The Diffusion of Development

Enrico Spolaore
Tufts University,
NBER and CESIfo

Romain Wacziarg[*]
Stanford University,
NBER and CEPR

March 2008

Abstract

We find that genetic distance, a measure associated with the amount of time elapsed since two populations' last common ancestors, has a statistically and economically significant effect on income differences, even when controlling for measures of geographical and climatic differences, transportation costs, linguistic distances, and other measures of cultural and historical differences. We provide an economic interpretation of these findings in terms of barriers to the diffusion of development from the world technological frontier, implying that income differences should be a function of relative genetic distance from the frontier. The empirical evidence strongly supports this barriers interpretation.

[*] Spolaore: Department of Economics, Tufts University, Medford, MA 02155-6722, enrico.spolaore@tufts.edu. Wacziarg: Graduate School of Business, Stanford University, Stanford CA 94305-5015, wacziarg@gsb.stanford.edu. We are grateful to Alberto Alesina, Robert Barro, Sam Bowles, Robert Boyd, Francesco Caselli, Dan Cox, Steven Durlauf, Jim Fearon, Oded Galor, Luigi Guiso, Peter Howitt, Yannis Ioannides, Larry Katz, Pete Klenow, Andros Kourtellos, Ed Kutsoati, Edward Leamer, John Londregan, Peter Lorentzen, Lisa Lynch, Paolo Mauro, Deborah Menegotto, Sharun Mukand, Louis Putterman, Gérard Roland, Fabio Schiantarelli, Antonio Spilimbergo, Susan Stokes, Chih Ming Tan, David Weil, Bruce Weinberg, Ivo Welch, three anonymous referees, and participants at numerous seminars and conferences for helpful comments. We also thank Jim Fearon for providing some data.

1 Introduction

What explains the vast differences in income per capita that are observed across countries? This paper provides new empirical evidence sheding light on this question.[1] At the center of our analysis is genetic distance, a measure based on aggregate differences in the distribution of gene variants across populations.[2] For the first time, we document and discuss the relationship between genetic distance and differences in income per capita across countries. We find that measures of genetic distance bear a statistically and economically significant relationship with income differences, and that this relationship is robust to controlling for a large number of measures of geographical and climatic differences, transportation costs, linguistic distance, and other measures of cultural and historical differences. The effect of genetic distance holds not only for contemporary income differences, but also for income differences measured since 1500. Moreover, the effect of genetic distance on income differences is quantitatively large, statistically significant and robust not only worldwide, but also within Europe, for which more precise measures of cross-country genetic distance are available.

In addition to establishing these facts, we provide an economic interpretation for our findings. What does genetic distance capture, and why is it correlated with income differences, even when controlling for geographical distance and other factors? Technically, genetic distance measures the difference in gene distributions between two populations, where the genes under considerations are *neutral*, i.e. they change randomly and independently of selection pressure. The rationale for this approach is that divergence in neutral genes provides information about lines of descent. Most random genetic change takes place regularly over time, as in a molecular clock. Therefore, genetic distance measures the time since two populations have shared common ancestors - i.e., the time since they have been the same population. In other words, genetic distance is a summary measure of general relateness between populations. An intuitive analogue is given by the familiar concept of relatedness between individuals: two siblings are more closely related than two cousins because they share more recent common ancestors - their parents rather than their grandparents.

[1] Contributions to the literature on the determinants of income per capita using cross-country regressions include Hall and Jones (1999), Acemoglu et al. (2001), Easterly and Levine (2003), Alcalá and Ciccone (2004), Glaeser et al. (2004), among many others.

[2] Our main source for genetic distances between human populations is Cavalli-Sforza, Menozzi and Piazza (1994). Recent textbook references on human evolution are Boyd and Silk (2003) and Jobling, Hurles and Tyler-Smith (2004). For a nontechnical discussion of these concepts see Dawkins (2004).

Since it is based on neutral genetic change, genetic distance is not meant to capture differences in specific genetic traits that directly matter for survival and fitness. Hence, our results provide *no* evidence for a direct effect of specific genes on income or productivity. That is, our findings are not about some societies having some specific genes that make them directly richer. Instead, our results provide strong evidence that a general measure of genealogical relatedness between populations can explain income differences today, even though it reflects mostly neutral genetic variation. Why? Our interpretation is that genetic distance captures barriers to the diffusion of development. More closely related societies are more likely to learn from each other and adopt each other's innovations. It is easier for someone to learn from a sibling than from a cousin, and easier to learn from a cousin than from a stranger. Populations that share more recent common ancestors have had less time to diverge in a wide range of traits and characteristics that are transmitted across generations with variation. Of course, in human populations many of those traits are transmitted across generations culturally rather than biologically.[3] Similarity in such traits would tend to facilitate communication and understanding, and hence the diffusion and adaptation of complex technological and institutional innovations.

What traits are captured by genetic distance? We argue that, by its very definition, genetic distance is an excellent *summary statistic* capturing divergence in the whole set of implicit beliefs, customs, habits, biases, conventions, etc. that are transmitted across generations - biologically and/or culturally - with high persistence. In a nutshell, human genetic distance can be viewed as a summary measure of very long-term divergence in intergenerationally-transmitted traits across populations. Our key hypothesis is that such long-term (and mainly random) divergence has created barriers to the diffusion of technological and institutional innovations across societies in more recent times. While we provide a general economic interpretation of genetic distance in terms of barriers to the diffusion of development from the frontier, we remain largely agnostic about specific mechanisms of technology diffusion, as well as about the specific traits and characteristics that create the barriers.

If our interpretation is correct, the relevant measure of genetic distance associated with economic distance between two societies should not be the absolute genetic distance between them, but their

[3] Classic references on cultural transmission and evolution are Cavalli-Sforza and Feldman (1981) and Boyd and Richerson (1985). See also Richerson and Boyd (2005). Recent economic analyses of cultural transmission across generations include Bisin and Verdier (2000, 2001) and others. This issue will be discussed in more detail in Section 2 of this paper.

relative distance from the world technological frontier. In our empirical analysis we test this central implication using Britain (in the 19th century) and the United States (in the 20th century) as the world technological frontiers. Consistent with our hypothesis, we find that the effect of relative genetic distance on economic distance is positive, and larger than the effect of absolute genetic distance, itself only an imperfect proxy for relative genetic distance to the frontier. We view this as important evidence in support of a barriers effect. The historical data also suggests that the effect of genetic distance on income differences, while always positive and significant since 1500, increased considerably between 1820 and 1870, consistently with a salient role for relative genetic distance during the gradual spread of the Industrial Revolution.[4] More broadly, our interpretation is consistent with the diffusion of economic development as emerging from the formation of a human web, gradually joined by different cultures and societies in function of their relative distance from the technological and institutional frontier.[5]

This paper is part of a small but growing set of contributions that use human genetic distance in empirical economic analyses.[6] Guiso et al. (2004) and Giuliano et al. (2006) employ genetic distance to study bilateral trade flows between European countries. Guiso et al. (2004) use genetic distance as an instrument for trust in trade gravity regressions.[7] Giuliano et al. (2006) study directly the effect of genetic distance on bilateral trade flows in gravity regressions, and argue that this effect is sensitive to controlling for measures of geographic isolation (in contrast, in our analysis the effect of genetic distance on income differences is robust to geographical controls, including transportation costs). Desmet et al. (2006) document the close relationship between genetic distance and cultural heterogeneity, and argue that genetic distance can be used to study nation formation in Europe. Their empirical analysis supports our interpretation of genetic distance as a broad measure of differences in intergenerationally-transmitted characteristics, including cultural

[4] However, it should be added that any changes in the size of the effect over past centuries can only be assessed with much caution, given the relatively small number of data points.

[5] For an excellent historical overview of the formation of the complex web of exchanges and interactions across human communities going back to the Neolithic period see McNeill and McNeill (2004).

[6] There also exists a different economic literature that uses genetic distances between species to evaluate biodiversity (for example, Weizman, 1992, and Brock and Xepapadeas, 2003) .

[7] A negative correlation between genetic distance and trust is consistent with our hypothesis that populations which are genetically more distant face higher barriers between them.

values. These contributions differ from our paper in several respects: we study income differences across countries, while these other authors focus on different dependent variables (bilateral trade flows or the formation of political borders). Moreover, all the other studies only use genetic distance between European populations, while we are the first economists to use worldwide measures of genetic distance in addition to European data.[8]

In Section 2 we present a simple framework in which genetic distance captures divergence in characteristics that are transmitted across generations within populations over the long run, and those differences act as barriers to the diffusion of development from the world technological frontier. The section also contains a general taxonomy of the mechanisms linking genetic distance and economic outcomes. In Section 3, we discuss our data on genetic distance and what it measures. Section 4 presents our empirical findings. Section 5 concludes.

2 A Conceptual Framework

In this section we present an analytic framework linking genetic distance, intergenerationally-transmitted traits, and the diffusion of economic development from the technological frontier. Our analysis leads to a testable prediction: income differences across societies should depend on their *relative* genetic distance from the technological frontier. In Section 4, we show that the empirical evidence strongly supports this prediction.

The main building block of our model is that genetic distance between populations captures the degree of genealogical relatedness of different populations over time. Thus, it can be interpreted as a general metric for average differences in characteristics transmitted across generations. In this paper we will call *vertically transmitted characteristics* (or vertical characteristics) the set of characteristics passed on across generations within a population over the very long run - i.e., over the time horizon along which populations have diverged.[9]

[8] A recent contribution by Fearon (2006) comments on a previous version of our work, and discusses the relationship between genetic distance and linguistic distance in a set of European countries. We are grateful to Jim Fearon for providing data on linguistic distance, which we use in the current version. We find that the effect of genetic distance on income differences is robust to controlling for all such measures of linguistic distance. This topic is covered in Section 4.

