Ancestry, Language and Culture

Enrico Spolaore, Romain Wacziarg

Ancestry, Language and Culture[*]

Enrico Spolaore

Tufts University and NBER

Romain Wacziarg

UCLA and NBER

May 2015

Abstract

We explore the interrelationships between various measures of cultural distance. We first discuss measures of genetic distance, used in the recent economics literature to capture the degree of relatedness between countries. We next describe several classes of measures of linguistic, religious, and cultural distances. We introduce new measures of cultural distance based on differences in average answers to questions from the World Values Survey. Using a simple theoretical model we hypothesize that ancestral distance, measured by genetic distance, is positively correlated with linguistic, religious, and cultural distance. An empirical exploration of these correlations shows this to be the case. This empirical evidence is consistent with the view that genetic distance is a summary statistic for a wide array of cultural traits transmitted intergenerationally.

1   Introduction

Populations that share a more recent common ancestry exchange goods, capital, innovations and technologies more intensively, but they also tend to fight more with each other.[1] Why does ancestral distance matter for these outcomes? In this paper, we argue that when populations split apart and diverge over the long span of history, their cultural traits also diverge. These cultural traits include language and religion but also a broader set of norms, values and attitudes that are transmitted intergenerationally and therefore display persistence over long stretches of time. In turn, these traits introduce barriers to interactions and communication between societies, in proportion to how far they have drifted from each other.

While the rate at which languages, religions and values diverged from each other over time varies across specific traits, we hypothesize and document a significant positive relationship between long-term relatedness between populations, measured by genetic distance, and a wide array of measures of cultural differences. In doing so, we provide support for the argument that the effect of genealogical relatedness on economic and political outcomes captures at least in part the effects of cultural distance. In sum, genetic relatedness is a summary statistic for a wide array of cultural traits transmitted vertically across generations. These differences in vertically transmitted traits introduce horizontal barriers to human interactions.

We begin our paper with a general discussion of measures of ancestral distance. We focus on genetic distance, a measure that has been used in a recent emerging literature on the deep roots of economic development. This measure captures how distant human societies are in terms of the frequency of neutral genes among them. It constitutes a molecular clock that allows us to characterize the degree of relatedness between human populations in terms of the number of generations that separate them from a common ancestor population. We next turn to measures of cultural differences. We consider three classes of such measures. The first is linguistic distance. Since these measures are described in great detail elsewhere, we keep our discussion brief.[2] The second class of measures is religious distance. We adopt an approach based on religious trees to

---

[1] For recent references on technological transmission, see Spolaore and Wacziarg (2009, 2012, 2013). On interstate wars, see Spolaore and Wacziarg (2015). On trade and financial flows, the literature documenting links with linguistic and cultural distance is vast. Salient references include Melitz (2008), Melitz and Toubal (2012), Guiso, Sapienza and Zingales (2009) and Egger and Toubal (2015).

[2] For instance, see Ginsburgh and Weber (2015).

characterize the distance between major world religions, and use these distances to calculate the religious distance between countries. Third, in the newest part of this paper, we define and compute a series of measures of differences in values, norms and attitudes between countries, based on the World Values Survey. We show that these classes of measures are positively correlated between each other, yet the correlations among them are not large. This motivates the quest for a summary measure of cultural differences.

We next argue that genetic distance is such a summary measure. We start with a simple model linking genetic distance to cultural distance, providing a conceptual foundation for studying the relationship between relatedness and cultural distance. The model shows that if cultural traits are transmitted from parents to children with variation, then a greater ancestral distance between populations should on average be related with greater cultural distance. This relationship holds in expectations and not necessarily in each specific case (it is possible for two genealogically distant populations to end up with similar cultural traits), but our framework predicts a positive relationship between genetic distance and cultural distance. We next investigate empirically the links between genetic distance and the aforementioned metrics of cultural distance, shedding some light on their complex interrelationships. We find that genetic distance is positively correlated with linguistic and religious distance as well as with differences in values and attitudes across countries, and is therefore a plausible measure of the average distance between countries along these various dimensions jointly.

This paper contributes to a growing empirical literature on the relationships between ancestry, language, and culture over time and space. This literature has expanded in recent years to include not only work by anthropologists, linguists, and population geneticists (such as, for instance, the classic contribution by Cavalli-Sforza, Menozzi and Piazza, 1994), but also those of economists and other social scientists interested in the effects of such long-term variables on current economic, political and social outcomes (for general discussions, see for example Spolaore and Wacziarg, 2013, and chapters 3 and 4 in Ginsburgh and Weber, 2011). Economic studies using measures of genetic and cultural distances between populations to shed light on economic and political outcomes include our own work on the diffusion of development and innovations (Spolaore and Wacziarg, 2009, 2012, 2013), international wars (Spolaore and Wacziarg, 2015) and the fertility transition (Spolaore and Wacziarg, 2014). Other studies using related approaches include Guiso, Sapienza and Zingales's (2009) investigation of cultural barriers to trade between European countries, Bai and Kung's

(2011) study of Chinese relatedness, cross-strait relations and income differences, Gorodnichenko and Roland's (2011) investigation of the relation between culture and institutions, and Desmet et al.'s (2011) analysis of the relations between genetic and cultural distances and the stability of political borders in Europe.

This paper is especially close to a section in the article by Desmet et al. (2011), where these authors provide an empirical analysis of the relationship between genetic distance and measures of cultural distance, using the World Values Survey. In particular, Desmet et al. (2011) find that European populations that are genetically closer give more similar answers to a broad set of 430 questions about norms, values and cultural characteristics included in the 2005 World Values Survey (WVS) sections on perceptions of life, family, religion and morals. They also find that the correlation between genetic distance and differences in cultural values remains positive and significant after controlling for linguistic and geographic distances. Our results here are consistent with their findings, but we use different empirical methods, a broader set of questions from all waves of the WVS, additional distances in linguistic and religious space, and a worldwide rather than European sample.

More broadly, this paper is also connected to the evolutionary literature on cultural transmission of traits and preferences and the coevolution of genes and culture (e.g., Cavalli-Sforza and Feldman, 1981; Boyd and Richerson, 1985; Richerson and Boyd, 2004; Bell, Richerson and McElreath, 2009; and in economics Bisin and Verdier, 2000, 2001, 2010; Seabright, 2010; and Bowles and Gintis, 2011), and to the growing empirical literature on the effects of specific genetic traits, measured at the molecular level, on economic, cultural and social outcomes.[3] However, as already mentioned, in our analysis we do not focus on the direct effects of intergenerationally transmitted traits subject to selection, but on general measures of ancestry based on neutral genes, which tend to change randomly over time, and capture long-term relatedness across populations. Finally, our work is connected to a different but related set of contributions focusing on the economic and political effects of genetic and cultural diversity not between populations, but within populations and societies (Ashraf and Galor, 2013a, 2013b; Arbatli, Ashraf and Galor, 2013, Desmet, Ortuño-Ortín and Wacziarg, 2014).

This paper is organized as follows. Section 2 addresses the measurement of ancestry using genetic distance. Section 3 discusses the constructions of each of our three classes of distances:

---

[3]For overviews and critical discussions, see for instance Beauchamp et al. (2011) and Benjamin et al. (2012).

linguistic, religious and values / norms / attitudes distances. Section 4 presents a simple theoretical framework linking genetic distance and distance in cultural traits. Section 5 reports patterns of correlations, both simple and partial, between genetic distance and cultural distance. Section 6 concludes.

2   Ancestry

2.1   Ancestry, relatedness, and genetic markers

Who is related to whom? The biological foundation of relatedness is ancestry: two individuals are biologically related when one is the ancestor of the other, or both have common ancestors. Siblings are more closely related than first cousins because they have more recent common ancestors: their parents, rather than their grandparents. It is well known that genetic information can shed light on relatedness and common ancestry at the individual level. People inherit their DNA from their parents, and contemporary DNA testing can assess paternity and maternity with great accuracy. By the same token, genetic information can help reconstruct the relations between individuals and groups who share common ancestors much farther in the past.

From a long-term perspective, all humans are relatively close cousins, as we all descend from a small number of members of the species *Homo sapiens*, originating in Africa over 100,000 years ago. As humans moved to different regions and continents, they separated into different populations. Genetic information about current populations allows us to infer the relations among them and the overall history of humankind. Typically, people all over the world tend to share the same set of gene variants (alleles), but with different frequencies across different populations. Historically, this was first noticed with respect to blood groups. The four main blood groups are A, B, AB and O, and are the same across different populations. These observable groups (phenotypes) are the outcome of genetic transmission, involving three different variants (alleles) of the same gene: A, B, and O. Each individual receives one allele from each parent. For instance, A-group people may be so because they have received two copies of allele A (homozygotes) or because they have received a copy of allele A and one of allele O (heterozygotes). In contrast, O-group people can only be homozygotes (two O alleles), and AB-group can only have an A from a parent and a B from the other parent.

By observing ABO blood groups, it is possible to infer the distribution of different alleles (A,

B and O) in a given population. The frequencies of such alleles vary across populations. For example, one of the earliest studies of blood group differences across ethnic groups, conducted at the beginning of the 20th century and cited in Cavalli-Sforza, Menozzi and Piazza (1994, p. 18) found that the proportions of A and B alleles among the English were 46.4 percent and 10.2 percent respectively, were 45.6 percent and 14.2 percent among the French, while these proportions were 44.6 percent and 25.2 percent among the Turks and 30.7 percent and 28.2 percent among the Malagasy. It is reasonably to assume that these gene frequencies have varied mostly randomly over time, as an effect of *genetic drift*, the random changes in allele frequency from one generation to the next due to the finite sampling of which specific individuals and alleles end up contributing to the next generation. Under random drift, it is unlikely that the French and the English have ended up with similar distributions of those alleles just out of chance, and more likely that their distributions are similar because they share recent common ancestors. That is, they used to be part of the same population in relatively recent times. In contrast, the English and the Turks are likely to share common ancestors farther in the past, and the English and the Malagasys even farther down the generations.

Genetic information about ABO blood groups alone would be insufficient to determine the relationships among different populations. More information can be obtained by considering a larger range of *genetic markers*, that is, genes that change across individuals, and are therefore useful to study their ancestry and relatedness. Blood groups belong to a larger set of classic genetic markers, which also include other blood-group systems (such as the RH and MN blood groups), variants of immunoglobulin (GM, KM, AM, etc.), variants of human lymphocyte antigens (HLA) and so on.

By considering a large number of classic genetic markers, pioneers in this area of human genetics, such as Cavalli-Sforza and his collaborators (e.g., see Cavalli-Sforza and Edwards, 1964; Cavalli-Sforza, Menozzi and Piazza, 1994) were able to measure global genetic differences across populations, and to use such measures to infer how different populations have separated from each other over time and space. More recently, the great advances in DNA sequencing have allowed the direct study of *polymorphisms* (that is, genetic information that differs across individuals) at the molecular level. In particular, human genetic differences can now be studied directly by looking at instances of *Single Nucleotide Polymorphism* or SNP (pronounced snip), a sequence variation in which a single DNA nucleotide – A, T, C or G – in the genome differs across individuals (for

example, Rosenberg et al., 2002; Seldin et. al., 2006; Tian et al., 2009; Ralph and Coop, 2013).[4]

2.2 Genetic distance between human populations

2.2.1 Definition of F_{ST}

In order to capture global differences in gene frequencies between populations, geneticists have devised summary measures, called *genetic distances*. One of the most widely used measures of genetic distance, first suggested by Sewall Wright (1951), is called F_{ST}. In general, it can be defined as:

$$F_{ST} = \frac{V_p}{\overline{p}(1 - \overline{p})} \tag{1}$$

where V_p is the variance between gene frequencies across populations, and $\overline{p}$ their average gene frequencies.

