Perspective. *Human Heredity*, 77(1–4):37–48, 2014.

Pemberton, T. J., Absher, D., Feldman, M. W., Myers, R. M., Rosenberg, N. A., et al. Genomic patterns of homozygosity in worldwide human populations. *The American Journal of Human Genetics*, 91(2):275–292, 2012.

Pickrell, J. K., Patterson, N., Barbieri, C., Berthold, F., Gerlach, L., et al. The genetic prehistory of southern Africa. *Nature Communications*, 3(1143):1–6, 2012.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, 2006.

Ramachandran, S., Deshpande, O., Roseman, C. C., Rosenberg, N. A., Feldman, M. W., et al. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *PNAS*, 102(44):15942–15947, 2005.

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

Reich Lab. Human origins dataset, 2020.

Rosenberg, N. A. and Kang, J. T. Genetic diversity and societally important disparities. *Genetics*, 201(1):1–12, 2015.

Sahoo, S. A., Zaidi, A. A., Anagol, S., and Mathieson, I. Long runs of homozygosity are correlated with marriage preferences across global population samples. Working Paper. Available at bioRxiv, 2021.

Schulz, J. F., Bahrami-Rad, D., Beauchamp, J. P., and Henrich, J. The Church, intensive kinship, and global psychological variation. *Science*, 366(6466), 2019.

University of East Anglia Climatic Research Unit, Harris, I. C., and Jones, P. D. CRU TS4. 01: Climatic Research Unit (CRU) Time-Series (TS) version 4.01 of high-resolution gridded data of month-by-month variation in climate (Jan. 1901–Dec. 2016). Centre for Environmental Data Analysis, 2017.

Weir, B. S. and Cockerham, C. C. Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, 38(6):1358–1370, 1984.

WHO. Global status report on blood safety and availability, 2016. *Geneva. Available from: apps. who. int/iris/handle/10665/254987*, 2016.

World GeoDatasets. World Language Mapping System (WLMS) Version 19. SIL International.