[9] This terminology is borrowed from the evolutionary literature on cultural transmission (Cavalli-Sforza and Feldman, 1981; Boyd and Richerson, 1985; Shennan, 2002; Richerson and Boyd, 2004). In our definition, such vertical transmission takes place across generations within a given population, and "oblique" transmission from other

This leads to our second main idea: distance in vertical characteristics act as barriers to the diffusion of productivity-enhancing innovations across societies.[10] We argue that populations that share a more recent common history, and are therefore closer in terms of vertical characteristics, face lower costs and obstacles to adopting each other's innovations.[11] We are interested primarily in the diffusion of economic development in historical times, and especially after the Industrial Revolution. Thus, in our empirical analysis we focus on differences in vertical characteristics as barriers to the diffusion of development from the modern technological frontier.[12]

2.1 A Simple Model

A stylized model illustrates our ideas in the simplest possible way.[13] Consider three periods (o for "origin," p for "prehistory," and h for "history"). In period o there exists only one population (population 0). In period p the original population splits in two populations (1 and 2). In period h each of the two populations split in two separate populations again (population 1 into 1.1 and 1.2, and population 2 into 2.1 and 2.2), as displayed in Figure 1. In this setting the genetic distance $d_g(i, j)$ between population i and population j can be simply measured by the number of periods

genetically-related people within the group is also part of our definition. Hence "vertical" in our context is not limited to parent-to-child transmission *stricto sensu*, but has a broader meaning, which is slightly different from the use of the term often found in the anthropological literature and related fields. We thank Robert Boyd for pointing out this distinction to us.

[10] Policy-induced barriers to the diffusion of technology are analyzed by Parente and Prescott (1994, 2002). In our framework we interpret barriers more broadly to include all long-term societal differences that are obstacles to the diffusion of development.

[11] The idea that differences in long-term societal and cultural characteristics may act as barriers to the diffusion of development is stressed in a large literature on the diffusion of innovations, including the classic book by Rogers (1962 and following editions). For an historical comparative analysis of managerial innovations and performances see Clark and Wolcott (1999).

[12] World technological leadership since the British Industrial Revolution (1700s) has been predominantly associated with Britain and, by the late 1800s, the United States (Brezis, Krugman, and Tsiddon, 1993). In the years right before the Industrial Revolution, the technological frontier was probably the Netherlands. According to Maddison (2003), in previous times the regions with the highest levels of income per capita were Italy (around 1500) and China (around 1000). We return to these issues in Section 4.

[13] In a previous version of this paper (available upon request), we presented a dynamic micro-founded extension of this model, built on Barro and Sala-i-Martin (1997). In that extension, imitation costs are a function of distance in vertical characteristics from the technological frontier, and hence income differences are a function of relative genetic distance in steady state.

since they were one population:

$$d_g(1.1, 1.2) = d_g(2.1, 2.2) = 1 \tag{1}$$

and:

$$d_g(1.1\ ,\ 2.1) = d_g(1.1\ ,\ 2.2) = d_g(1.2\ ,\ 2.1) = d_g(1.2\ ,\ 2.2) = 2 \tag{2}$$

For simplicity, all vertical characteristics of a population are summarized as a point on the real line (i.e., a population i has vertical characteristics v_i, where v_i is a real number). Populations inherit characteristics from their ancestor populations with variations. In general, a population i' descending from a population i will have characteristics:

$$v_{i'} = v_i + \eta_{i'} \tag{3}$$

Consider the simplest possible mechanism for variation: vertical change as a random walk - for every population i', $\eta_{i'}$ takes value $\eta > 0$ with probability 1/2 and $-\eta$ with probability 1/2.[14] Consequently, the distance in vertical characteristics between two populations $d_v(i,j) \equiv |v_j - v_i|$ is, on average, increasing in their genetic distance $d_g(i,j)$.[15] This captures our first main idea.

Our second main idea can also be captured within this simplified setting. Assume that in periods o and p all populations produce output using the basic technology $Y = A_0 L$, so that all populations have the same income per capita $y_0 = A_0$. In period h a population happens to find a more productive technology $A_1 = A_0 + \Delta$ where $\Delta > 0$.[16] Denote this population, the technological frontier, as f.[17] We assume that populations farther from population f in terms of

[14] This simplification is consistent with the molecular-clock interpretation of genetic distance itself. While more complex processes could be considered, this formalization has two advantages: it is economical ("Occam's razor"), and illustrates how neutral random changes are sufficient to generate our theoretical predictions.

[15] Specifically, in period h the expected difference in vertical characteristics between populations at a genetic distance equal to 2 and populations at a genetic distance equal to 1 is given by: $E\{d_v(i,j) \mid d_g(i,j) = 2\} - E\{d_v(i,j) \mid d_g(i,j) = 1\} = \frac{\eta}{2} > 0$. Of course, this is not a deterministic relationship. Some pairs of populations that are genealogically more distant may end up with more similar vertical characteristics than two more closely related populations, but that outcome is less likely to be observed than the opposite. On average, genetic distance and distance in vertical characteristics go hand in hand.

[16] We abstract from the possibility that the likelihood of finding the innovation may itself be a function of a society's vertical characteristics. Such direct effects of vertical characteristics would strengthen the links between genetic distance and economic outcomes, but are not necessary for our results.

[17] The model can be viewed as a very reduced form of a dynamic process in which the frontier economy produces

vertical characteristics face higher barriers to adopt the new technology. To fix ideas, assume that a society i at a vertical distance from the frontier equal to $d_v(i,f)$ can improve its technology only by:

$$\Delta_i = [1 - \beta d_v(i,f)]\Delta \tag{4}$$

where the parameter $\beta > 0$ captures the barriers to the horizontal diffusion of innovations due to distance in vertical characteristics.[18] Hence, income per capita in society i is given by:

$$y_i = A_0 + [1 - \beta d_v(i,f)]\Delta \tag{5}$$

This implies that the economic distance between population i and population j, measured by their income difference $d_e(i,j) \equiv |y_i - y_j|$, is a function of their *relative* vertical distance from the frontier $|d_v(i,f) - d_v(j,f)|$:

$$d_e(i,j) \equiv |y_j - y_i| = \beta\Delta|d_v(i,f) - d_v(j,f)| \tag{6}$$

As we have shown, vertical difference $d_v(i,j)$ and genetic distance $d_g(i,j)$ are positively correlated. Therefore, on average, income differences across societies are increasing in their relative genetic distance from the frontier society. Formally:

$$\begin{aligned} E\{d_e(i,j)||d_g(i,f) - d_g(j,f)| &= 2\} - \\ E\{d_e(i,j)||d_g(i,f) - d_g(j,f)| &= 1\} = \frac{\eta\beta\Delta}{3} > 0 \end{aligned} \tag{7}$$

This result is intuitive. As we increase relative genetic distance from the frontier, the expected income gap increases. The size of the effect is a positive function of the extent of divergence in vertically transmitted characteristics (η), the extent to which this divergence constitutes a barrier to the horizontal diffusion of innovations (β), and the size of the improvement in productivity at the frontier (Δ).

Our framework predicts a positive correlation between economic distance $|y_j - y_i|$ and relative genetic distance from the frontier $|d_g(i,f) - d_g(j,f)|$. It also accounts for a positive correlation between economic distance and simple genetic distance $d_g(i,j)$ as long as $|d_g(i,f) - d_g(j,f)|$ and

several innovations, including improvements to the innovation process itself, in the spirit of the observation that "the greatest invention of the 19th century was the invention of the method of invention" (Alfred North Whitehead, 1931, p.38; quoted in Howitt and Mayer-Foulkes, 2005).

[18] Without loss of generality, we assume that β is lower than 1/2. Alternatively, the formula could be re-written as $\Delta_i = \max\{[1 - \beta d_v(i,f)]\Delta, 0\}$

$d_g(i,j)$ are positively correlated.[19] At the same time, our theory predicts that relative genetic distance from the frontier should have a stronger impact on economic distance than absolute genetic distance, because relative distance is a more accurate measure of relative distance from the frontier in terms of vertical characteristics. In fact, the expected economic distance associated with an absolute genetic distance $d_g(i,j) = 1$ is $E\{d_e(i,j)|d_g(i,j) = 1\} = \eta\beta\Delta$, while the expected economic distance associated with an equivalent level of relative genetic distance $|d_g(i,f) - d_g(j,f)| = 1$ is higher:[20]

$$E\{d_e(i,j)||d_g(i,f) - d_g(j,f)| = 1\} = \frac{7\eta\beta\Delta}{6} > E\{d_e(i,j)||d_g(i,j) = 1\} \tag{8}$$

In summary, our theory has the following testable implications:

Implication 1. *Relative genetic distance from the frontier is positively correlated with differences in income per capita (economic distance).*

and:

Implication 2. *The effect on income differences associated with relative genetic distance from the frontier is larger than the effect associated with absolute genetic distance.*

As we will see in Section 4, both predictions are consistent with the empirical evidence.

2.2 A General Taxonomy

To clarify the nature of the links between genetic distance and income differences, it is useful to introduce a broader classification of different mechanisms through which the transmission of characteristics across generations may in principle affect economic outcomes.

In general, traits can be transmitted across generations through DNA (call it "genetic transmission", or GT - e.g. eye color) or through pure cultural interactions (call it "cultural transmission", or CT - e.g., a specific language). Moreover, vertical characteristics, whether passed on through GT or CT, may affect income differences because of a direct (D) effects on productivity or because they constitute barriers (B) to the transmission of innovations across populations. There are four

[19] It is easy to verify that the two measures are positively correlated in our theoretical framework. More importantly, relative genetic distance from the frontier and absolute genetic distance are also positively correlated in the actual data, as we show in Section 4. Our framework provides an explanation for the observed positive correlation between economic distance and absolute genetic distance in the data: absolute genetic distance is an imperfect proxy of the economically relevant variable, i.e. relative genetic distance.

[20] An analogous relationship exists between $E\{d_e(i,j) \mid |d_g(i,f) - d_g(j,f)| = 2\}$ and $E\{d_e(i,j) \mid |d_g(i,j) = 2\}$.

possible combinations of mechanisms through which intergenerationally-transmitted characteristics may affect income differences. The following chart summarizes the four possibilities.