For example, consider two populations (a and b) of equal size, and one *biallelic* gene - i.e., a gene that can take only two forms: allele 1 and allele 2. Let p_a and $q_a = 1 - p_a$ be the gene frequency of allele 1 and allele 2, respectively, in population a.[5] By the same token, p_b and $q_b = 1 - p_b$ are the gene frequency of allele 1 and allele 2, respectively, in population b. Without loss of generality, assume $p_a \geq p_b$ and define:

$$p_a \equiv \overline{p} + \sigma \tag{2}$$

$$p_b \equiv \overline{p} - \sigma \tag{3}$$

where $\sigma \geq 0$. Then, we have:

$$F_{ST} = \frac{V_p}{\overline{p}(1 - \overline{p})} = \frac{(p_a - \overline{p})^2 + (p_b - \overline{p})^2}{2\overline{p}(1 - \overline{p})} = \frac{\sigma^2}{\overline{p}(1 - \overline{p})} \tag{4}$$

In general, $0 \leq F_{ST} \leq 1$. In particular, $F_{ST} = 0$ when the frequencies of the alleles are identical across populations ($\sigma = 0$), and $F_{ST} = 1$ when one population has only one allele and the other

---

[4] A *haplogroup* is a group of similar haplotypes (collection of specific alleles) that share a common ancestor having the same SNP mutation. Among the most commonly studied human haplogroups are those passed only down the matrilineal line in the mitochondrial DNA (mtDNA) and those passed only in the patrilineal line in the Y-chromosome. While the analysis of the distribution of these specific haplogroups across populations is extremely informative to study the history of human evolution and human migrations, measures of overall genetic distance and relatedness between populations require the study of the whole genome. The measures of genetic distance that we discuss and use in the rest of this paper capture this more comprehensive notion of relatedness between populations.

[5] Note that since $p_a + q_a = 1$ we also have $(p_a + q_a)^2 = p_a^2 + q_a^2 + 2p_a q_a = 1$.

population has only the other allele - that is, when $\sigma = \overline{p}$. In that case, we say that the gene has reached *fixation* in each of the two populations - that is, there is no *heterozygosity* within each population.

In fact, F_{ST} is part of a broader class of measures called *fixation indices*, and can be reinterpreted in terms of a comparison between heterozygosity within each population and heterozygosity in the *sum* of the two populations.[6] The probability that two randomly selected alleles at the given locus are *identical* within the population (homozygosity) is $p_a^2 + q_a^2$, and the probability that they are different (heterozygosity) is:

$$h_a = 1 - \left(p_a^2 + q_a^2\right) = 2p_aq_a \tag{5}$$

By the same token, heterozygosity in population b is:

$$h_b = 1 - \left(p_b^2 + q_b^2\right) = 2p_bq_b \tag{6}$$

The average gene frequencies of allele 1 and 2 in the two populations are, respectively:

$$\overline{p} = \frac{p_a + p_b}{2} \tag{7}$$

and:

$$\overline{q} = \frac{q_a + q_b}{2} = 1 - \overline{p} \tag{8}$$

Heterozygosity in the *sum* of the two populations is:

$$h = 1 - \left(\overline{p}^2 + \overline{q}^2\right) = 2\overline{p}\overline{q} \tag{9}$$

Average heterozygosity is measured by:

$$h_m = \frac{h_a + h_b}{2} \tag{10}$$

F_{ST} measures the variation in the gene frequencies of populations by comparing h and h_m:

$$F_{ST} = 1 - \frac{h_m}{h} = 1 - \frac{p_aq_a + p_bq_b}{2\overline{p}\overline{q}} = \frac{1}{4}\frac{(p_a - p_b)^2}{\overline{p}(1-\overline{p})} = \frac{\sigma^2}{\overline{p}(1-\overline{p})} \tag{11}$$

In sum, if the two populations have identical allele frequencies ($p_a = p_b$), F_{ST} is zero. On the other hand, if the two populations are completely different at the given locus ($p_a = 1$ and $p_b = 0$, or

---

[6]More generally, the study of genetic distance between populations is part of the broader study of human genetic variation and diversity between and within populations. Interesting discussions of the economic effects of genetic diversity *within* populations and of the relationship between genetic and cultural diversity and fragmentation are provided in Ashraf and Galor (2013a, 2013b).

$p_a = 0$ and $p_b = 1$), F_{ST} takes value 1. In general, the higher the variation in the allele frequencies across the two populations, the higher is their F_{ST} distance. The formula can be extended to account for L alleles, S populations, different population sizes, and to adjust for sampling bias. The details of these generalizations are provided in Cavalli-Sforza, Menozzi and Piazza (1994, pp. 26-27).

2.2.2 Genetic distance and separation time

F_{ST} genetic distance has a very useful interpretation in terms of *separation time*, defined as the time since two populations shared their last common ancestors - that is, since they were the same population. Consider two populations whose ancestors were part of the same population t generations ago: t is the separation time between the two populations. Assume, for simplicity, that both populations have the same effective population size N.[7] Assume also that allele frequencies change over time only as the result of random genetic drift. Then it can be shown that:[8]

$$F_{ST} = 1 - e^{-\frac{t}{2N}} \tag{12}$$

For a small F_{ST}, we can approximate it with $-\ln(1 - F_{ST})$, which implies that:

$$F_{ST} \simeq \frac{t}{2N} \tag{13}$$

This means that the genetic distance between two cousin populations is roughly proportional to the time since the ancestors of the two populations split and formed separate populations. In this respect, we can therefore interpret genetic distance as a measure of the time since two populations shared a common ancestry.

2.2.3 Empirical estimates of genetic distance

In their landmark study *The History and Geography of Human Genes*, Cavalli-Sforza, Menozzi and Piazza (1994) provide some of the most detailed and comprehensive estimates of genetic distances between human populations, within and across continents. Their initial database contains $76,676$

---

[7]Effective population size only includes active breeders, and is generally smaller than actual census size. More precisely, effective population size is the number of breeding individuals that would produce the actual sampling variance, or rate of inbreeding, if they bred in a way consistent with a series of idealized benchmark assumptions (e.g., see Falconer and Mackay, 1996, chapter 4, or Hamilton, 2009, chapter 3).

[8]See Cavalli-Sforza et al. (1994, p. 30 and references).

gene frequencies, corresponding to $6,633$ samples in different locations. By culling and pooling such samples, they restrict their analysis to 491 populations. They focus on 'aboriginal populations that were at their present location at the end of the fifteenth century when the great European migrations began' (Cavalli-Sforza et al., 1994, p. 24). When studying genetic difference at the world level, the number is reduced to 42 representative populations, aggregating subpopulations characterized by a high level of genetic similarity. For these 42 populations, Cavalli-Sforza and coauthors report bilateral distances computed from 120 alleles.

Among this set of 42 world populations, the greatest genetic distance observed is between Mbuti Pygmies and Papua New-Guineans, where the F_{ST} distance is 0.4573, while the smallest genetic distance (0.0021) is between the Danish and the English. When considering more disaggregated data for 26 European populations, the smallest genetic distance (0.0009) is between the Dutch and the Danes, and the largest (0.0667) is between the Lapps and the Sardinians. The mean genetic distance among the 861 available pairs in the world population is 0.1338. Figure 1, reproduced from Cavalli-Sforza et al. (1994, Figure 2.3.2B, p. 78), is a phylogenetic tree, constructed from genetic distance data, that visually shows how different human populations have split apart over time. The phylogenetic tree is constructed to maximize the correlation between Euclidian distances to common nodes (measured along the branches) and F_{ST} genetic distance computed from allele frequencies. Hence, the tree is a simplified summary of (but not a substitute for) the matrix of F_{ST} genetic distances between populations. Cavalli-Sforza et al. (1994) also calculated estimates of Nei's distance, which is a different measure of genetic distance between populations. While F_{ST} and Nei's distance have different analytical definitions and theoretical properties, they capture the same basic relationships, and their correlation is 93.9 percent. Therefore, in the rest of this paper we only use F_{ST} measures.

Cavalli-Sforza et al. (1994) provide genetic distance data at the population level, not at the country level. Therefore, economists and other social scientists interested in studying country-level data need to match populations to countries. In Spolaore and Wacziarg (2009), we did so using ethnic composition data by country from Alesina et al. (2003), who list $1,120$ country-ethnic group categories. We matched ethnic group labels with population labels in Appendices 2 and 3 from Cavalli-Sforza et al. (1994). For instance, according to Alesina et al. (2003), India is composed of 72 percent of "Indo-Aryans" and 25 percent "Dravidians." These groups were matched, respectively, to "Indians" and "Dravidhans" (S.E. Indians) from Cavalli-Sforza et al.

(1994). Another example is Italy, where the ethnic groups labelled "Italians" and "Rhaetians" (95.4 percent of Italy's population) in Alesina et al. (2003) were matched to the genetic category "Italian" in Cavalli-Sforza et al. (1994), and the "Sardinians" ethnic group (2.7 percent of Italy's population) was matched to the "Sardinian" genetic group.

Using these matching rules, we constructed two measures of F_{ST} genetic distance between countries.[9] The first was the distance between the plurality ethnic groups of each country in a pair, i.e. the groups with the largest shares of each country's population. For instance, the plurality genetic distance between India and Italy is the genetic distance between the Indian genetic group and the Italian genetic group ($F_{ST} = 0.026$). This resulted in a dataset of $21,321$ pairs of countries (207 underlying countries and dependencies) with available genetic distance data.[10] The second was a measure of weighted genetic distance. Many countries, such as the United States or Australia, are made up of sub-populations that are genetically distant, and for which both genetic distance data and data on the shares of each genetic group are available. Assume that country 1 contains populations $i = 1, ..., I$ and country 2 contains populations $j = 1, ..., J$, denote by s_{1i} the share of population i in country 1 (similarly for country 2) and d_{ij} the genetic distance between populations i and j. The weighted F_{ST} genetic distance between countries 1 and 2 is then:

$$F_{ST}^W = \sum_{i=1}^{I} \sum_{j=1}^{J} (s_{1i} \times s_{2j} \times d_{ij}) \tag{14}$$

The interpretation of this measure is straightforward: it represents the expected genetic distance between two randomly selected individuals, one from each country.[11] Weighted genetic distance

[9] We also constructed genetic distance for populations as they were in 1500, based again on data from Cavalli-Sforza et al. (1994). For this variable, for instance, the United States is matched to the North Amerindian population. This measure of genetic distance in 1500 can either be used as an instrument for contemporary genetic distance (Spolaore and Wacziarg, 2009), or as an independent variable in applications that seek to explain pre-Industrial economic outcomes (Spolaore and Wacziarg, 2013). However we do not make use of this variable in this paper, since we focus on the contemporary relationship between ancestry and culture.

[10] For 27 countries, the data on group shares was missing from Alesina et al.'s (2003) database, but a match to genetic groups based on plurality groups was possible through information from *Encyclopedia Britannica*. Thus, our weighted measure of genetic distance covers $16,110$ pairs, or 180 countries, whereas for the plurality match we have data on $21,321$ pairs from 207 countries.

[11] Therefore, the weighted measure is not to be interpreted as F_{ST} genetic distance between the *whole* population of a country (say, all Australians) and the *whole* population of another country (say, all Americans), as if each country were formed by one randomly-mating population (a *deme*). Instead, to each pair of individuals in each

is very highly correlated with genetic distance based on dominant groups: the correlation is 93 percent. In the rest of this paper we will mostly use weighted F_{ST} distance, which is a more precise measure of expected genetic distance between countries. Table 1 presents summary statistics for F_{ST} and F_{ST}^W.