	Direct Effect (D)	Barrier Effect (B)
Genetic Transmission (GT)	Quadrant I	Quadrant II
Cultural Transmission (CT)	Quadrant III	Quadrant IV

For instance, genetic traits affecting the trade-off between quality and quantity of children in the theoretical framework proposed by Galor and Moav (2002) to explain the Industrial Revolution would be examples of GT direct effects (Quadrant I).[21] GT barrier effects (Quadrant II) could stem from visible characteristics (say, physical appearance) that do not affect productivity directly, but introduce barriers to the diffusion of innovations by reducing exchanges and learning across populations that perceive each other as different. This effect is related to the already cited study by Guiso et al. (2004), who argue that differences in physical characteristics may affect the extent of trust across populations, and that trust affects bilateral trade between different societies. Visible differences across ethnic groups also play an important role in the theory of ethnic conflict by Caselli and Coleman (2002). Direct economic effects of cultural characteristics have been emphasized in a vast sociological literature that goes back at least to Max Weber.[22] A recent empirical study of the relationship between cultural values and economic outcomes that is consistent with the mechanisms of Quadrant III is provided by Tabellini (2005). The link between differences in vertically-transmitted characteristics - including cultural characteristics, as in Quadrant IV - is at the core of our own model. In the model presented in Section 2.1, differences in neutral characteristics (i.e., traits that do not have a direct effect on productivity) explain income differences by acting as barriers to the diffusion of innovation across populations.

It is worth pointing out that the distinction between GT and CT may be useful to fix ideas, but is not a clear-cut dichotomy.[23] In fact, this distinction (related to the distinction between nature and nurture), if taken too literally, may be misleading from an economic perspective, as well as

[21] For a discussion of related ideas, see also the recent book by Gregory Clark (2007) on the causes of the Industrial Revolution.

[22] Recent references can be found in the edited volume by Harrison and Huntington (2000).

[23] For a recent general discussion of the interactions between biological and cultural transmission, see Richerson and Boyd (2004). Recent results in genetics that are consistent with complex gene-culture interactions are provided by Wang et al. (2006).

from a biological perspective. Generally, the economic effects of human characteristics are likely to result from interactions of cultural and genetic factors, with the effects of genetic characteristics on economic outcomes changing over space and time depending on cultural characteristics, and vice versa. To illustrate this point, consider differences across individuals within a given population (say, the US). Consider a clearly genetic characteristic, for instance having two X chromosomes. This purely genetic characteristic is likely to have had very different effects on a person's income and other economic outcomes in the year 1900 and in the year 2000, because of changes in culturally transmitted characteristics over the century. This is a case where the impact of genes on outcomes varies with a change in cultural characteristics.[24] By the same token, one can think of the differential impact of a given cultural characteristic (say, the habit of drinking alcohol) on individuals with different genetic characteristics (say, genetic variation in alcohol dehydrogenase, the alcohol-metabolizing enzyme). An example of a complex interactions in which culture affects genes is the spread of the gene for lactose tolerance in populations that domesticated cows and goats. Hence, in interpreting our empirical results we do not dwell much on the distinction between genetic and cultural transmission of traits, but instead interpret genetic distance as an overall measure of differences in the whole set of intergenerationally transmitted characteristics.[25]

3 The Genetic Distance Data

3.1 Measuring Genetic Distance

Since the data on genetic distance that we use as a measure of distance in vertical characteristics is not commonly used in the economics literature, and constitutes a central contribution of our paper, we describe it in some detail. Genetic distance measures genetic differences between two

[24] This is a variation on an example by Alison Gopnik in her comment to the Pinker vs. Spelke debate at http://www.edge.org/discourse/science-gender.html#ag. Pinker's response is also available at http://www.edge.org/discourse/science-gender.html.

[25] That said, we do find clues pointing to cultural transmission, rather than purely biological transmission, as a likely mechanism behind our results. For instance, we find large effects of genetic distance on income differences within Europe, among populations that are geographically close, have shared very similar environments, and have had a very short time to diverge genetically. The view that cultural transmission trumps genetic transmission in explaining differences within human populations is standard among geneticists and anthropologists. For nontechnical discussions of these issues, see Diamond (1992, 1997), Cavalli-Sforza and Cavalli-Sforza (1995), and Richerson and Boyd (2004).

populations. The basic unit of analysis is the *allele*, which is a particular form taken by a gene.[26] By sampling populations for specific genes that can take different forms, geneticists have compiled data on allele frequencies.[27] Differences in allele frequencies are the basis for computing summary measures of distance between populations. Following Cavalli-Sforza et al. (1994), we use measures of F_{ST} distance, also known as "coancestor coefficients". F_{ST} distances, like most measures of genetic differences, are based on indices of heterozygosity, the probability that two alleles at a given locus selected at random from two populations will be different. F_{ST} takes a value equal to zero if and only if the allele distributions are identical across the two populations, while it is positive when the allele distributions differ. A higher F_{ST} is associated with larger differences.[28]

Measures of genetic distance can be used to reconstruct phylogenies (or family trees) of human populations. F_{ST} is strongly related to how long two populations have been isolated from each other.[29] In principle, when two populations split apart, their genes can start to change as a result either of random genetic drift or natural selection. When calculating genetic distances in order to study population history and phylogenesis, geneticists concentrate on *neutral* characteristics that are not affected by strong directional selection (Cavalli-Sforza et al., 1994, p. 36). The term "neutral markers" refers to genes affected only by random drift.[30] It is important to stress that our measures of genetic distance are based on such neutral markers only, and not on selected traits. When populations become separated, the process of random drift will take them in different

[26] A gene is commonly defined as a DNA sequence that encodes for a protein. The genetic data in Cavalli-Sforza et al. (1994) have been obtained from "classical analysis," which focuses on protein polymorphism. More recent approaches look directly at the DNA. So far those studies, which include the Human Genome Diversity Project (http://www.stanford.edu/group/morrinst/hgdp.html) and the International HapMap Project (http://www.hapmap.org/), have confirmed the results from classical protein analysis, but are not yet available for extensive cross-regional analysis.

[27] Allele frequencies for various genes and for most populations in the world can be found at http://alfred.med.yale.edu/

[28] Appendix A provides an illustration of the construction of F_{ST} for the simple case of two populations of equal size, and one gene that can take only two forms (i.e., two alleles).

[29] Isolation here refers to the bulk of the genetic heritage of a given population. As stressed by Cavalli-Sforza et al. (1994), small amounts of intermixing between members of different populations do not affect measured genetic distance.

[30] The classic reference for the neutral theory of molecular evolution is Kimura (1968). For more details on the neutral theory, the *molecular clock* hypothesis, and the construction and interpretation of measures of genetic distance, see Jobling et al. (2004).

directions, raising their genetic distance. The longer the period of separation, the greater genetic distance becomes. If drift rates are constant, genetic distance can be used as a molecular clock - that is, the time elapsed since two populations separated can be measured by the genetic distance between them.[31] Consequently, F_{ST} is a measure of distance to the most recent common ancestors of two populations, or, equivalently, as their degree of genealogical relatedness.

To summarize, we use F_{ST} distance as a measure of genealogical relatedness between populations. A larger F_{ST} distance reflects a longer separation between populations, and hence, on average, a larger difference in vertical characteristics.

3.2 The World Sample

The genetic distance data is from Cavalli-Sforza et al. (1994), p. 75-76. Our main focus on the set of 42 world populations for which they report all bilateral distances, computed from 120 alleles.[32] These populations are aggregated from subpopulations characterized by a high level of genetic similarity. However, measures of bilateral distance among these subpopulations are available only regionally, not for the world as a whole. Among the 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, and the smallest is between the Danish and the English, for which the genetic distance is 0.0021.[33] The mean genetic distance among the 861 available pairs is 0.1338. Figure 2, from Cavalli-Sforza et al. (1994, Figure 2.3.2B, p. 78), is a phylogenetic tree illustrating the process by which different human populations have split apart over time.[34] Such phylogenetic trees,

[31] When genetic distance is based on neutral markers, and populations are sufficiently large, geneticists have shown that drift rates are indeed constant (very small populations are generally subject to faster random genetic drift).

[32] Cavalli-Sforza et al. (1994) also provide a different measure of genetic distance (Nei's distance). Nei's distance, like F_{ST}, measures differences in allele frequencies across a set of specific genes between two populations. F_{ST} and Nei's distance have slightly different theoretical properties, but the differences are unimportant in practice as their correlation is 93.9% (Table 1) We show below that the choice of measures does not affect our results.

[33] Among the more disaggregated data for Europe which we also gathered, the smallest genetic distance (equal to 0.0009) is between the Dutch and the Danish, and the largest (equal to 0.0667) is between the Lapp and the Sardinians. The mean genetic distance across European populations is 0.013. Genetic distances are roughly ten times smaller on average across populations of Europe than in the World dataset.

[34] The figure was constructed to maximize the correlation between Euclidian distances to common nodes, measured along the branches, and the F_{ST} genetic distance computed directly from allele frequencies. Hence, the tree-diagram is a simplified summary of (but not a substitute for) the matrix of genetic distances between populations, organized

constructed from genetic distance data, are the population analogs of family trees for individuals.

Genetic distance data is available at the population level, not at the country level. It was thus necessary to match populations to countries. We did so using ethnic composition data by country from Alesina et al. (2003). In many cases, it was possible to match ethnic group labels with population labels from Cavalli-Sforza et al. (1994), using their Appendices 2 and 3 to identify the ethnic groups sampled to obtain genetic distances. This was supplemented with information from Encyclopedia Britannica when the mapping of populations to countries was not achievable from ethnic groups data. Obviously, many countries feature several ethnic groups. Whenever the shares of these groups were available from Alesina et al. (2003) and the match to a genetic group was possible, we matched each of a country's ethnic group to a genetic group. For instance, the Alesina et al. (2003) data on ethnic groups has India composed of 72% of "Indo-Aryans" and 25% "Dravidians". These groups were matched, respectively, to the Cavalli-Sforza groups labeled "Indians" and "Dravidhans" (i.e. S.E. Indian in Figure 4).[35]

This match served as the basis for constructing measures of genetic distance between countries, rather than groups. We constructed two such measures. 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. This resulted in a dataset of $21,321$ pairs of countries (207 underlying countries and dependencies) with available genetic distance data. The second was a measure of weighted genetic distance. Some 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

by clusters. It is important to notice that the organization of populations by tree does not imply that genetic distance establishes a "linear relation" among all of them, either along to x-axis (abscissa) or along the y-axis (ordinate). The abscissa at the bottom of the diagram can be used to read the genetic distance between pairs of populations in the tree only when they share direct common ancestors. For example, the genetic distance between New Guineans and Australians can be calculated by reading the position of the node that separates the two populations, which is approximately at 0.1. It is also possible to measure average genetic distance between clusters of populations, by reading the position of the node that separates two clusters on the ascissa. For example, the average genetic distance between African populations and the rest of the world is approximately 0.2. However, in order to read the genetic distance between any pair of populations, one should use (as we do) the complete matrix of genetic distances, which is provided in Cavalli-Sforza et al. (1994, Table 2.3.1A, p. 75).

[35] The complete match of genetic groups to ethnic groups, and in turn to countries, is available upon request.

(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:

$$F_{ST}^{W} = \sum_{i=1}^{I} \sum_{j=1}^{J} (s_{1i} \times s_{2j} \times d_{ij}) \tag{9}$$

where s_{ki} is the share of group i in country k, d_{ij} is the F_{ST} genetic distance between groups i and j.[36] The interpretation of this measure is straightforward: it represents the expected genetic distance between two randomly selected individuals, one from each country. Weighted genetic distance is very highly correlated with genetic distance based on dominant groups (the correlation is 94%), so for practical purposes it does not make a big difference which one we use. We will use the weighted F_{ST} distance as the baseline measure throughout this study, as it is a more precise measure of average genetic distance between countries.

Error in the matching of populations to ethnic groups should lead us to understate the correlation between genetic distance and income differences. Several regions may be particularly prone to matching errors. One is Latin America, where it is sometimes difficult to identify whether populations are predominantly of European descent or of Amerindian descent. This is particularly problematic in countries with large proportions of Mestizos, i.e. populations of mixed descent, such as Colombia (in this specific case the country's dominant group was matched to the South Amerindian category). Another is Europe, where countries can only be matched to one of four genetic groups (Danish, English, Greek and Italian). As a strict rule, we matched countries to groups that were the closest genetically to that country's population, using the regional genetic distance data in Cavalli-Sforza et al. (1994).

The ethnic composition in Alesina et al. (2003) refers to the 1990s. This is potentially endogenous with respect to current income differences if the latter are persistent and if areas with high income potential tended to attract European immigration since 1500. This would be the case for example under the view that the Europeans settled in the New World due to a favorable geographical environment.[37] In order to construct genetic distance between countries as of 1500, we also mapped populations to countries using their ethnic composition as of 1500, i.e. prior to the major

[36] Due to missing data on group shares, the weighted measure only covers 16,110 pairs, or 180 countries.

[37] In fact, income differences are not very persistent at a long time horizon such as this - see Acemoglu et al. (2002). Our own data shows that pairwise log income differences in 1500 are uncorrelated with the 1995 series in the common sample (Table 2).

colonizations of modern times. Thus, for instance, while the United States is classified as predominantly populated with English people for the current match, it is classified as being populated with North Amerindians for the 1500 match. This distinction affected mostly countries that were colonized by Europeans since 1500 to the point where the dominant ethnic group is now of European descent (New Zealand, Australia, North America and some countries in Latin America). Since we do not have data on ethnic composition going back to 1500, the corresponding match only refers to plurality groups. Genetic distance in 1500 can be used as a convenient instrument for current genetic distance. The matching of countries to populations for 1500 is also more straightforward than for the current period, since Cavalli-Sforza et al. (1994) attempted to sample populations as they were in 1500, likely reducing the extent of measurement error.

3.3 The European Sample

Cavalli-Sforza et al. (1994) also present matrices of genetic distance among populations within several regions. These sub-matrices cannot be merged with the world data, because they are based on sets of underlying genes distinct from the 120 genes used for the 42 populations in the world sample, and because the genetic distance between most groups in the regional samples and in the World sample are unavailable. They can, however, be used separately. We assembled a dataset of genetic distances between 26 European populations, a much finer classification than the world sample which only featured 4 distinct (non-minority) European populations (English, Danish, Italian and Greek).[38] Matching populations to countries is more straightforward for the European sample than for the world sample, because the choice of sampled European populations generally corresponds to nation-state boundaries. This should reduce the incidence of measurement error. The populations were matched to 26 countries, resulting in 325 country pairs.[39] The largest F_{ST} genetic distance among those pairs was 0.032, between Iceland and Slovenia. The smallest, among countries matched to distinct genetic groups, was between Denmark and the Netherlands

[38] Minority populations in the world sample also include Basque, Lapp and Sardinian.

[39] These 26 countries are Austria, Belgium, Croatia, Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Macedonia, Netherlands, Norway, Poland, Portugal, Russian Federation, Slovak Republic, Slovenia, Spain, Sweden, Switzerland the United Kingdom and Serbia/Montenegro. The Basque, the Lapp and the Sardinian populations were not matched to any country, and some countries were matched to the same groups (Croatia, Slovenia, Macedonia and Serbia/Montenegro were all matched to the Yugoslavian population, while the Czech Republic and Slovak Republics were both matched to the Czech population).

($F_{ST} = 0.0009$).

4 The Empirics of Income Differences

In this section we test the empirical implications of our model. We investigate the relationship between genetic distance and economic distance. Genetic distance is considered both relative to the technological frontier and in absolute terms. In line with our theory, we use the log income per capita as a metric of economic performance. The data on per capita income is purchasing power-parity adjusted data from the World Bank, for the year 1995.[40]

4.1 Genetic Distance to the Frontier

We start with a simple descriptive approach. Suppose that we can pinpoint the technological frontier. Does a country's genetic distance to the frontier correlate with its income level? To investigate this hypothesis, we ran income level regressions, and for now confine our attention to the World sample, where we have data on all our variables for 137 countries. We assumed the US was the technological frontier. This choices seems reasonable a priori: in the World sample, only Luxemburg and Norway had incomes per capita higher than the US in 1995.[41] Few would dispute that the US is the world's major technological innovator. We measure distance to the US using our weighted measure, which is more appropriate since the US is a genetically diverse country (variation in this measure is dominated by the distance to the English population).

Table 1 presents the results. In column 1, genetic distance to the US is entered alone, and the coefficient has the expected negative sign and is highly significant statistically, with a t-statistic of about 9.3. In this specification, genetic distance entered alone accounts for 39% of the variation in log income levels. Figures 3 displays the univariate results of column (1) graphically, so the reader can evaluate which countries drive the result. Columns 2 and 3 add several controls for geographic distance from the US and transport costs (column 2) as well as linguistic and religious differences (column 3). We will say a lot more about these control variables below, but for now it suffices to

[40] We also used data from the Penn World Tables version 6.1 (Heston, Summers and Aten, 2002), which made little difference in the results. We focus on the World Bank data for 1995 as this allows us to maximize the number of countries in our sample.

[41] It is likely both countries do not owe their economic superiority primarily to their technological inventiveness, but to natural resource wealth (Norway) and a specific specialization pattern (Luxemburg).

note that the coefficient on genetic distance is barely affected by the inclusion of the geographic distance controls, and that its magnitude is only reduced by 20% when including linguistic and religious distance. The latter variables, incidentally, represent the type of culturally-transmitted vertical characteristics that are part of the range of factors captured by genetic distance. That their inclusion reduces the effect of genetic distance is therefore consistent with our story. Overall, these results provide preliminary evidence that a country's genetic distance to the US is significantly associated with lower per capita income, even after controlling for a variety of metrics of geographic and other distances.

4.2 Bilateral Approach

To generalize the results of the previous subsection, we consider a specification in which the absolute difference in income between pairs of countries is regressed on measures of distance between the countries in this pair. In addition to being closer to our theoretical specification, this has two advantages. First, we no longer always need to choose a technological leader to investigate the correlation between absolute genetic distance and income differences (our model led to predictions about the sign of this correlation, see Section 2.1). Second, we can make more efficient use of a wealth of bilateral distance data as regressors. We will use this bilateral approach for the rest of this paper.

We computed income differences between all pairs of countries in our sample for which income data and other was available, i.e. $9,316$ pairs (based on 137 underlying countries) in the World sample, and 325 pairs (based on 26 underlying countries) in the Europe sample. Define G^D_{ij} as the *absolute* genetic distance between countries i and j. Denote G^R_{ij} the genetic distance between i and j *relative* to the technological frontier. In most of what follows, we continue to assume the technological frontier is the United States. Then, by definition $G^R_{ij} = |G^D_{i,US} - G^D_{j,US}|$.[42] Our

[42] Absolute and relative genetic distance are algebraically the same when one of the two countries in a pair is the frontier economy, and the two measures are also closely correlated when considering pairs involving one country that is very close genetically to the frontier economy. The measures differ most for countries that are relatively genetically far from each other (e.g. Ethiopia and Nigeria) but roughly equally distant from the frontier – in this case absolute distance is large, and relative distance is small. Relative genetic distance is meant to capture the fact that *per se* the distance between Nigeria and Ethiopia does not matter to explain their income difference, since they are unlikely to learn frontier technologies from each other. Rather what matters is their relative distance from the US.

baseline specifications are:

$$|\log y_i - \log y_j| = \beta_0 + \beta_1 G_{ij}^D + \beta_2' X_{ij} + \varepsilon_{ij} \tag{10}$$

and:

$$|\log y_i - \log y_j| = \gamma_0 + \gamma_1 G_{ij}^R + \gamma_2' X_{ij} + \nu_{ij} \tag{11}$$

where X_{ij} is a set of measures of geographic and cultural distance and ε_{ij} and ν_{ij} are disturbance terms.[43]

The reason our empirical specification must involve income differences rather than a single country's income level on the left hand side is that this makes the use of bilateral measures of distance possible. Conceptually, therefore, we depart from existing methodologies: our regression is not directional, i.e. our specification is not simply obtained by differencing levels regressions across pairs of countries.[44] We should also stress that our specifications are reduced forms. That is, differences in income are presumably the result of differences in institutions, technologies, human capital, savings rates, etc., all of which are possibly endogenous with respect to income differences, and themselves a function of geographic and human barriers.