3   Culture

To capture cultural distance we adopt a three-pronged approach. We first focus on a salient dimension of culture, language, likely to be strongly related with genetic distance because language, like genes, is transmitted from parents to children within populations, and because linguistic differentiation, like genetic differentiation, results over time from horizontal separation between populations. Religion is another salient characteristic of human societies, also transmitted intergenerationally with variations. Finally, in the most novel part of this paper we use answers to the World Values Survey to construct broader metrics of distance in values, norms and attitudes. Jointly, these three classes of measures are referred to as *memetic distance*, by analogy with genetic distance, using a distinction between culturally transmitted traits (*memes*) and genetically transmitted traits (*genes*) that goes back to Dawkins (1976). We describe in turn the methods by which each of these measures were constructed, and provide descriptions of these variables, before turning to their interrelationships.

3.1   Linguistic distance

To capture linguistic distance, we employ two methods, one based on *language trees*, and the other based on *lexicostatistics*. These are arguably the most widely used in the social sciences, but there exist other types of measures of linguistic distance, discussed in Ginsburgh and Weber (2015).

The classification of languages into trees is based on a methodology borrowed from *cladistics*. Linguists group languages into families based on perceived similarities between them.[12] For instance, in one commonly used classification of languages, from Ethnologue, French is classified as

country is assigned their respective ancestrally inherited distance – that is, the distance corresponding to their respective ancestral groups – which may vary across individuals within each country when these countries are formed of different genetic groups.

[12]For a further discussion of linguistic trees, see Desmet, Ortuño-Ortín and Wacziarg (2015) and Ginsburgh and Weber (2015).

"Indo-European - Italic - Romance - Italo-Western - Western - Gallo-Iberian - Gallo-Romance - Gallo-Rhaetian - Oil - Français." Similarly, Italian is classified as "Indo-European - Italic - Romance - Italo-Western - Italo-Dalmatian." This can serve as the basis for characterizing the linguistic distance between French and Italian, because Italian shares 4 nodes with French. Variation in the number of common nodes corresponds to variation in linguistic distance. French and Italian, for instance, share no common nodes with non Indo-European languages, and are therefore at a higher linguistic distance from them than they are with each other.

We use data from Fearon (2003), who assembled data on the prevalence of different languages for a large set of countries in the world from a variety of sources, and used the linguistic trees provided in Ethnologue to capture the distance between these languages. As we did with genetic distance, we compute two different measures: the number of common nodes between the two plurality languages of each country in a pair, CN, and the expected or weighted number of common nodes, CN^W. The latter exploits the fact that countries can be linguistically heterogeneous, and consists of computing the expected number of common linguistic nodes between two randomly chosen individuals, one from each country. More formally, for each country in a pair:

$$CN^W = \sum_{i=1}^{I} \sum_{j=1}^{J} \left(s_{1i} \times s_{2j} \times c_{ij} \right) \tag{15}$$

where s_{ki} is the share of linguistic group i in country k and c_{ij} is the number of common nodes between languages i and j.[13] Both CN and CN^W range from 0 to 15. From the two measures of linguistic proximity, following Fearon (2003) we use the following transformation to obtain corresponding measures of linguistic distance ranging from 0 to 1:

$$TLD = \sqrt{\frac{15 - CN}{15}} \tag{16}$$

Here TLD refers to tree-based linguistic distance and we similarly define the weighted measures TLD^W by replacing CN with CN^W in equation (16). The main advantage of this approach is that distances can be computed for a wide range of countries: we have $12,246$ observations for TLD and TLD^W, from 157 underlying countries (Table 1 provides summary statistics). The drawback of tree-based measures is that linguistic distance is calculated on a discrete number of common nodes, which could be an imperfect measure of separation times between languages. A single split

---

[13] CN^W is, in fact, the B-index of Greenberg (1956), applied here to measuring the expected distance between two individuals chosen from each country in a pair, rather from within a single society. See Ginsburgh and Weber (2015) for further details on Greenberg's B-index.

between two languages that occurred a long time ago would result in the same measure of distance than a more recent single split, but the languages in the first case may in fact be more distant than in the second. Similarly, numerous recent splits may result in two languages sharing few nodes, while a smaller number of very distant linguistic subdivisions could make distant languages seem close. This drawback justifies looking at an alternative measure.

This second measure of linguistic distance is based on lexicostatistics, the branch of quantitative linguistics classifying language groups based on whether words used to convey some common meanings – such as "mother" or "table" – are *cognate*, i.e. stem from the same ancestor word. Two languages with many cognate words are linguistically closer than those with non-cognate words. For instance, the words "tavola" in Italian and "table" in French both stem from the common Latin term "tabula". They are therefore cognate. Replicating this over a large number of meanings, the percentage of cognate words is a measure of linguistic proximity. We rely on data from Dyen et al. (1992), who use 200 underlying meanings. In the same way as before, we compute two measures of the percentage of cognate words: the percentage of cognate words between the plurality languages spoken in each country in a pair, CLD, and the weighted percentage, CLD^W, which represents the expected percentage of cognate words between two individuals randomly chosen from each country in a pair.[14] Once again, Table 1 provides summary statistics, showing that CLD and CLD^W vary between 0 and 0.92, with the sample mean equal to roughly 0.6.

The big advantage of the lexicostatistical approach is that it approximates linguistic differences in a more continuous way than the cladistic approach.[15] The most widely used source of lexicostatistical distance data in the social sciences is Dyen et al. (1992), which we use here. This particular source only covers Indo-European languages, and therefore metrics of linguistic distance are only available for country pairs where these languages are spoken. However, new data from the Automated Similarity Judgment Program (ASJP), combining lexicostatistical methods with

---

[14]In cases of pairs composed of countries, like India, where Indo-European languages are spoken by a plurality, but non-Indo-European languages are spoken by a large minority, CLD may be available but not CLD^W. Indeed we have 63 underlying countries (1, 953 pairs) for CLD^W and 85 countries (3, 570 pairs) for CLD.

[15]Under the assumption that linguistic drift is constant across languages, i.e. that the rate of linguistic innovation over time is similar across languages, lexicostatistical distance can be argued to be correlated with separation times between languages. This insight gave rise to the field of *glottochronology,* the attempt to back out the dates of separation of population based on linguistic similarities between them. The assumption of common linguistic drift has been heavily debated.

measures of phonological, grammatical and lexical similarity between languages, covers a wider set of languages (Ginsburgh and Weber, 2015, provide further references and details on this recent database). While we do not use this data here, its recent development opens new avenues to study the effect of linguistic distance on socioeconomic outcomes since it consists of continuous metrics of linguistic similarity available for a broad set of languages.

The tree-based and cognate-based measures of linguistic distance, in the limited sample of Indo-European speaking countries for which the two sets of measures are available, are relatively highly correlated. The correlation between the two weighted measures is 0.82, while the correlation between the plurality measures is 0.78.

3.2 Religious distance

To capture religious distance between countries, we adopt an approach analogous to the tree-based linguistic distance. We consider trees that describe the relationship between world religions. One such tree is from Mecham, Fearon and Laitin (2006), displayed in Figure 2, and another is from the World Christian Database (2007, henceforth WCD), displayed in Figure 3. We make use of both in the empirical work that follows.

The trees consist of grouping religions into broad categories. For instance, "Near-Eastern Monotheistic Religions" is one broad category common to both trees we use. These broad categories are further divided into finer classifications. For instance Near Eastern monotheistic religions are subdivided into Christianity, Islam and Judaism. These are further refined into yet greater levels of disaggregation. The number of common nodes between religions is a metric of religious proximity. For instance Lutherans are closer in religious space to Baptists than they are to the Greek Orthodox.

In the Mecham, Fearon and Laitin dataset there can be up to 5 common nodes between religions, while the WCD data is less finely disaggregated, so there can be up to 3 common nodes only.[16] Each source provides data on the frequency of each religion in each country, so distances between religions can be mapped to religious distance between countries. As before, we calculate the number of common nodes between the plurality religions of each country in a pair, as well as the expected number of common nodes (following a formula analogous to equation 15). Finally, to

---

[16] Due to its finer level of disaggregation the Fearon, Mecham and Laitin classification and data is preferred. However for the sake of completeness we present results pertaining to both datasets below.

obtain measures of religious distance, we implement a transformation analogous to that in equation (16). Summary statistics for the 4 resulting metrics are displayed in Table 1.

3.3 Cultural distance based on the World Values Survey

Answers to questions from social surveys can be used as indicators of a respondent's cultural norms, values and attitudes. By analogy with genetics, questions correspond to *gene loci* while the specific answers given are the *alleles*. Differences across populations in the answer shares to a specific question can be used to calculate the cultural distance between countries on that specific question. Finally, aggregating over questions allows the computation of indices of cultural distance in values, norms and attitudes space.

There are three major challenges when computing these indices. The first challenge is the choice of questions. Rather than choosing questions arbitrarily, which would be open to criticism, we consider the set of *all* values-related questions appearing in the World Values Survey 1981-2010 Integrated Questionnaire, i.e. those listed by the WVS as categories A through G.[17] All 740 questions can be considered when computing distances question by question. When calculating summary indices of cultural distance that aggregate across questions, however, it is important to have a sample that is balanced across country pairs, i.e. to have the same number of questions for each pair. Some of the questions were only asked in a subset of countries, sometimes a small subset. There is a trade-off between maintaining a large set of questions, in which case the number of country pairs shrinks, or maintaining a broad sample of country pairs, in which case the set of questions is reduced. In what follows we chose to do the latter, to maximize the representativeness of the sample of countries. This led to keeping 98 questions out of the original set. Data availability is the only concern that governs which questions remain. Yet since the remaining questions are those that were asked in the broadest set of countries, they constitute the core questions of the WVS. Focusing on these questions, that were asked in at least one wave of the WVS in 74 countries, we are left with distances computed for $2,701$ pairs.[18]

---

[17]These categories are as follows: Category A: Perceptions of Life. Category B: Environment. Category C: Work. Category D: Family. Category E: Politics and Society. Category F: Religion and Morale. Category G: National Identity. Additional categories, S, X and Y are not considered here since they relate either to the demographic characteristics of the respondent or characteristics of the survey (wave, year, etc).

[18]We also implemented different choices in terms of the mix of country pairs and questions, increasing the number of questions at the cost of losing some country pairs, leading to no appreciable change in the results. These results

The second challenge is the choice of a functional form for computing distances for each question. There are many possible choices, but we focus on the simplest one, which is to calculate the Euclidian distance. In further empirical work that is available in the online appendix, we used Manhattan and F_{ST} cultural distances instead of Euclidian distance, finding results that are very similar to those reported here.[19] Consider countries 1 and 2 and question i from the WVS, which admits answers $j = 1, ..., J$. Some questions are binary ($J = 2$) and others admit more than two answers ($J > 2$).[20] Let s_{ij}^c denote the share of respondents in country $c \in \{1, 2\}$ giving answer j to question i. Then for binary questions, cultural distance CD_i^{12} between countries 1 and 2 is simply:

$$CD_i^{12} = \left| s_{i1}^1 - s_{i1}^2 \right| \tag{17}$$

while for non-binary questions:

$$CD_i^{12} = \sqrt{\sum_{j=1}^{J} \left(s_{ij}^1 - s_{ij}^2 \right)^2} \tag{18}$$

The third challenge is to aggregate question-specific distances in order to obtain summary measures of cultural distance. To create summary indices we first standardize the question-specific distances to have a mean of zero and a standard error equal to one. This ensures equal weighing of questions in every summary index. We next simply sum the question specific indices, to compute several indices of cultural distance. We first sum across all 98 questions, to obtain an overall index. Next, we sum question-specific distances for each of the six categories of questions, as specified by the WVS.[21] Finally, we created an index for the whole group of binary questions and another one for the whole group of non-binary questions.