Before turning to the results, we must address a technical point regarding the disturbances ε_{ij} and ν_{ij}. In principle, if one is willing to assume that the measures of barriers are exogenous, equations (10) and (11) can be estimated using least squares. However, in this case usual methods of inference will be problematic due to spatial correlation resulting from the construction of the dependent variable.[45] Appendix B illustrates why introducing the difference in log income on the left hand side results in spatial correlation, due to the definition of the dependent variable.

To address the problem of spatial correlation, we rely on two-way clustering of the standard errors, following the approach in Cameron, Gelbach and Miller (2006). In our application, clustering

[43] We also estimated an alternative specification where the distance measures were all entered in logs. This did not lead to appreciable differences in the economic magnitude or statistical significance of any of the estimates. Since several countries were matched to the same genetic group, so that the corresponding pairs had a genetic distance of zero, taking logs resulted in the loss of valuable observations, so we omit these results here.

[44] Our methodology is as much akin to gravity regressions in the empirical trade literature as it is to levels or growth regressions in the literature on comparative development.

[45] This, of course, was not a concern in the simple regressions presented in Section 4.1. These results featured t-statistics in excess of 9 in the world sample, much larger than the t-statistics that we find using the bilateral approach with two-way clustering. This reinforces our confidence that our results are not driven by standard errors that are too low due to spatial correlation.

arises at the level of country 1 and at the level of country 2, and is non-nested: each individual observation on income differences, say $|\log y_i - \log y_j|$ belongs to the group that includes country i and the group that includes country j. The estimator in Cameron, Gelbach and Miller (2006) allows for an arbitrary correlation between errors that belong to "the same group (along either dimension)" (p. 7). Their method is therefore directly applicable to the specific econometric issue we face (on page 3 of their manuscript the authors specifically mention spatial correlation as a possible application of their estimator). Results obtained with this method feature standard errors that are an order of magnitude larger than those obtained with simple OLS with heteroskedasticity-robust standard errors, suggesting that spatial correlation was indeed an important issue. However, as we show below, genetic distance remains statistically significant even after correcting the standard errors for spatial correlation.[46]

4.3 Unconditional Results

Table 2 presents some summary statistics for our variables. Throughout, we use a baseline sample of $9,316$ country pair observations obtained from 137 underlying countries. We consider various measures of genetic distance. As already mentioned, our baseline measure is weighted F_{ST} genetic distance. We also used the weighted Nei genetic distance.[47] These measures bear high correlations among themselves (0.94), and in practice it matters little which one we use. On the other hand,

[46] There are in principle several other ways to address the problem of spatial correlation. One approach would be to do feasible GLS by explicitly estimating the elements of the covariance matrix, and introducing the estimated covariance matrix as a weighing matrix in the second stage of the GLS procedure. This is computationally very demanding as the dimensionality of the matrix is large - in our application we have over $9,316$ country pairs with available data on the variables of interest, and up to 137 covariance terms to estimate (for the same reason, it is difficult to implement tests of spatial correlation in our context). Another approach, which we pursued in a previous version of this paper, is to include in our regressions common country fixed-effects, meant to soak up the spatial correlation. For this we relied on well-known results cited in Case (1991), showing that fixed effects soak up spatial correlation, though in a context quite different from ours. Following this insight, we modeled : $\varepsilon_{ij} = \sum_{k=1}^{N} \gamma_k \delta_k + \eta_{ij}$ where $\delta_k = 1$ if $k = i$ or $k = j$, $\delta_k = 0$ otherwise, and η_{ij} is a well-behaved disturbance term. We treated δ_k as fixed effects, i.e. we introduced in the regression a set of N dummy variables δ_k each taking on a value of one $N-1$ times - δ_j takes a value of 1 whenever country j appears in a pair. This did not affect the qualitative and quantitative nature of our results compared to the solution we pursue here - but the standard errors were smaller than the ones we report in this version.

[47] In past work we also used both F_{ST} and Nei genetic distance based on plurality groups, with results very similar to those reported here.

the theoretically more appropriate measure of relative distance to the US bears a correlation of only 0.63 with absolute F_{ST} genetic distance. Finally, we considered F_{ST} genetic distance with countries matched to populations as they were in 1500. The correlation between this variable and the current measure is 0.83.

Our measure of absolute F_{ST} genetic distance, G^D, bears a positive correlations of 0.20 with the absolute value of log income differences in 1995. Genetic distance relative to the frontier, G^R, bears a higher correlation with income differences, equal to 0.34, which is directly in line with our model's prediction (these correlations are higher in the European sample, respectively 0.33 and 0.41).

Table 3 presents univariate regressions of income differences on various measures of genetic distance for the World sample. As a measure of the magnitude of the coefficients, we report the standardized beta coefficient on genetic distance for each regression.[48] Column 1 shows that, when entered alone in the regression, one standard deviation in F_{ST} genetic distance between plurality groups accounts for 16.79% of a standard deviation of income differences. This effect rises in magnitude to 26.98% when we consider genetic distance (also between plurality ethnic groups of each country in a pair) relative to the frontier (column 2). This means that the effect of genetic distance relative to the frontier is larger than the effect of bilateral genetic distance, exactly as predicted by Implication 2 of our model. Turning to the weighted measure, similar effects are found, with slightly larger magnitudes (columns 3 and 4), respectively 19.71% and 33.65%. The larger magnitudes are consistent with the idea that weighted measures are better proxies for the expected genetic distance between countries. The effect is also larger when Nei genetic distance is used instead of F_{ST} (column 5).

We next make use of data from Cavalli-Sforza et al. (1994) on the standard deviation of the genetic distance estimates. Since these data are based on allele frequencies collected from samples of different sizes, they are estimated more or less precisely depending on population pairs. We have data on the standard errors of each estimate of genetic distance, obtained from bootstrap analysis. In column 6, we linearly downweigh observations with higher standard errors on genetic distance. As expected, the magnitude of the resulting weighted least squares effect of F_{ST} genetic distance is larger than under simple OLS, consistent with the idea that measurement error is greater for

[48] The standardized beta is defined as the effect of a one standard deviation change in the regressor expressed as a percentage of one standard deviation of the dependent variable

pairs with high standard errors on genetic distance. Similar results are obtained using alternative measures of genetic distance.

While providing suggestive evidence in favor of implications 1 and 2 of our theoretical model, these unconditional results may confound the effect of barriers linked to vertical characteristics with geographic barriers. In the next subsection, we control for a large number of measures of geographic distance. In everything that follows we will focus on the weighted relative genetic distance to the frontier as the baseline measure of genetic distance, i.e. the measure used in column (4) of Table 3, since this is the more theoretically appropriate measure.

4.4 Controlling for Geographic Factors

Genetic distance and geographic isolation are likely to be highly correlated. The more isolated two groups become, the more they will drift apart genetically, since genetic admixture is made difficult by geographic barriers. It is therefore important to adequately control for geographic isolation: failing to do so would ascribe to genetic distance an effect that should be attributed to geographic distance. In this subsection, we control for a vast array of measures of geographic isolation.

Distance metrics. Our first set of measures of geographic isolation between countries consists of various measures of distance. We consider a measure of the greater circle (geodesic) distance between the major cities of the countries in our sample, from a dataset compiled by researchers at CEPII.[49] We also include latitudinal distance - i.e. simply the absolute value of the difference in latitude between the two countries i and j in each pair: $G_{ij}^{LA} = |\text{latitude}_i - \text{latitude}_j|$. Latitude could be associated with climatic factors that affect income levels directly, as in Gallup, Mellinger and Sachs (1998) and Sachs (2001). Latitude differences would also act as barriers to technological diffusion: Diamond (1997) suggests that barriers to the transmission of technology are greater along the latitude direction than along the longitude direction, because similar longitudes share the same climate, availability of domesticable animal species, soil conditions, etc. We should therefore expect countries at similar latitudes to also display similar levels of income. Third, we use a measure of

[49] The data is available at http://www.cepii.fr/anglaisgraph/bdd/distances.htm. This dataset features various measures of distance (between major cities, between capitals, weighted using several distances between several major cities, etc.), all of which bear pairwise correlations that exceed 99%. The dataset also includes other useful geographical and historical controls, such as whether pairs of countries whether the countries are contiguous, whether they had a common colonizer, were ever part of a single country, etc. We use these below.

longitudinal distance, $G_{ij}^{LO} = \left|\text{longitude}_i - \text{longitude}_j\right|$, to capture possible geographic isolation along this alternative axis.

Perhaps surprisingly, raw correlations between genetic distance and these simple measures of geographic distance are not as high as we might have expected. For instance, the correlation between geodesic distance and weighted F_{ST} genetic distance relative to the US is only 9.54% - though unsurprisingly it rises to 31.4% if genetic distance is measured based on populations as they were in 1500, because the colonization era acted to weaken the link between genetic distance and geographic distance by shuffling populations across the globe. The largest correlation is found with the absolute difference in latitudes, bearing a correlation of 37.86% with weighted genetic distance relative to the US.

Table 4, column 2 includes these three measures jointly with F_{ST} genetic distance relative to the frontier. The effect of relative genetic distance barely changes at all compared to the baseline univariate regression replicated in column (1). We find evidence that latitudinal distance matters - the standardized beta on this variable is 11.97%, consistent with Jared Diamond's hypothesis.

Microgeographic factors. In addition to these straightforward distance measures, we controlled for other measures of isolation between countries. In the context of gravity regressions for Europe, Giuliano et al. (2006) argued that genetic distance was likely correlated with features of the terrain that raise transport costs. These "microgeographic" features may not be well captured by simple metrics of distance. To account for this possibility, we included dummy variables taking a value of 1 if countries in a pair were contiguous, if they had access to a common sea or ocean, if any country in a pair was an island or was landlocked.[50] These measures are meant to capture ease of communication and travel between countries, which may be associated both with barriers to technological diffusion and to population isolation (and thus genetic distance). Column 3 of Table 4 shows these variables have the expected signs, but their inclusion does not affect the coefficient on genetic distance. To summarize, although the additional controls have explanatory power for income differences, we found no evidence that the inclusion of microgeographic factors modifies the

[50] The common sea variable is the same as that used in Giuliano et al. (2006). These authors also used a measure of the average elevation of the countries that lie between any two countries in a pair, as a measure of how hard it is to travel from one to the other. While we calculate and use this variable for the Europe sample, where it is relatively straightforward to do so, there are simply too many possible paths between any two countries in the world for this to be practical in the broader sample of world countries.

effect of genetic distance.