---

are available in an online appendix at:

http://www.anderson.ucla.edu/faculty_pages/romain.wacziarg/downloads/ancestry_appendix.pdf

[19] Desmet et al. (2011) used the F_{ST} functional form to calculate cultural distance based on answers from the World Values Survey for a sample of European countries, and explored the relationship between the resulting matrix of cultural distance and F_{ST} genetic distance in Europe, finding like we do a strong association between the two.

[20] We call these non-binary. Non-binary questions are further divided into those that admit an ordering on a scale (e.g. happiness on a scale from 1 to 10), and those that do not (e.g. do you prefer option 1, option 2 or option 3). This distinction is not relevant here. For an in-depth discussion of question types in the WVS, see Desmet, Ortuño-Ortín and Wacziarg (2014).

[21] Category B, questions relating to the environment, is dropped as no question from this category was asked in all of the **74** countries. Category A features **32** questions, category C features **14** questions, category D features **7** questions, category E features **30** questions, category F features **12** questions and category G features **3** questions. There were **35** binary questions, and **63** non-binary questions.

Summary statistics for these 9 indices appear at the bottom of Table 1. By construction each index has mean zero, and is available for all 2, 701 pairs.[22]

4   Ancestry and culture: A simple conceptual framework

As we discussed in Section 2, genetic distance measures relatedness between populations and is roughly proportional to time since two populations shared the same ancestors, that is, since they were the same population. Over time, ancestors transmit a large number of traits to their descendants, not only biologically (through DNA), but also culturally. This transmission takes place with variation and change over time. Therefore, on average, populations that are more closely related will have had less time to diverge from each other on a large set of culturally transmitted traits, such as language, religion, traditions, habits, and values. This process establishes a close connection between ancestry, measured by genetic distance, and culturally transmitted traits: genetic distance and memetic distance should be positively correlated. A stylized formal model, adapted from Spolaore and Wacziarg (2009, 2012), can illustrate this relationship in a simplified and concise way.

For simplicity, we consider three populations, $i = 1, 2, 3$, living at the present time. Population 1 and population 2 descend from the same last common ancestor population, which lived one period ago. In contrast, population 3 only shares common ancestors with populations 1 and 2 going further in time, back to two periods ago. That is, population 3 is less closely related with populations 1 and 2 than these are with each other. Using the analogy discussed in Section 2, we can say that populations 1 and 2 are like siblings, while population 3 is a more distant cousin. The phylogenetic tree of the three populations is illustrated in Figure 4.

Building on the results described in Section 2.2.2, we can approximate the genetic distance $d_g(i, j)$ between population i and population j as the time since they were one population. Therefore, genetic distance $d_g(1, 2)$ between population 1 and population 2 is smaller than genetic distance between population 1 and population 3, and also smaller than genetic distance between population 2 and population 3:

$$d_g(1, 2) = F < d_g(1, 3) = d_g(2, 3) = F' \tag{19}$$

---

[22]However one country, Puerto Rico, drops out in our regression analysis due to missing data on genetic distance, leaving us with 2, 628 observations in the regressions of Section 5.

How far in cultural traits are these populations from each other? While, in general, their cultural divergence may depend on complex processes of transmission across generations of a large set of cultural traits, it is useful to focus on the simplest possible mechanism of cultural transmission with variation, whereby culture is captured by just one trait (or *meme*), which we assume can be represented as a point on the real line. In each period t, a population i has cultural traits $c_i(t)$ that are inherited with variation from its ancestor population, which had traits $c_i(t-1)$, according to:

$$c_i(t) = c_i(t-1) + \varepsilon_i(t) \tag{20}$$

where $\varepsilon_i(t)$ measures random variation between time $t-1$ and time t. We assume the simplest possible mechanism for variation: cultural change as a random walk. That is, $\varepsilon_i(t) = \varepsilon > 0$ with probability $1/2$ and $\varepsilon_i(t) = -\varepsilon < 0$ with probability $1/2$. In addition, we assume that such shocks are independent across different populations ($\varepsilon_i(t)$ is independent of $\varepsilon_j(t)$ for $j \neq i$).

Let $d_c(i,j) \equiv |c_j - c_i|$ denote the distance in cultural traits between population i and population j. The expected memetic distance between population 1 and population 2, which share their last common ancestors only one period ago, is denoted by $E[d_c(1,2)]$, and given by:[23]

$$E[d_c(1,2)] = \varepsilon \tag{21}$$

All variation between populations 1 and 2 is given by cultural change that took place between one period ago and now. In contrast, expected memetic distance between population 1 and population 3, and between population 2 and population 3, comes from shocks that took place both between one period ago and now, and between two periods ago and one period ago. On average, such shocks are associated with a larger distance in culturally transmitted traits:[24]

$$E[d_c(1,3)] = E[d_c(2,3)] = \frac{3\varepsilon}{2} > \varepsilon = E[d_c(1,2)] \tag{22}$$

Therefore, on average a larger genetic distance is associated with greater distance in cultural traits. This relation is not deterministic. Some pairs of populations that are more distant cousins may end up with more similar cultural traits than two more closely related populations, but that outcome is less likely to be observed than the opposite. Therefore, genetic distance and distance in culturally transmitted traits, such as language, religion, and values, are expected to be positively correlated.

---

[23] For the derivation, see the Appendix.

[24] The derivation of this result, and of a generalization, are provided in the Appendix.

5 Ancestry and culture: Empirical evidence

In this section we conduct an empirical exploration of the relationship between genetic distance and our various measures of memetic distance, to test the hypothesis that longer separation times are in fact positively related with differences in language, religion and norms, values and attitudes.

5.1 Genetic distance and linguistic distance

Measures of linguistic and genetic distances should be positively correlated. Cavalli-Sforza et al. (1994, pp. 98-105) observed that there is usually little genetic admixture between linguistic groups. Languages are generally transmitted intergenerationally. Thus, phylogenetic trees and linguistic trees tend to resemble each other. At the same time, we should not expect a perfect relationship, for several reasons. Firstly, as already mentioned, linguistic data based on trees feature a discrete number of nodes, whereas genetic distance based on a large number of alleles, as we use, is a continuous measure of separation times. Second, the functional forms for measures of genetic distance (F_{ST}) and linguistic distance (a nonlinear transformation of the number of different nodes, or the percentage of non-cognate words, depending on the measure) are different. Third, successful groups conquering the territories of distinct linguistic groups can impose their language without necessarily imposing their genes. Such was the case, for instance, with the Magyar conquest of Hungary: the resulting language was of the Uralic family, but the Magyar genetic admixture was so limited that the Hungarians are genetically very close to other Slavic populations, such as the Poles. An even more stark example comes from the population movements that followed the discovery of the New World, in particular the slave trade: the current descendants of former slaves do not speak the original West African languages of their ancestors. Similarly, current inhabitants of the United States predominantly speak English, whereas their ancestors came from a diverse set of linguistic groups. Thus, modern migrations served to break the link between genetic and linguistic distance.

Table 2, Panel A explores the basic correlations. We find that our various measures of linguistic distance are highly correlated among themselves. For instance, the correlation between weighted TLD and weighted CLD is 0.82. Weighted TLD is also positively correlated with weighted genetic distance, with a correlation equal to 0.22. However, CLD is not strongly correlated with genetic distance, in all likelihood because the sample is limited to Indo-European speaking countries, which tend to also be genetically close: there is not enough variation in the data to detect a significant correlation.

19

Table 3 presents regressions of our various measures of linguistic distance on genetic distance, with or without controls for a wide range of measures of geographic separation - including geodesic distance, the absolute difference in longitudes and latitudes, etc. Indeed one concern is that genetic distance merely reflects geographic proximity, and that genetic and linguistic distance are positively correlated simply because the relationship goes through geographic distance. We find that this is not the case, and that genetic distance is significantly related to tree-based measures of linguistic distance (TLD). In fact, the standardized beta coefficient on F_{ST} genetic distance, reported in the last line of Table 2, suggests that a one standard deviation increase in genetic distance is associated with a $0.15 - 0.22$ standard deviation increase in linguistic distance, depending on the measure and specification.[25] For the cognate-based measures (CLD), the relationship is negative, but not robustly significant statistically.

As mentioned already, the population movements that followed the discovery of the New World were important factors breaking the link between genetic and linguistic distance. To investigate this issue, Table 4 isolates the sample consisting of Old World countries. This excludes any country pair containing a country from the Americas or Oceania. We find much larger correlations than in Table 3. For instance, the standardized betas on weighted TLD now range between 0.29 and 0.41. Moreover, the correlations between genetic distance and CLD turn positive, and significant in the univariate cases. These results show there exists a strong correlation between genetic and linguistic distance for country pairs least likely to have experienced language replacement over the course of the last 500 years.

5.2 Genetic distance and religious distance

Like language, religious beliefs tend to be transmitted intergenerationally, leading us to expect a positive correlation between religious distance and genetic distance. However, several factors may limit the extent to which religious distance correlates with genealogical distance. First, while they may find their sources in ancient religious beliefs, several major world religions appeared relatively recently. For instance, one of the oldest monotheistic religions, Judaism, appeared only $3,500$ years ago. Second, in line with the first observation, the rate of drift of religious beliefs is likely to be much faster than that of genes, so that populations that are genetically similar often espouse different

---

[25] It is well-known that the standardized beta is equal to the correlation coefficient for the univariate case. This can be verified by comparing the standardized betas in columns (1) and (3) to the corresponding ones in Table 2 Panel A.

religious beliefs. In fact, two recent religious innovations, Christianity and Islam, occurred among closely related populations in the Middle East. Third, religious beliefs are transmitted horizontally through conquests and conversions, perhaps to a faster extent than even languages, as it is easier to change one's religion than one's language. Thus, the emergence and horizontal diffusion of new religions is likely to weaken the link between religious distance and genetic distance perhaps to a greater extent than for linguistic distance. Fourth, the aforementioned functional form differences between metrics of linguistic and genetic distance apply with the same force to measures of religious distance.

Despite these caveats, we do find that religious distance is positively correlated with genetic distance. The first piece of evidence in presented in Panel B of Table 2. There we see, for instance, that weighted religious distance based on the Mecham, Fearon and Laitin religious tree ($F - RD$) bears a 0.18 correlation with weighted genetic distance. Correlations are smaller using measures based on the World Christian Database tree ($WCD - RD$), which are less finely disaggregated. We also find substantial positive correlations among our various measures of religious distance, but these correlations are not sufficiently high to justify looking at only one measure.

Tables 5 and 6 present regression evidence, again with or without controls for geographic distance for each of the 4 measures of religious distance. In all but one of the specifications, genetic distance comes out with a positive statistically significant coefficient. The standardized magnitude of the effect of genetic distance is generally smaller than for linguistic distance, in line with the observation above. Yet, in particular for $F - RD$, we find standardized effects comprised between 8.3 percent and 18.1 percent, again consistent with our model of cultural drift. Moreover, unlike for language, we do not find a particular tendency for the effect to be more pronounced among Old World countries (Table 6).

5.3 Genetic distance and cultural distance

Our final exploration concerns the relationship between genetic distance and distance in norms, values and attitudes. We start with an analysis of the relationship between genetic distance and question-specific distances, for all available questions from the WVS. Under the null hypothesis of no relationship between genetic and cultural distances we would expect 5 percent of the correlations to be significant (2.5 percent positive and significant), and the distribution of correlations to be centered around zero. Figure 5 presents a histogram of sample correlations between bilateral

distance for each question, and weighted genetic distance, for the full set of 740 questions.[26] The mode of the distribution is well to the right of zero, with a mean of about 10 percent. 71.6 percent of the correlations were positive. In 53.1 percent of the cases the correlation with genetic distance is both positive and significant, far in excess of what we would expect under the null. A substantial subset of the questions feature correlations that are quite large - for 22.4 percent of the questions, the correlations are in excess of 0.20 and statistically significant at the 5 percent level.[27]

These simple correlations could confound the effects of geographic distance with those of genetic distance. To address this issue, we ran regressions, for each question, of WVS distance on genetic and geodesic distance. Figure 6 presents a histogram of the standardized beta coefficient on genetic distance, representing the effect of a one standard deviation change in genetic distance as a share of a standard deviation in the dependent variable. 66.9 percent of the standardized betas are positive, and 47.2 percent are both positive and significant at the 5 percent level. We also find a number of large effects, with 20 percent of the standardized betas greater than 0.20.[28] Controlling for geodesic distance does not modify the conclusion reached earlier.

While these results are informative, they conflate questions on very different subjects, and of different types (binary versus non-binary). So we now turn to the relationship between our 9 indices of cultural distance, and genetic distance. The analysis is now limited to the 98 questions available for 74 countries. Table 7 presents simple correlations. Genetic distance bears a correlation of 0.27 with our summary measure of cultural distance. The last line of the table shows that genetic distance is positively and significantly correlated with 8 of our 9 measures of cultural distance based on the WVS. The only category for which this is not the case is category D, pertaining to questions about family. Among the other categories, the correlation with genetic distance varies between 7.4 percent (questions on work) and 29.9 percent (questions on politics and society).

In Tables 8 and 9 we turn to regression analysis, following the same format as earlier: for each index we present a univariate regression and one that controls for geographic barriers. Table 8 focuses on the aggregate index covering all 98 questions, and then the indices for binary and

---

[26] The underlying sample varies across questions, which could introduce some bias. However, the results are no different when we focus on the set of 98 questions for which we have a balanced sample of 2,701 country pairs.

[27] For the restricted set of 98 questions covering a balanced set of countries, 63.3 percent of the correlations were positive and significant, and 75.5 percent of them were positive.

[28] For the restricted set of 98 questions, 67.3 percent of the standardized beta coefficients on genetic distance were positive, and 53.1 percent of the effects were positive and significant at the 5 percent level.

non-binary questions. We find a large, statistically significant positive relationship between genetic distance and cultural distance. In the specification with controls (column 2), the standardized effect of genetic distance is 25.5 percent. While the effect remains positive and significant for both binary and non-binary questions, it is largest for the latter - with a standardized effect of 30.2 percent. Interestingly, the inclusion of geographic distance controls serves to weaken the effect of genetic distance only a little bit.

Table 9, finally, breaks things down by question category. In the bottom panel, with geographic controls, we see positive and significant effects of genetic distance on cultural distance for all but category D (Family). The largest effects, quantitatively, are for categories A (Perceptions of Life), E (Politics and Society) and F (Religion and Morale). Future work should seek to delve more deeply into the characteristics of questions most closely associated with ancestral distance.

6   Conclusion

What does genetic distance measure? In this paper we argued that genetic distance is a summary statistic for differences in a wide range of intergenerationally transmitted human traits. We focused on language, religion and values, finding empirical evidence of a positive correlation between genetic distance and linguistic, religious and cultural distances. It is important to note that genetic distance is not strongly correlated with only a small and specific subset of differences in cultural traits. On the contrary, genetic distance tends to be broadly and significantly correlated with a vast range of differences in cultural traits. Thus, while specific correlations with individual sets of traits are typically moderate in magnitude, there is an overall relation between ancestry and culture, consistent with a conceptual framework in which a broad range of cultural traits are transmitted with variation across generations over time. Genetic distance is a useful summary statistic capturing differences in this broad range of cultural traits.

Future research should seek to improve on this evidence. Recent progress in the measurement of linguistic differences, using systematic quantitative methods, will allow for a more precise evaluation of the effects of linguistic distance on political economy outcomes. Similarly, improvements in the gathering of genetic data should lead to improvements in our ability to detect effects of ancestral distance on cultural distance and in turn on political economy outcomes. As more genetic data on more finely defined populations become available, more granular analyses of the relationship between genetic and cultural distance will become possible. Third, alternative data sets on values,

norms and attitudes also exist, either regionally or worldwide, and could be used to complement our analysis.

Research seeking to quantify human barriers to socioeconomic interactions across populations is in its infancy. With this paper we have sought to clarify what observable traits are captured by ancestral distance, but much remains to be done.

References

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

Ashraf, Q., and O. Galor (2013b) 'Genetic Diversity and the Origins of Cultural Fragmentation', *American Economic Review Papers and Proceedings*, 103(3), 528–33.

Alesina, A., A. Devleeschauwer, W. Easterly, S. Kurlat, and R. Wacziarg (2003) 'Fractionalization', *Journal of Economic Growth*, 8(2), 55–194.

Arbatli, E., Q. Ashraf and O. Galor (2013) 'The Nature of Civil Conflict', *working paper*, Brown University.

Bai, Y. B. and J. K. Kung (2011) 'Genetic Distance and Income Difference: Evidence from Changes in China's Cross-Strait Relations.' *Economics Letters,* 110(3), 255–258.

Beauchamp, J. P., D. Cesarini, M. Johannesson, M. J. H. M. van der Loos, P. D. Koellinger, P. J. F. Groenen, J. H. Fowler, J. N. Rosenquist, A. R. Thurik, and N. A. Christakis (2011) 'Molecular Genetics and Economics', *Journal of Economic Perspectives*, 25(4), 57–82.

Bell, A. V., P. J. Richerson, and R. McElreath (2009) 'Culture Rather than Genes Provides Greater Scope for the Evolution of Large-Scale Human Prosociality', *Proceedings of the National Academy of Sciences of the United States*, October, 106(42), 17671–17674.

Benjamin, D. J., D. Cesarini, M. J. H. M. van der Loos, C. T. Dawes, P. D. Koellinger, P. K. E. Magnusson, C. F. Chabris, D. Conley, D. Laibson, M. Johannesson and P. M. Vissche (2012) 'The Genetic Architecture of Economic and Political Preferences', *Proceedings of National Academy of Sciences of the United States*, 109(21) May 22, 8026–8031.

Bisin, A. and T. Verdier (2000) 'Beyond The Melting Pot: Cultural Transmission, Marriage, and the Evolution of Ethnic And Religious Traits', *Quarterly Journal of Economics,* 115, 955–988.

Bisin, A. and T. Verdier (2001) 'The Economics of Cultural Transmission and the Evolution of Preferences', *Journal of Economic Theory,* 97(2), 298–319.

Bisin, A. and T. Verdier (2010) 'The Economics of Cultural Transmission and Socialization', *NBER Working Paper* #16512.

Bowles, S. and H. Gintis (2011) *A Cooperative Species: Human Reciprocity and Its Evolution* (Princeton: Princeton University Press).

Boyd, R. and P. J. Richerson (1985) *Culture and the Evolutionary Process* (Chicago: University of Chicago Press).

Cavalli-Sforza, L. L. and A. W. F. Edwards (1964) Analysis of Human Evolution, *Proc. 11th International Congress Genetics,* 2, 923–933.

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

Cavalli-Sforza, L. L., P. Menozzi and A. Piazza (1994) *The History and Geography of Human Genes* (Princeton: Princeton University Press).

Dawkins, R. (1976) *The Selfish Gene* (Oxford: Oxford University Press).

Desmet, K., I. Ortuño-Ortín and R. Wacziarg (2014) 'Culture, Ethnicity and Diversity', *working paper*, UCLA, May.

Desmet, K., I. Ortuño-Ortín and R. Wacziarg (2015) 'Linguistic Cleavages and Economic Development', forthcoming in Victor Ginsburgh and Shlomo Weber, eds., *The Palgrave Handbook of Economics and Language,* London: Palgrave Macmillan.

Desmet, K., M. Le Breton, I. Ortuño-Ortín and S. Weber (2011) 'The Stability and Breakup of Nations: A Quantitative Analysis', *Journal of Economic Growth*, 16(3), 183–213

Dyen, I., J. B. Kruskal, and P. Black (1992) 'An Indoeuropean Classification: A Lexicostatistical Experiment', *Transactions of the American Philosophical Society*, 82, 1–132.

Egger, P. and F. Toubal (2015) 'Common Spoken Language and International Trade', forthcoming in Victor Ginsburgh and Shlomo Weber, eds., *The Palgrave Handbook of Economics and Language,* London: Palgrave Macmillan.

Falconer, D. S. and T. F.C. Mackay (1996) *Introduction to Quantitative Genetics* (Harlow: Pearson Education Limited).

Fearon, J. (2003) 'Ethnic and Cultural Diversity by Country', *Journal of Economic Growth*, 8(2), 195–222.

Ginsburgh, V. and S. Weber (2011) *How Many Languages Do We Need? The Economics of Linguistic Diversity.* Princeton: Princeton University Press

Ginsburgh, V. and S. Weber (2015) 'Linguistic Distances and Their Use in Economics', forthcoming

in Victor Ginsburgh and Shlomo Weber, eds., *The Palgrave Handbook of Economics and Language,* London: Palgrave Macmillan.

Gorodnichenko, Y. and G. Roland (2011) 'Culture, Institutions and the Wealth of Nations', *working paper*, UC Berkeley.

Greenberg, J. (1956) 'The measurement of linguistic diversity', *Language*, 32, 109–115.

Guiso, L., P. Sapienza and L. Zingales (2009) 'Cultural Biases in Economic Exchange', *Quarterly Journal of Economics*, 124(3), 1095–1131.

Hamilton, M. B. (2009) *Population Genetics* (Chichester: Wiley-Blackwell).

Mecham, R. Q., J. Fearon, and D. Laitin (2006) 'Religious Classification and Data on Shares of Major World Religions', *unpublished*, Stanford University.

Melitz, J. and F. Toubal (2012) 'Native Language, Spoken Language, Translation and Trade', *CEPR Discussion Paper* No. DP8994.

Melitz, J. (2008) 'Language and Foreign Trade', *European Economic Review,* 52(4), 667–699.

Richerson, P. J., and R. Boyd (2004) *Not By Genes Alone: How Culture Transformed Human Evolution* (Chicago: University of Chicago Press).

Rosenberg, N. A., Pritchard J. K., Weber J. L., Cann H. M., Kidd K. K., Zhivotovsky L. A., Feldman M. W. (2002) 'Genetic Structure of Human Populations', *Science*, December 20; 298 (5602), 2381–2385.

Seabright, P. (2010) *The Company of Strangers: A Natural History of Economic Life* (Princeton: Princeton University Press).

Seldin, M. F., R. Shigeta, P. Villoslada, C. Selmi, J. Tuomilehto, G. Silva, J. W. Belmont, L. Klareskog, and P. K. Gregersen (2006) 'European Population Substructure: Clustering of Northern and Southern Populations', *PLoS Genet.* September, 2(9).

Spolaore, E. and R. Wacziarg (2009) 'The Diffusion of Development', *Quarterly Journal of Economics*, 124(2), 469–529.

Spolaore, E. and R. Wacziarg (2012) 'Long-Term Barriers to the International Diffusion of Innovations', in Jeffrey Frankel and Christopher Pissarides, eds., *NBER International Seminar on Macroeconomics 2011*, (Chicago: University of Chicago Press) Chapter 1, 11–46.

Spolaore, E. and R. Wacziarg (2013) 'How Deep Are the Roots of Economic Development?' *Journal of Economic Literature*, 51(2), 325–369.

Spolaore, E. and R. Wacziarg (2014) 'Fertility and Modernity', *working paper*, UCLA and Tufts University.

Spolaore, E. and R. Wacziarg (2015) 'War and Relatedness', *working paper*, UCLA and Tufts University.

Ralph, P. and G. Coop (2013) 'The Geography of Recent Genetic Ancestry Across Europe', *PLOS Biology*, May 7, 11(5).