Transport costs. A good summary measure of geographic isolation is transportation costs. Giuliano et al. (2006) use a new measure of transport costs based directly on freight rates for surface transport (sea or land) between European countries. We have obtained the same data as the one they used, for the World sample.[51] Column 4 of Table 4 adds this measure of freight costs to our specification. We find that freight costs bear a positive relationship with income differences, as expected. However, this effect is not significant statistically and does not affect the signs or magnitudes of the other included variables, particularly genetic distance. We find no evidence that genetic distance captures the effect of geographic isolation or transportation costs in our application.[52]

Continent effects. The largest genetic distances observed in our worldwide dataset occur between populations that live on different continents. One concern is that genetic distance may simply be picking up the effect of cross-continental barriers to the diffusion of development, i.e. continent effects. If this were the case, it would still leave open the question of how to interpret economically these continent effects, but to test explicitly for this possibility, we added to our baseline specification two sets of continent dummies. We included one set of 7 dummies (one for each continent) taking on a value of one if the two countries in a pair are on the same continent. We also included a set of 7 dummies each equal to one if exactly one country belongs to a given continent, and the other not. The results are in column 5 of Table 4. The inclusion of continent dummies reduces by about one third the magnitude of the genetic distance effect, but the latter remains statistically significant. Its magnitude is still large, with a standardized beta of 21.88%.

Figure 3 shows that many countries the most genetically and economically distant countries from the US are in Sub-Saharan Africa. To examine whether this drives our results, we excluded any pair involving a Sub-Saharan African country from our sample. In the resulting regression (available upon request), the standardized beta on genetic distance was 32.11%, and was highly

[51] The data is available from http://www.importexportwizard.com/. The measure we used referred to 1000kg of unspecified freight transported over sea or land, with no special handling. This is the same definition used in Giuliano et al. (2006). The data on $10,825$ pairs of countries were downloaded from the website using a Perl script.

[52] In the previous version of this paper, we also used the approach in Limao and Venables (2001) and Hummels and Lugovskyy (2006) to measure trade costs indirectly through the matched partner technique, using the ratio of CIF to FOB exports. The measure of indirect trade costs is $ITC_{ij} = (CIF_{ij}/FOB_{ij}) - 1$. Results with this alternative measure of trade costs featured a much smaller sample, but were unchanged compared to the ones reported here.

significant statistically. We therefore find no evidence that our results are driven by the inclusion of Africa in our sample. We will provide further evidence on the within-continent effects of genetic distance using the European dataset in Section 4.7.

Climatic similarity. Next, we constructed measures of climatic similarity based on 12 Koeppen-Geiger climate zones.[53] One measure is the average absolute value difference, between two countries, in the percentage of land area in each of the 12 climate zones. Countries have identical climates, under this measure, if they have identical shares of their land areas in the same climates. As a simpler alternative, we used the absolute difference in the percentage of land areas in tropical climates. As with latitude, climate may have direct effects on productivity, or barrier effects on technological diffusion: countries located in different climates may have experienced difficulties in adopting each other's mode of production, particularly in the agrarian era. Columns 6 and 7 of Table 4 report the results. As expected, climatic differences are associated with greater income differences, even controlling for latitude differences. However, the inclusion of these variables hardly affects the coefficient on genetic distance.

Relative geographic distance. One concern with the above regressions is that geographic distance is entered in absolute terms, not relative to the frontier. We did this because some geographic variables (for instance latitude) can bear a direct effect on income levels, in addition to a barrier effect on income differences. If these variables acted as barriers, by a reasoning analogous to the one that applied to genetic distance, what would matter would be geographic distance relative to the frontier, rather than absolute distance. In Table 5, all the measures of distance and transport costs are entered relative to the USA. Columns 2, 3 and 4 replicate, respectively, columns 3, 4 and 5 of Table 4. If anything, the coefficient on genetic distance is now larger. Interestingly, transport costs relative to the US become significant at the 10% level. We conclude that our results are robust to controlling for a wide variety of measures of geographic distance, micro-geographic measures of isolation, continent effects, climatic differences and transportation costs, whether entered in absolute terms or relative to the frontier.

[53] The 12 Koeppen-Geiger climate zones are: tropical rainforest climate (Af), monsoon variety of Af (Am), tropical savannah climate (Aw), steppe climate (BS), desert climate(BW), mild humid climate with no dry season (Cf), mild humid climate with a dry summer (Cs), mild humid climate with a dry winter (Cw), snowy-forest climate with a dry winter (Dw), snowy-forest climate with a moist winter (Ds), tundra/polar ice climate (E) and highland climate (H). The data, compiled by Gallup, Mellinger and Sachs, is available at http://www.ciesin.columbia.edu/eidata/.

4.5 Endogeneity and the Diamond Gap

Possible endogeneity of current genetic distance. We attempted to control for the possible endogeneity of genetic distance with respect to income differences. While differences in (neutral) allele frequencies between the populations of two countries do not result causally from income differences, migration could lead to a pattern of genetic distances today that is closely linked to current income differences. The first order issue arises from the pattern of colonization of the New World starting after 1500. Europeans tended to settle in larger numbers in the temperate climates of North America and Oceania. If geographic factors bear a direct effect on income levels, and were not properly accounted for in our regressions through included control variables, then genetic distance today could be positively related to income distance not because genetic distance precluded the diffusion of development, but because similar populations settled in regions prone to generating similar incomes.

To assess this possibility, we use our data on F_{ST} genetic distance as of 1500, relative to the English population, as an instrument for current genetic distance in column 1 of Table 6. This variable reflects genetic distance between populations as they were before the great migrations of the modern era, i.e. as determined since the Neolithic era, and yet is highly correlated (61.12%) with current genetic distance relative to the US, so it fulfills the conditions of a valid instrument. Here, the magnitude of the genetic distance effect is raised by one third, with a standardized beta reaching 49.75%. As is usual in this type of application, the larger estimated effect may come from a lower incidence of measurement error under two-stage least squares - as explained above the matching of populations to countries is much more straightforward for the 1500 match.

To further assess whether our results are driven by endogeneity of the sort discussed above, column 2 of Table 6 excludes from the sample any pairs involving one or more countries from the New World (defined as countries in North America, Latin America, the Caribbean and Oceania), where the endogeneity problem is likely to be most acute. The effect of genetic distance falls by about one third, but remains both statistically and economically significant. The difference in latitudes becomes much larger, an observation to which we shall return below.

The Diamond Gap. Jared Diamond's (1997) influential book stressed that differences in latitude played an important role as barriers to the transfer of technological innovations in early human history, and later in the pre-industrial era, an effect that could have persisted to this day. Our estimates of the effect of latitudinal distance provided evidence that this effect was still at play: in our regressions we found evidence that differences in latitudes help explain income differences across countries, and this effect was much larger when excluding the New World from our sample. However, Diamond took his argument one step further, and argued that Eurasia enjoyed major advantages in the development of agriculture and animal domestication because a) it had the largest number of potentially domesticable plants and animals, and b) had a predominantly East-West axis that allowed an easier and faster diffusion of domesticated species. By contrast, differences in latitudes in the Americas and Africa created major environmental barriers to the diffusion of species and innovations. More generally, Eurasia might have enjoyed additional benefits in the production and transfer of technological and institutional innovations because of its large size.[54] It is important to properly control for Diamond's geography story as it is either a substitute or a complement to ours.

To test and control for a Eurasian effect, we constructed a dummy variable that takes on a value of 1 if one and only one of the countries in each pair is in Eurasia, and 0 otherwise (the "Diamond gap").[55] In order to test Diamond's hypothesis, we added the Diamond gap to regressions explaining income differences in 1995 (column 3 of Table 6) and, using Maddison's historical income data, in 1500 (column 4). For the former regression, we restricted our sample to the Old World.[56] As expected, in the regression for 1995 income differences, the Diamond gap enters with a positive and significant coefficient, and its inclusion reduces (but does not come close to eliminating) the effect of genetic distance. In column 4, using 1500 income differences as a dependent variable, the Diamond gap is also significant and large in magnitude, despite the paucity of observations. Again, the effect of genetic distance relative to the English population using the 1500 match remains large in magnitude, with a standardized beta of 37.96%. This provides suggestive quantitative evidence in favor of Diamond's observation that the diffusion of development was faster in Eurasia. We also conclude that genetic distance between populations plays an important role in explaining income

[54] This point is stressed in Kremer (1993). See also Masters and McMillan (2001).

[55] For further tests providing statistical support for Diamond's observations, see Olsson and Hibbs (2005).

[56] It is appropriate to exclude the New World from the sample when using 1995 incomes because Diamond's theory is about the geographic advantages that allowed Eurasians to settle and dominate the New World. If we were to include the New World in a regression explaining income differences today, we would include the higher income per capita of non-aboriginal populations who are there because of guns, germs and steel, i.e. thanks to their ancestors' Eurasian advantage.

differences even when controlling for the environmental advantages and disadvantages associated with Eurasia. Diamond's hypothesis on the long-term diffusion of development is complementary to ours.

4.6 Controlling for Common History and Cultural Distance

In this subsection we control for additional possible determinants of income differences.[57] We first consider common history variables: countries that have shared a common history, for instance a common colonial past, may be closer genetically, culturally and economically. Second, we include mesaures of cultural distance, which is also potentially correlated with both genetic distance and income differences. As argued in Cavalli-Sforza et al. (1994) there is usually very little genetic admixture between populations that speak different languages, for instance, so irrespective of their physical distance linguistically distant populations should also be genetically distant. Moreover, as we have argued in Section 2, genetic distance may be associated with differences in vertical characteristics transmitted culturally, rather than genetically. The types of barriers captured by genealogical relatedness (i.e. genetic distance) likely include slow-moving cultural barriers, such as linguistic barriers and difference in norms or values. Are there specific cultural traits that are correlated with genetic distance, and that are directly measurable? How much of the estimated effect of genetic distance is attributable to differences in specific measurable cultural characteristics?