Tian, C., R. Kosoy, R. Nassir, A. Lee, P. Villoslada, L. Klareskog, L. Hammarström, H.-J. Garchon, A. E. Pulver, M. Ransom, P. K. Gregersen, and M. F. Seldin (2009) 'European Population Genetic Substructure: Further Definition of Ancestry Informative Markers for Distinguishing among Diverse European Ethnic Groups', *Molecular Medicine*, November, 15(11), 371–383.

World Christian Database (2007) http://www.worldchristiandatabase.org/wcd/

Wright, S. (1951) 'The Genetical Structure of Populations', *Annals of Eugenics,* 15, 323–354.

Appendix: Derivations of the results in Section 4

First, we show that $E[d_c(1,2)] = \varepsilon$. The result is immediate. With probability $1/4$ both populations experience a positive shock ε, and with probability $1/4$ both populations experience a negative shock $-\varepsilon$. Hence, with probability $1/2$, their vertical distance is zero. With probability $1/2$ one population experiences a positive shock ε and the other a negative shock $-\varepsilon$, implying a cultural distance equal to $|\varepsilon - (-\varepsilon)| = 2\varepsilon$. On average, the expected cultural distance is

$$E[d_c(1,2)] = \frac{1}{2}0 + \frac{1}{2}2\varepsilon = \varepsilon \tag{23}$$

Second, we show that

$$E[d_c(1,3)] = E[d_c(2,3)] = \frac{3\varepsilon}{2} > \varepsilon = E[d_c(1,2)] \tag{24}$$

In fact, this is a special instance of the more general case in which the shock between today and a period ago is given by $\varepsilon > 0$ with probability $1/2$ and $-\varepsilon$ with probability $1/2$, while the shock to cultural traits between two periods ago and one period ago is $\varepsilon' > 0$ with probability $1/2$ and ε' with probability $1/2$. In Section 4, we show the result for the special case $\varepsilon = \varepsilon'$.

In general, with probability $1/4$ population 1's ancestor populations and population 2's ancestor populations experienced identical shocks both between two periods ago and one period ago, and between one period ago and now. That is, with probability $1/4$ we have $d_v(1,3) = 0$. By the same token, with probability $1/4$ the two populations experienced identical shocks between two periods ago and one period ago, but different shocks between one period ago and now, implying $d_c(1,3) = 2\varepsilon$, and with probability $1/4$ identical shocks between one period ago and now, but different shocks between two periods ago and one period ago, implying $d_c(1,3) = 2\varepsilon'$. With probability $1/8$, one population lineage has experienced two positive shocks $(\varepsilon' + \varepsilon)$ while the other has experienced two negative shocks $(-\varepsilon' - \varepsilon)$, therefore leading to a distance equal to $2\varepsilon' + 2\varepsilon$. Finally, with probability $1/8$ one population lineage has experienced a positive shock ε' and a negative shock $-\varepsilon$ while the other population lineage has experienced $-\varepsilon'$ and ε. In this latest case, we have $d_c(1,3) = |2\varepsilon - 2\varepsilon'|$. In sum, expected cultural distance is given by

$$E[d_c(1,3)] = \frac{1}{4}2\varepsilon' + \frac{1}{4}2\varepsilon + \frac{1}{8}(2\varepsilon' + 2\varepsilon) + \frac{1}{8}|2\varepsilon - 2\varepsilon'|$$

which is equal to $\varepsilon + \frac{\varepsilon'}{2}$ if $\varepsilon \geq \varepsilon'$ and equal to $\varepsilon' + \frac{\varepsilon}{2}$ if $\varepsilon \leq \varepsilon'$, or, equivalently

$$E[d_c(1,3)] = \max\{\frac{\varepsilon'}{2} + \varepsilon, \varepsilon' + \frac{\varepsilon}{2}\} \tag{25}$$

The same expected cultural distance holds between populations 2 and 3.

In the special case $\varepsilon = \varepsilon'$, condition (25) simplifies to $E[d_c(1,3)] = E[d_c(2,3)] = 3\varepsilon/2$, the result shown in Section 4.

Table 1. Simple Summary Statistics for Measures of Ancestral, Linguistics and Cultural Distances

Variable	# pairs	# countries	Mean	Std. Dev.	Min	Max
			Genetic Distance			
F_{ST} distance, weighted	16,110	180	0.115	0.070	0.000	0.355
F_{ST} distance, plurality	21,321	207	0.117	0.081	0.000	0.338
			Linguistic Distance			
Tree-based linguistic distance, weighted	12,246	157	0.970	0.100	0.000	1.000
Tree-based linguistic distance, plurality	12,246	157	0.965	0.137	0.000	1.000
Cognate-based linguistic distance, weighted	1,953	63	0.599	0.268	0.000	0.918
Cognate-based linguistic distance, plurality	3,570	85	0.636	0.284	0.000	0.920
			Religious Distance			
Religious distance, weighted, Fearon	12,246	157	0.853	0.144	0.089	1.000
Religious distance, plurality, Fearon	12,246	157	0.786	0.313	0.000	1.000
Religious distance, weighted, WCD	19,306	197	0.747	0.163	0.127	0.997
Religious distance, plurality, WCD	19,306	197	0.628	0.358	0.000	1.000
		Cultural Distance based on the World Values Survey				
Overall cultural distance measure	2,701	74	0.000	33.015	-89.818	118.294
Category A – Perceptions of Life	2,701	74	0.000	12.110	-35.146	46.542
Category C – Work	2,701	74	0.000	7.460	-14.102	37.550
Category D – Family	2,701	74	0.000	3.436	-7.749	18.479
Category E – Politics and Society	2,701	74	0.000	11.872	-28.001	45.561
Category F – Religion and Morale	2,701	74	0.000	6.962	-13.140	26.975
Category G – National Identity	2,701	74	0.000	1.967	-4.544	7.345
All binary questions	2,701	74	0.000	14.049	-36.103	57.345
All non-binary questions	2,701	74	0.000	21.911	-58.630	62.481

Table 2 - Simple Correlations between Linguistic, Religious and Genetic Distances

Panel A - Linguistic Distance

	Weighted TLD	Plurality TLD	Weighted CLD	Plurality CLD	Weighted F$_{ST}$
Plurality TLD	0.926* (12,246)	1			
Weighted CLD	0.817* (1,035)	0.798* (1,035)	1 (1,953)		
Plurality CLD	0.740* (2,145)	0.776* (2,145)	0.979* (1,953)	1 (3570)	
Weighted F$_{ST}$	0.220* (11,476)	0.195* (11,476)	-0.034 (1,378)	-0.058* (2,701)	1 (16,110)
Plurality F$_{ST}$	0.232* (12,246)	0.210* (12,246)	0.011 (1,953)	-0.031 (3,570)	0.939* (16,110)

(Number of observations in parentheses; * denotes 5% significance)

Panel B - Religious Distance

	Weighted F-RD	Plurality F-RD	Weighted WCD-RD	Plurality WCD-RD	Weighted F$_{ST}$
Plurality F-RD	0.839* (12,246)	1			
Weighted WCD-RD	0.784* (11,325)	0.622* (11,325)	1 (19,306)		
Plurality WCD-RD	0.698* (11,325)	0.640* (11,325)	0.819* (19,306)	1 (19,306)	
Weighted F$_{ST}$	0.181* (11,476)	0.121* (11,476)	0.091* (15,400)	0.064* (15,400)	1 (16,110)
Plurality F$_{ST}$	0.168* (12,246)	0.114* (12,246)	0.056* (19,306)	0.034* (19,306)	0.939* (16,110)

(Number of observations in parentheses; * denotes 5% significance)

Table 3 - Linguistic Distance and Contemporary Genetic Distance, Full Sample
(OLS Regressions, dependent variable displayed in the second row)

	(1) Weighed TLD	(2) Weighed TLD	(3) Plurality TLD	(4) Plurality TLD	(5) Weighted CLD	(6) Weighted CLD	(7) Plurality CLD	(8) Plurality CLD
Weighted Fst	**0.320**	**0.242**	**0.385**	**0.284**	**-0.173**	**-0.260**	**-0.325**	**-0.643**
Genetic Distance	**(24.12)*****	**(16.54)*****	**(21.24)*****	**(14.14)*****	**(1.27)**	**(1.65)***	**(3.03)*****	**(5.73)*****
Geodesic Distance		-0.001		-0.002		-0.018		-0.010
(1000s of km)		(2.03)**		(1.73)*		(2.56)**		(2.72)***
Absolute difference		0.030		0.037		0.214		0.197
in longitudes		(5.09)***		(4.59)***		(3.87)***		(6.70)***
Absolute difference		0.001		0.000		-0.012		-0.059
in latitudes		(0.23)		(0.03)		(0.19)		(1.68)*
1 for contiguity		-0.087		-0.122		-0.131		-0.215
		(14.15)***		(14.44)***		(3.14)***		(7.10)***
=1 if either country		0.011		0.013		0.047		0.046
is landlocked		(5.60)***		(4.88)***		(2.54)**		(3.94)***
=1 if either country		0.010		0.010		0.052		0.071
is an island		(4.12)***		(3.06)***		(3.08)***		(6.04)***
=1 if pair shares at		-0.059		-0.071		-0.176		-0.160
least one sea or ocean		(18.07)***		(15.85)***		(9.61)***		(10.59)***
Constant	0.935	0.937	0.922	0.925	0.624	0.646	0.672	0.648
	(536.89)***	(371.07)***	(388.29)***	(267.22)***	(48.97)***	(34.36)***	(68.14)***	(45.15)***
Adjusted R²	0.05	0.12	0.04	0.10	0.00	0.14	0.00	0.17
N	11,476	11,476	11,476	11,476	1,378	1,378	2,701	2,701
Standardized beta	0.220	0.166	0.195	0.143	-0.034	-0.051	-0.058	-0.115

(t-statistics in parentheses; * significant at 10%; ** significant at 5%; *** significant at 1%)
Key: TLD = Tree-based linguistic distance. CLD = %Cognate-based linguistic distance (Indo-European languages only).

Table 4 - Linguistic Distance and Contemporary Genetic Distance, Old World
(OLS Regressions, dependent variable displayed in the second row)

	(1) Weighed TLD	(2) Weighed TLD	(3) Plurality TLD	(4) Plurality TLD	(5) Weighted CLD	(6) Weighted CLD	(7) Plurality CLD	(8) Plurality CLD
Weighted Fst	**0.293**	**0.255**	**0.344**	**0.300**	**0.771**	**0.399**	**1.334**	**0.423**
Genetic Distance	**(39.25)*****	**(24.25)*****	**(30.44)*****	**(18.51)*****	**(2.59)****	**(1.05)**	**(5.39)*****	**(1.48)**
Geodesic Distance		-0.004		-0.006		-0.011		0.036
(1000s of km)		(7.23)***		(6.41)***		(0.45)		(5.09)***
Absolute difference		0.037		0.050		0.232		-0.031
in longitudes		(9.26)***		(8.12)***		(1.40)		(0.73)
Absolute difference		0.038		0.049		0.241		-0.278
in latitudes		(7.73)***		(6.56)***		(1.02)		(3.27)***
1 for contiguity		-0.073		-0.097		-0.173		-0.203
		(22.77)***		(19.77)***		(4.74)***		(7.27)***
=1 if either country		-0.005		-0.003		0.006		-0.012
is landlocked		(4.77)***		(1.85)*		(0.30)		(0.87)
=1 if either country		0.004		0.004		-0.116		-0.009
is an island		(2.68)***		(1.78)*		(3.24)***		(0.44)
=1 if pair shares at		-0.021		-0.020		-0.035		-0.051
least one sea or ocean		(9.95)***		(6.14)***		(1.21)		(2.15)**
Constant	0.948	0.955	0.941	0.948	0.681	0.667	0.694	0.687
	(957.57)***	(650.87)***	(628.54)***	(419.09)***	(52.95)***	(32.85)***	(65.49)***	(45.66)***
Adjusted R²	0.17	0.26	0.11	0.18	0.02	0.16	0.03	0.20
N	7,626	7,626	7,626	7,626	351	351	780	780
Standardized beta	0.410	0.357	0.329	0.288	0.137	0.071	0.190	0.060

(t-statistics in parentheses; * significant at 10%; ** significant at 5%; *** significant at 1%)
Key: TLD = Tree-based linguistic distance. CLD = %Cognate-based linguistic distance (Indo-European languages only).

Table 5 - Religious Distance and Genetic Distance, Full Sample
(OLS Regressions, dependent variable displayed in the second row)

	(1) Weighed FRD	(2) Weighed FRD	(3) Plurality FRD	(4) Plurality FRD	(5) Weighted WCD-RD	(6) Weighted WCD-RD	(7) Plurality WCD-RD	(8) Plurality WCD-RD
Weighted Fst	**0.376**	**0.307**	**0.542**	**0.373**	**0.208**	**0.085**	**0.329**	**0.010**
Genetic Distance	**(19.70)*****	**(14.38)*****	**(13.10)*****	**(8.01)*****	**(11.29)*****	**(4.14)*****	**(7.95)*****	**(0.21)**
Geodesic Distance		-0.002		-0.003		-0.002		0.008
(1000s of km)		(1.85)*		(1.24)		(2.43)**		(3.48)***
Absolute difference in longitudes		0.042		0.072		0.077		0.033
		(4.85)***		(3.85)***		(9.33)***		(1.76)*
Absolute difference in latitudes		-0.024		0.009		-0.023		-0.100
		(2.54)**		(0.44)		(2.75)***		(5.19)***
1 for contiguity		-0.081		-0.208		-0.095		-0.203
		(8.87)***		(10.41)***		(9.60)***		(9.03)***
=1 if either country is landlocked		0.002		-0.025		0.014		-0.017
		(0.83)		(3.96)***		(5.00)***		(2.67)***
=1 if either country is an island		0.038		0.075		-0.016		0.016
		(10.56)***		(9.52)***		(5.43)***		(2.42)**
=1 if pair shares at least one sea or ocean		-0.069		-0.120		-0.052		-0.102
		(14.37)***		(11.53)***		(11.93)***		(10.44)***
Constant	0.809	0.813	0.725	0.732	0.735	0.730	0.604	0.601
	(318.81)***	(217.33)***	(131.71)***	(89.83)***	(294.46)***	(195.12)***	(107.71)***	(70.93)***
Adjusted R^2	0.03	0.09	0.01	0.06	0.01	0.07	0.00	0.04
Number of pairs	11,476	11,476	11,476	11,476	15,400	15,225	15,400	15,225
Standardized effect	0.181	0.148	0.121	0.083	0.091	0.037	0.064	0.002

(t-statistics in parentheses; * significant at 10%; ** significant at 5%; *** significant at 1%)
Key: FRD = Fearon, Meecham and Laitin religious distance. WCD-RD = World Christian Database religious distance.

Table 6 - Religious Distance and Genetic Distance, Old World
(OLS Regressions, dependent variable displayed in the second row)

	(1) Weighed FRD	(2) Weighed FRD	(3) Plurality FRD	(4) Plurality FRD	(5) Weighted WCD-RD	(6) Weighted WCD-RD	(7) Plurality WCD-RD	(8) Plurality WCD-RD
Weighted Fst	**0.363**							
	(19.29)*							
Genetic Distance		**0.058**	**0.525**	**0.005**	**0.360**	**-0.018**	**0.494**	**-0.293**
		(2.13)	**(11.93)***	**(0.08)**	**(16.44)***	**(0.59)**	**(9.76)***	**(4.02)***
Geodesic Distance (1000s of km)		-0.003 (2.14)**		-0.008 (2.41)**		-0.003 (1.71)*		0.006 (1.59)
Absolute difference in longitudes		0.112 (10.81)***		0.180 (7.34)***		0.197 (16.63)***		0.219 (7.75)***
Absolute difference in latitudes		0.135 (10.96)***		0.289 (9.90)***		0.104 (7.52)***		0.190 (5.78)***
1 for contiguity		-0.052 (6.25)***		-0.184 (9.24)***		-0.066 (6.46)***		-0.171 (7.05)***
=1 if either country is landlocked		0.000 (0.04)		-0.031 (4.37)***		0.022 (6.39)***		0.001 (0.12)
=1 if either country is an island		0.020 (4.92)***		0.058 (5.98)***		0.002 (0.62)		0.060 (6.18)***
=1 if pair shares at least one sea or ocean		-0.031 (5.50)***		-0.044 (3.37)***		0.008 (1.21)		0.015 (0.98)
Constant	0.830 (326.57)***	0.800 (210.25)***	0.758 (127.48)***	0.726 (80.43)***	0.743 (251.17)***	0.681 (155.53)***	0.620 (90.64)***	0.522 (49.88)***
Adjusted R²	0.05	0.13	0.02	0.08	0.03	0.16	0.01	0.08
Number of pairs	7,626	7,626	7,626	7,626	8,778	8,646	8,778	8,646
Standardized beta	0.216	0.034	0.135	0.001	0.173	-0.009	0.104	-0.061

(t-statistics in parentheses; * significant at 10%; ** significant at 5%; *** significant at 1%)
Key: FRD = Fearon, Meecham and Laitin religious distance. WCD-RD = World Christian Database religious distance.

Table 7 - Correlations between Genetic Distance and Cultural Distance measures

	CD, all questions	CD, cat. A	CD, cat. C	CD, cat. D	CD, cat. E	CD, cat. F	CD, cat. G	CD, binary questions	CD, non-binary questions
CD, category A	0.879*	1							
CD, category C	0.644*	0.481*	1						
CD, category D	0.619*	0.528*	0.373*	1					
CD, category E	0.802*	0.598*	0.284*	0.392*	1				
CD, category F	0.728*	0.544*	0.394*	0.398*	0.476*	1			
CD, category G	0.431*	0.315*	0.306*	0.203*	0.298*	0.275*	1		
CD, binary questions	0.870*	0.837*	0.816*	0.584*	0.515*	0.544*	0.305*	1	
CD, non-binary questions	0.949*	0.787*	0.448*	0.559*	0.878*	0.749*	0.453*	0.670*	1
Weighted F_{ST} gen. dist.	0.269*	0.245*	0.074*	0.028	0.299*	0.216*	0.130*	0.147*	0.312*

(Correlations based on 2,628 observations; * denotes 5% significance).
Key for WVS question categories: Category A: Perceptions of Life. Category C: Work. Category D: Family. Category E: Politics and Society.
Category F: Religion and Morale. Category G: National Identity.

36

Table 8 - Cultural Distance and Genetic Distance
(Dependent variable: Cultural Distance, based on Euclidian distance metric)

	(1) Total	(2) Total	(3) Binary	(4) Binary	(5) Non-binary	(6) Non-binary
F_{ST} genetic distance, weighted, current match	143.891 (14.33)***	138.547 (12.24)***	33.291 (7.60)***	29.873 (6.02)***	110.600 (16.82)***	108.675 (14.65)***
Geodesic Distance, 1000s of km		-1.088 (2.18)**		-0.334 (1.52)		-0.754 (2.30)**
Absolute difference in longitudes		0.072 (2.00)**		0.019 (1.21)		0.053 (2.24)**
Absolute difference in latitudes		0.108 (2.16)**		0.037 (1.67)*		0.071 (2.17)**
1 for contiguity		-34.347 (9.92)***		-13.900 (9.16)***		-20.447 (9.00)***
Number of landlocked countries in the pair		-8.833 (7.35)***		-4.027 (7.65)***		-4.806 (6.10)***
Number of island countries in the pair		-3.325 (2.12)**		-2.704 (3.93)***		-0.622 (0.60)
1 if pair shares at least one sea or ocean		-12.034 (5.23)***		-2.999 (2.98)***		-9.034 (5.99)***
Constant	-10.570 (10.88)***	-3.688 (2.50)**	-2.368 (5.58)***	0.702 (1.09)	-8.202 (12.88)***	-4.390 (4.54)***
Adjusted R^2	0.07	0.14	0.02	0.08	0.10	0.16
# observations	2,628	2,513	2,628	2,513	2,628	2,513
Standardized beta	0.269	0.255	0.147	0.130	0.312	0.302

(t-statistics in parentheses; * significant at 10%; ** significant at 5%; ** significant at 1%)
The sample includes 73 countries (2,628 country pairs) and the total index of cultural distance is based on 98 WVS questions.

Table 9 - Cultural Distance and Genetic Distance, by question category

	(1) Category A	(2) Category C	(3) Category D	(4) Category E	(5) Category F	(6) Category G
			Univariate Specification			
F_{ST} genetic distance, weighted	47.613 (12.92)***	8.947 (3.82)***	1.555 (1.43)	57.536 (16.06)***	24.133 (11.32)***	4.106 (6.71)***
R-squared	0.06	0.01	0.00	0.09	0.05	0.02
Standardized beta	0.244	0.074	0.028	0.299	0.216	0.130
			Multivariate specification [a]			
F_{ST} genetic distance, Weighted	47.591 (11.55)***	7.010 (2.62)***	1.468 (1.20)	55.279 (13.59)***	24.836 (10.09)***	2.363 (3.39)***
Adjusted R-squared	0.14	0.05	0.08	0.14	0.08	0.04
Standardized beta	0.241	0.058	0.026	0.284	0.218	0.075

(t-statistics in parentheses; * significant at 10%; ** significant at 5%; *** significant at 1%)

The univariate specification is based on 2,628 observations (country pairs). The multivariate specification is based on 2,513 observations.

All specifications include an intercept.

Key for WVS question categories: Category A: Perceptions of Life. Category C: Work. Category D: Family. Category E: Politics and Society. Category F: Religion and Morale. Category G: National Identity.

[a]: The multivariate specification includes the following geographic controls: Geodesic distance, absolute difference in longitudes, absolute difference in latitudes, dummy for contiguity, number of landlocked countries in the pair, number of island countries in the pair, dummy=1 if pair shares at least one sea or ocean.