Measures of linguistic and religious distance. Measuring cultural distance is fraught with difficulties. In a well-known survey over fifty years ago Kroeber and Kluckhohn (1952) listed 164 definitions of culture proposed by historians and social scientists. We adopt a more parsimonious approach, confining our attention to measures of linguistic and religious distance. The most salient example of a slowly changing culturally transmitted characteristic is language: it is transmitted across generations within a population, yet there is no gene for speaking specific languages. We attempt to measure cultural distance in three ways.

Our first approach follows Fearon (2003). Fearon used data from Ethnologue to create linguistic trees, classifying languages into common families and displaying graphically the degree of relatedness of world languages. The linguistic tree in this dataset contains up to 15 nested classifications. If two languages share many common nodes in the tree, these languages are more likely to trace

[57] Throughout this subsection we will use the specification in column 4 of Table 4 as the baseline - i.e. we include a large array of geographic isolation controls.

their roots to a more recent common ancestor language. The number of common nodes in the linguistic tree, then, is a measure of linguistic similarity. For instance, according to this measure, French and Italian share 4 common nodes - both belong to the Indo-European / Italic / Romance / Italo-Western linguistic groupings. Using data on the linguistic composition of countries (also from Fearon, 2003), and matching languages to countries, we can construct indices of linguistic distance between countries. We did so, as for genetic distance, in two ways: first, we computed a measure of the number of common nodes shared by languages spoken by plurality groups within each country in a pair. Second, we computed a weighted measure of linguistic similarity, representing the expected number of common linguistic nodes between two randomly chosen individuals, one from each country in a pair (the formula is the same as that of equation 9).[58] Following Fearon (2003), we transformed each of these series so that they are increasing in linguistic distance (LD) and bounded by 0 and 1:

$$LD = \sqrt{\frac{(15 - \#\text{ Common Nodes})}{15}} \tag{12}$$

Our second measure of linguistic distance is based on work in the field of lexicostatistics (a branch of linguistics). We use data from Dyen, Kruskal and Black (1992). They assembled data on 200 common "meanings" from all Indo-European languages. For each language, they compiled lists of words expressing these meanings. When words from two languages expressing a given meaning originated from a common source, these words were labelled as "cognate". For instance, the word "table" in French and "tavola" in Italian are cognate because both stem from the word "tabula" in Latin. Aggregating over the 200 meanings, our measure of linguistic difference is the percentage of non-cognate words, and as before we can compute an expected (weighted) measure and a measure based on the percentage of cognate words between the languages spoken by the plurality linguistic groups in each country in a pair. Again, the greater the percentage of cognate words, the more recently the languages shared a common ancestor language. In contrast to the linguistic trees data, this measure has the advantage of being a continuous measure of similarity. Its main drawback is that it is only available for Indo-European languages, so the geographic coverage is reduced to 62 countries (when considering the % cognate between plurality languages).

Our third measure of cultural distance is based on religious differences. Religion also tends to be transmitted intergenerationally within populations. We followed an approach similar to

[58] Using the measure based on the plurality language or the weigthed measure did not make any difference for the results. In keeping with what we did for genetic distance, we focus on weighted measures in our empirical work.

that used for linguistic distance. We relied on a nomenclature of world religions obtained from Mecham, Fearon and Laitin (2006).[59] This nomenclature was broken down into religious families, first distinguishing between monotheistic religions of Middle-Eastern origin, Asian religions and "others", then subdividing each group into finer groups (such as Christians, Muslims and Jews) and so on. The number of common classifications (up to 5 in this dataset) is a measure of religious proximity. We matched religions to countries using Mecham, Fearon and Laitin (2006)'s data on the prevalence of religions by country and transformed the data in a manner similar to that in equation (12).

Pairwise correlations between measures of genetic, linguistic and religious distances are displayed in Table 7. These correlations are generally positive, as expected, but they are not very large in magnitude. For instance, the correlation between F_{ST} genetic distance and weighted linguistic distance is 22.7%. The two alternative measures of linguistic distance bear a correlation of 81%. Religious distance bears a correlation of 43.8% with linguistic distance, and 17.1% with genetic distance. These correlations are generally weaker when considering measures relative to the US, rather than absolute distances. In keeping with our theory, we use the measures relative to the US in our empirical analysis (in past work, we used the absolute distance measures, and this did not impact our results).

Regression results. Tables 8 and 9 present results when including these cultural variables in our baseline regression. We first control for variables representing a pair's common historical experience. These are dummy variables for pairs that were ever part of the same country (for example Austria and Hungary), were ever in a colonial relationship, have shared a common colonizer since 1945 and are currently in a colonial relationship (such as France and French Polynesia). These variables all bear the expected signs and have statistically significant coefficients (Table 8, column 2). For instance having had a common colonizer or having been part of the same country are associated with smaller income differences. The inclusion of these variables in the regression does not affect the magnitude of the genetic distance effect.

Turning to linguistic and religious distance, in Table 8, columns 3 and 4, both linguistic distance and religious distance enter with the expected positive signs and are statistically significant at the 5% level when entered individually. Their standardized betas are, respectively, 15.10% and 20.17%,

[59] An alternative classification obtained from the World Christian Database, with only 3 nested classifications, did not lead to appreciably different results.

so these variables can help account for a sizable fraction of the variation in income differences. When the two variables are entered jointly, only religious distance remains significant (column 5). More importantly from our perspective, the inclusion of these variables, either alone or together, slightly reduces the effect of genetic distance on income differences. Comparing column (5) with column (2), the reduction in the coefficient on genetic distance (and in the standardized beta), is 11.5%. This is very consistent with our story: when we can measure some vertically transmitted traits, in this case linguistic and religious distance, we obtain a reduction in the size of the coefficient on genetic distance, suggesting that genetic distance was capturing part of the effect of these vertical characteristics. The reduction is not large enough to suggest that genetic distance *only* captures the effect of linguistic and religious distance, but it is large enough to provide some support to the idea that genetic distance captures a wide array of traits, norms, values, and other slow-moving vertically transmitted traits (including language and religion) that hinder the diffusion of development.

The specific measures of linguistic and religious distance used in Table 8 may include a lot of measurement error, a point confirmed by Fearon (2006) in a short paper that suggests, for European countries, that genetic distance is robust to the inclusion of a variety of measures of linguistic distance in a regression seeking to explain income differences. Consider linguistic distance. It is well-known that linguistic and genetic trees look very similar, as demonstrated in Cavalli-Sforza et al. (1994, p. 98-105). The reason is straightforward: genes, like languages, are transmitted intergenerationally, and this insight is the basis for our interpretation of genetic distance as capturing the full set of vertical characteristics. Yet Table 7 shows that the correlation between our measure of genetic distance and linguistic distance is quite low. This may be partly due to the fact that genetic distance is a *continuous* measure, whereas the number of nodes is a discrete measure of linguistic distance.[60]

The lexicostatistical measure may partly addresses this problem, since it is more continuous. Table 9 presents the results obtained using the weighted and plurality measures of linguistic distance based on the percentage of cognate words. Using these data comes at a cost of losing all of the non-Indoeuropean speaking countries.[61] To allow comparisons within a common sample, for each

[60] Populations may share few common nodes but linguistic splits occurred recently, in which case one is overestimating distance, or they may share lots of common nodes but the last split occurred a long time ago, in which case one is underestimating distance.

[61] In addition, when using the weighted measure of lexicostatistical distance, we lose further countries with sizeable minorities of non-Indoeuropean speaking populations, such as India. For this variable, only 43 countries remain.

measure (weighted or not), we include a baseline regression controlling for geographic distance, transport costs and common history variables, for the sample for which the lexicostatistical measures are available. We find results broadly consistent with the ones obtained using Fearon's discrete measure of linguistic distance. For instance, comparing columns 3 and 4 of Table 9, for which most observations are available, the effect of genetic distance falls by 12.5% when controlling for lexicostatistical distance.

To conclude, using the best available measures of linguistic and religious distance, we are able to reduce the effect of genetic distance by $10 - 15\%$, but the effect remains large and significant.

4.7 Historical Income Data

In this subsection we examine the time variation in the effect of genetic distance in the 500 years that surrounded the Industrial Revolution. We find a pattern of coefficients supportive of our model of diffusion. In Table 10, we use income per capita data since 1500 from Maddison (2003), and repeat our basic reduced form regression for 1500, 1700, 1820, 1870, 1913 and 1960.[62] Our measure of genetic distance is now F_{ST} genetic distance between plurality groups, relative to the English population.[63] This is both the group to which the plurality genetic group in the US is matched for the modern period, and (conveniently) the group located in the birthplace of the Industrial Revolution. For the 1500 and 1700 regressions, we use the early match for genetic distance, i.e. genetic distance between populations as they were in 1492, prior to the discovery of the Americas and the great migrations of modern times.[64] For the subsequent periods we use the current match.

Table 10 shows that across all periods, the coefficient on relative genetic distance is statistically

These are mostly countries in Europe and the Americas.

[62] These dates were chosen to maximize the number of observations. The data on income for 1960 is from the Penn World Tables version 6.1. For comparison, column 7 reproduces the results for 1995.

[63] We also used Italy as the refernce point for the early periods, since there is evidence that Italy was th etechnological leader in Europe during the Renaissance. This led to no appreciable difference in the results. The Italians and the English are genetically very close relative to average worldwide genetic distance - the genetic distance between the English and the Italians is 0.0072 while the average genetic distance between World populations is 0.13.