Figure 1 - Genetic distance among 42 populations.
Source: Cavalli-Sforza et al., 1994.
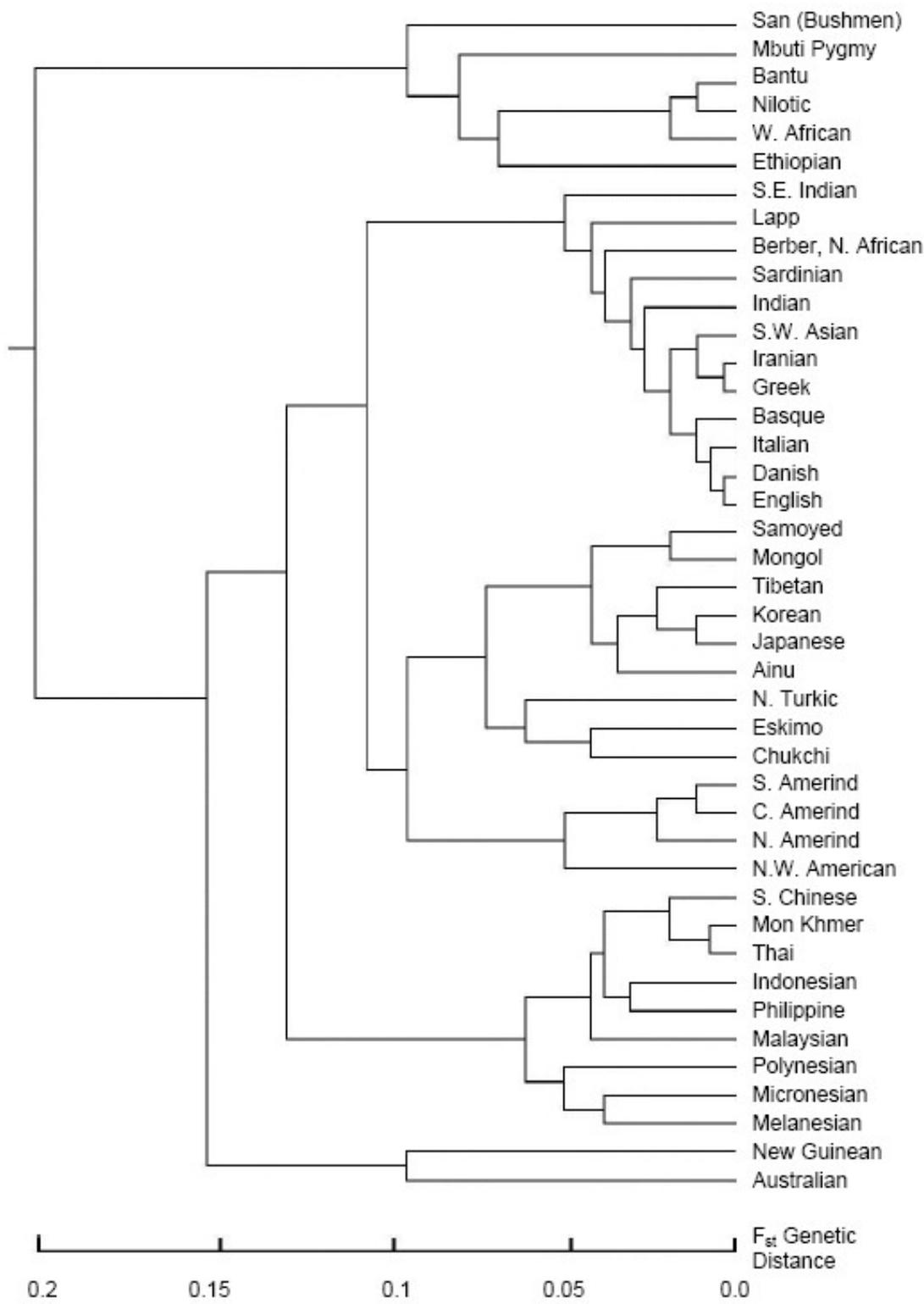
Figure 2 – Mecham, Fearon and Laitin Religious Tree

1.0   Asia-born Religion
       1.1 South Asian Religions
             1.11   Hinduism
       1.2 Far Eastern Religions
             1.21   Taoism
             1.22   Buddhism
                    1.221   Therevada
                    1.222   Cao Dai
                    1.223   Hoa Hao
2.0   Near Eastern Monotheistic Religion
       2.1 Christianity
             2.11 Western Catholicism
                    2.111 Roman Catholic
                    2.112 Protestant
                            2.1121 Anglican
                            2.1122 Lutheran
                            2.1123 Presbyterian
                            2.1124 Methodist
                            2.1125 Baptist
                            2.1126 Calvinist
                            2.1127 Kimbanguist
                            2.1128 Church of Ireland
             2.12 Eastern Orthodox
                    2.121 Greek Orthodox
                    2.122 Russian Orthodox
                            2.1221 Old Believers
                    2.123 Ukranian Orthodox
                            2.1231 Russian Patriarchy
                            2.1232 Kiev Patriarchy
                    2.124 Albanian Orthodox
                    2.125 Armenian Orthodox
                    2.126 Bulgarian Orthodox
                    2.127 Georgian Orthodox
                    2.128 Macedonian Orthodox
                    2.129 Romanian Orthodox
       2.2 Islam
             2.21 Sunni Islam
                    2.211 Shaf'i Sunni
             2.22 Shi'I Islam
                    2.221 Ibadi Shi'i
                    2.222 Alevi Shi'i
                    2.223 Zaydi Shi'i
             2.23 Druze
       2.3 Judaism
3.0   Traditional
4.0   Other
5.0   Assorted
6.0   None

Figure 3 - World Christian Database Religious Tree

1.0   Asia-born Religion
       1.1 South Asian Religions
              1.11   Jains
              1.12   Hindus
              1.13   Sikhs
              1.14   Zoroastrians
       1.2 Far Eastern Religions
              1.21   Confucianists
              1.22   Shintoists
              1.23   Taoists
              1.24   Buddhists
              1.25   Chinese Universists
2.0   Near Eastern Monotheistic Religion
       2.1 Christians
              2.11   Anglicans
              2.12   Independents
              2.13   Marginals
              2.14   Orthodox
              2.15   Protestants
              2.16   Roman Catholics
              2.17   Disaffiliated/Unaffiliated Christians
              2.18   Doubly-Affiliated Christians
       2.2 Muslims
       2.3 Jews
3.0   Ethnoreligionists
4.0   Spiritists
5.0   Bha'is
5.0   Doubly Professing
6.0   Other Religionists
7.0   Nonreligionists/Atheist
       7.1 Nonreligionists
       7.2 Atheist

Figure 4 - Population Tree
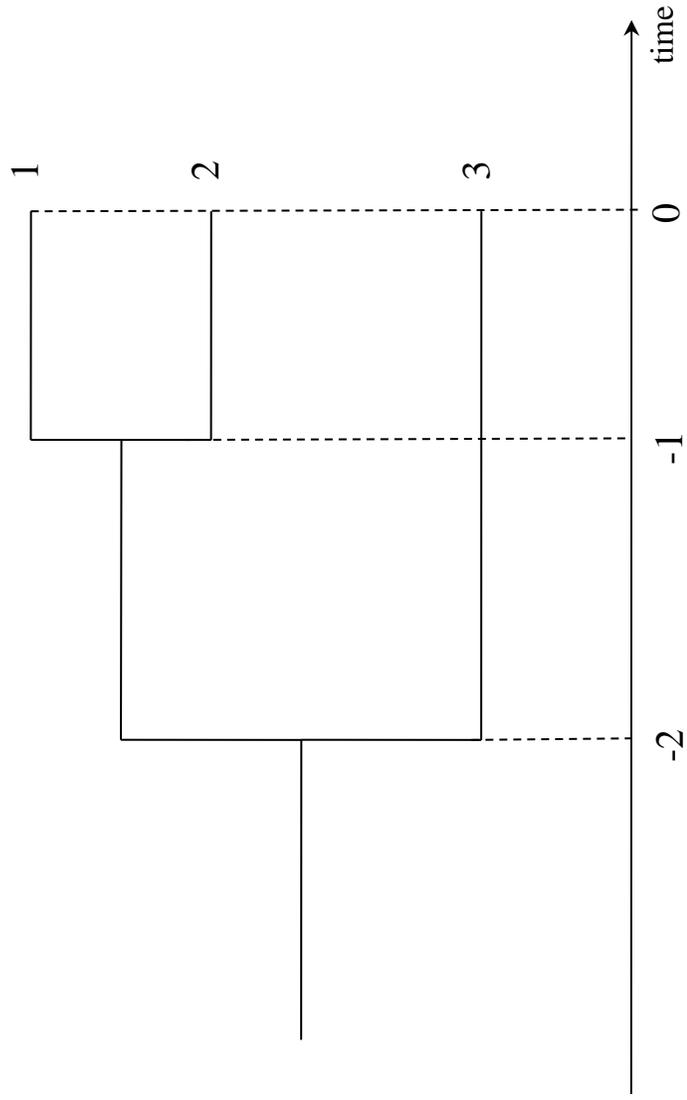
Figure 5 - Distribution of Correlations between Cultural and Genetic Distances (740 questions)
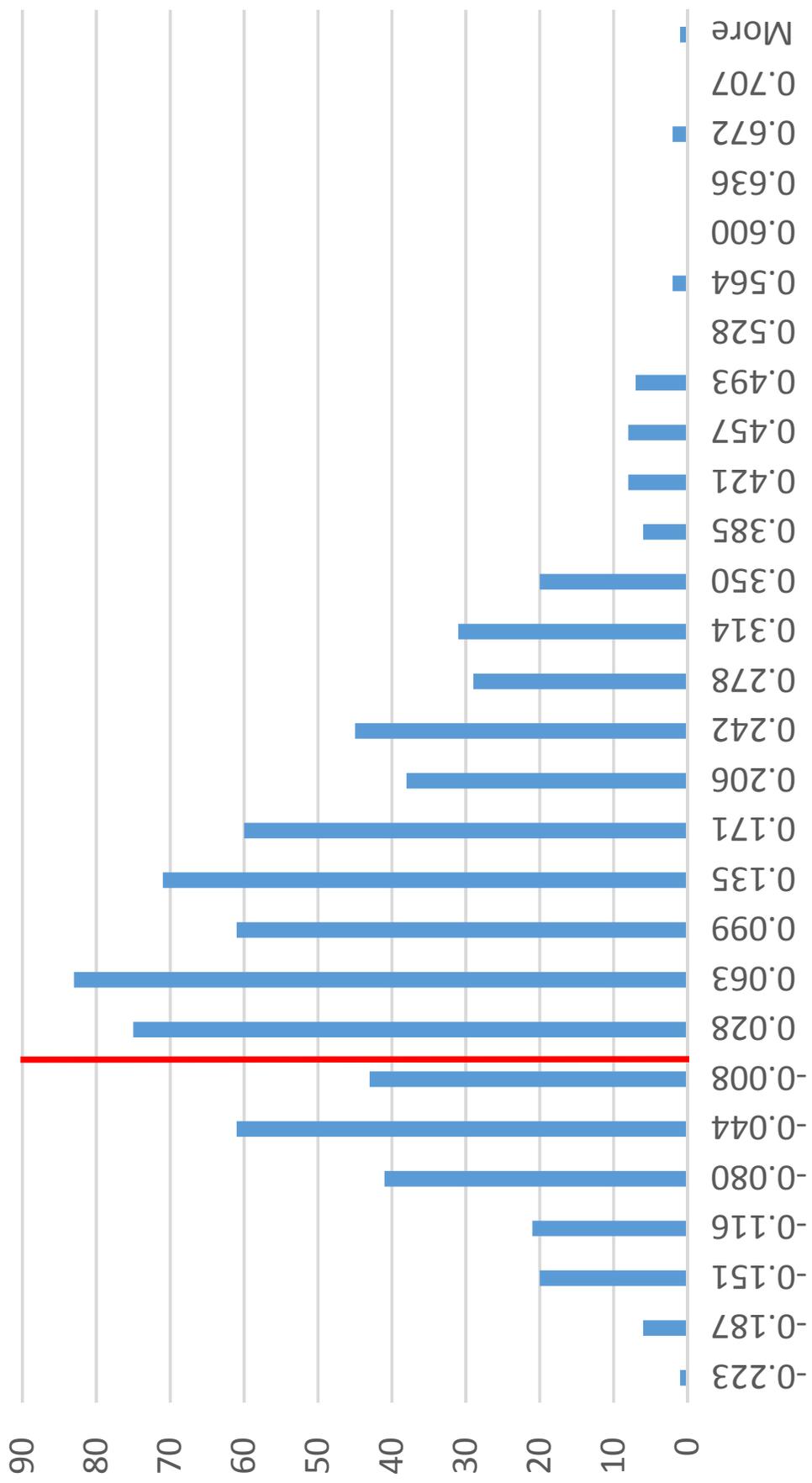
Figure 6 - Distribution of Beta Coefficients on Genetic Distance (740 questions, with geodesic distance control)