[64] Regressions for these early periods feature at most 29 countries. These countries are Australia, Austria, Belgium, Brazil, Canada, China, Denmark, Egypt, Finland, France, Germany, Greece, India, Indonesia, Ireland, Italy, Japan, Korea, Mexico, Morocco, Netherlands, New Zealand, Norway, Portugal, Spain, Sweden, Switzerland, the United Kingdom and the United States. There were 325 pairs (26 countries) with available data for 1500 income, and 406 pairs (29 countries) for 1700.

significant and positive. Moreover, the magnitudes are much larger than for the current period: in regressions obtained from a common sample of 26 countries (325 pairs) for which data is continuously available, standardized beta coefficients range from 25.35% (in 1995) to 45.01% (in 1500). Thus, genetic distance is strongly positively correlated with income differences throughout modern history. It is worth noting that genetic distance bears a large, positive and significant effect on income differences for the past five centuries, even though income differences in 1500 and in 1995 are basically uncorrelated (Table 2 shows this correlation to be -0.051 for the 325 country pairs for which data is available). This fact is consistent with our interpretation of genetic distance as a barrier to the diffusion of innovations across populations: genetic distance remains significant despite significant reversals of fortune since 1500.

The time pattern of the effect in the common sample of 26 countries provides additional clues that support our interpretation. The standardized beta on genetic distance decreases from 1500 to 1820, then increases significantly in 1870 during the Industrial Revolution, and declines gradually thereafter. The shape of such time path during the 19th century is consistent with the view that the effect captures the diffusion of economic development from the world technological frontier - in particular, the gradual spread of the Industrial Revolution: a major shift in the growth regime (the Industrial Revolution) initially results in large income discrepancies. These discrepancies persist in proportion to genealogical relatedness. As more and more countries adopt the major innovation, the impact of genetic distance progressively declines.[65] At the same time, the slight decrease of the effect in recent times suggests that the impact of genetic distance may progressively decline, as more and more countries adopt the frontier innovations, and/or intersocietal barriers to the diffusion of development decrease through globalization and other forces.[66]

[65] In terms of the comparative statics of our simplified reduced-form model, an increase in the effect of genetic distance on income differences may be expected right after a big jump in the parameter Δ at the technological frontier. A more general interpretation is that the effect should increase after a series of big positive shocks to technology at the frontier, including possibly to the R&D technology itself. See the discussion in Howitt and Meyer-Foulker (2005). An analytical formalization of these ideas within a dynamic extension of our framework is available upon request.

[66] Within our simplified model, globalization and other forces that reduce inter-societal barriers can be interpreted as a reduction in the parameter β.

4.8 Genetic Distance across European Countries

As the last step in our empirical investigation, we provide a detailed analysis of the Europe sample. Analyzing the European data can be informative for several reasons. First, it constitutes a robustness check on the worldwide results. Second, matching populations to countries is much more straightforward for Europe than for the rest of the world, because the choice of sampled populations happens to match nation-state boundaries. This should reduce the incidence of measurement error. Third, genetic distances are orders of magnitude smaller across countries of Europe, and genetic specificities there have developed over the last few thousand years (and not tens of thousands of years). It is very unlikely that any genetic traits have risen to prominence within Europe as the result of strong natural selection over such a short period of time, so a finding that genetic distance based on neutral markers within Europe is associated with income differences would be evidence that barriers to the diffusion of development are primarily induced by differences in culturally transmitted traits.

In addition, existing studies using genetic distance in the economics literature use the European data exclusively (this is the case in Guiso et al., 2006, Giuliano et al., 2006 and Fearon, 2006). In particular, Giuliano et al. (2006) argue that the effect of genetic distance on bilateral trade across countries of Europe is sensitive to the inclusion of measures of geographic isolation and transport costs. Our application is different - we study income differences and not trade - and we have already shown that the effect of genetic distance is robust to controlling for geographic isolation and transport costs in the World sample. As we will see, the effect of genetic distance on income differences is robust to geographic and transportation-cost controls in the European dataset also.

We maintain the choice of the US as the frontier country. This requires us to use measures of genetic distance based on plurality groups, because we lack the data to calculate weighted genetic distance from European countries to the US.[67] To maintain consistency throughout, we also use measures of linguistic and religious distance based on plurality languages and plurality religions (this choice does not matter in terms of our empirical results). Since the plurality population of the US is matched to the English population, genetic distance is entered relative to the English. Similarly, linguistic distance is relative to the English language, and religious distance is relative to

[67] We do not have data on genetic distance between West Africans, Central Ameridians and Chinese, on the one hand, and European populations at the level of precision of the European dataset, on the other. These would be required to compute weighted genetic distance from European countries to the US.

Lutherans.[68]

Summary statistics. Table 11 presents summary statistics for the Europe sample. Several observations are in order. First, as stressed by Giuliano et al. (2006), genetic distance is indeed correlated with geodesic distance and transportation costs (geodesic distance and freight costs themselves bear a 96.8% correlation with each other). This remains true, though the correlations are weaker, when considering relative genetic distance. Second, even more so than in the World sample, relative genetic distance is only weakly correlated with relative linguistic and religious distances. Third, the correlation of absolute log income differences with absolute genetic distance (32.8%) is smaller that its correlation with relative genetic distance to the English (40.9%): Implication 2 of our model holds in the European sample as well.

Univariate regressions and geographic controls In Table 12 we present univariate regressions and regressions controlling for geography and transport costs. Genetic distance is again positively and significantly associated with income differences. Columns 1 and 2 confirm empirically Implication 2 of our model - the coefficient on relative genetic distance is about 30% larger than the coefficient on absolute genetic distance. These effects also are about 30% larger in magnitude than the corresponding effects found in the World sample (Table 3, columns 1-4). While genetic distance across European countries are smaller than in the World sample, so are the income differences to be explained.

We then add distance metrics as defined in Section 4.4.[69] The main direction of geographic inequality in Europe seems to be longitudinal - between the former Soviet Bloc countries and the West. We then add a large set of micro-geography controls. This set includes *exactly* the variables used in Giuliano et al. (2006), namely: a dummy variable taking a value of 1 if countries in a pair share access to the same sea or ocean; a variable that measures the average elevation of countries

[68] Our results are robust to using Germany or the UK as the frontier countries instead of the US. Results with the Germany and UK baseline are available upon request. It is not surprising that the results using the UK as the baseline would be similar to those using the US (the genetic and linguistic groups are the same, and only the religious plurality groups differ). Germans are very genetically close to the English, and like the US the plurality religion is Lutheran, so again the results change little when using Germany as the baseline. In fact, any genetic group in Northwestern Europe is close to the English.

[69] Again, introducing distance metrics and freight costs relative to the frontier (either the UK, the US or Germany) did not change our results concerning genetic distance. These regressions are available upon request.

that lies on the direct path between two countries (for instance the average elevation between France and Austria is the average elevation of France, Germany and Austria); and their measure of freight costs from the Import Export Wizard. In addition to these controls, we included additional measures of isolation: a dummy for contiguity, a dummy taking a value of 1 if at least one country in a pair is landlocked, and a similar dummy variable for islands. Together, the inclusion of these variables reduces the standardized effect of relative genetic distance from 41.08% to 34.61%, but the effect remains larger than in the World sample, and statistically significant at the 5% level (despite the much smaller sample). While it is clearly crucial to control for geographic factors here, these do not seem to play nearly as important a role on income differences as they apparently do on bilateral trade.

The last column of Table 12 uses income differences in 1870 as the dependent variable. While we lose 7 countries for lack of income data in the Maddison source, the results on genetic distance relative to the English are quantitatively and statistically much stronger than those for the contemporary period, consistent with the findings reported in Section 4.7. Upon impact, a major new innovation such as the Industrial Revolution diffuses in proportion to how genetically distant countries are from the frontier.[70]

Controlling for cultural distance In a short comment on our work, Fearon (2006) examined the interrelationships between genetic and linguistic distance within Europe. Regressing income levels on genetic distance from the English, geodesic distance to the UK and linguistic distance from the English language (using both his data based on linguistic trees and the lexicostatistical data), he found that genetic distance was generally robust to the inclusion of these variables.[71] We reexamine this issue using our bilateral methodology and with our full set of geographic controls.

[70] It is convenient (and surely not a coincidence) that the baseline population for calculating relative genetic distance, the English, is the plurality group both in the US (the frontier in 1995) and in the birthplace of the Industrial Revolution (the frontier in 1870).

[71] Fearon (2006) presents one regression with 22 observations using the lexicostatistical data where t-statistic on the coefficient on genetic distance to the English falls to 1.5. We have replicated this regression with all 23 European countries for which the lexicostatistical and linguistic trees data are available, and found that genetic distance remained statistically significant at the 5% level (Iceland, Hungary and Finland are missing from these regressions due to lack of data on one or the other linguistic distance measure). These results are available upon request.

Table 13 presents the results.[72] The bottom line is that the magnitude of the genetic distance effect is not affected by the inclusion of the cultural variables. Moreover, these variables themselves generally do not enter significantly, in contrast with the results obtained in the world sample.

5 Conclusion

In this paper we make two contributions: (1) For the first time, we document a statistically and economically significant positive relationship between measures of genetic distance and cross-country income differences, even when controlling for measures of geographical and climatic distances, transportation costs, linguistic distances, and other cultural and historical differences. (2) We provide an economic interpretation of these findings, in terms of barriers to the diffusion of development from the world technological frontier.

Our interpretation is based on two main ideas. The first idea is that, on average, genetic distance captures divergence in characteristics that are transmitted across generations within populations over the very long run. Genetic distance, measuring the time since two populations shared common ancestors, provides an ideal summary of divergence in slowly-changing genealogically-transmitted characteristics, including habits, customs, etc. Our second main idea is that such differences in long-term vertically-transmitted characteristics act as barriers to the diffusion of development from the world technological frontier.

The empirical evidence over time and space is consistent with the barriers interpretation. In line with our framework, the effect on economic distance associated with *relative* genetic distance from the world technological frontier is larger than the effect of absolute genetic distance. We also found that the effect has varied across time and space in ways that supports our diffusion-from-the-frontier interpretation: the effect has increased in the first part of the 19th century, peaked in 1870, and slightly decreased afterwards, consistent with the view that relative genetic distance captures barriers to the diffusion of the Industrial Revolution. Some evidence, particularly the results for European countries, suggests that these differences may stem in substantial part from cultural (rather than purely biological) transmission of characteristics across generations.

[72] While our baseline sample features 26 countries, we lack observations on linguistic and religious distance for Iceland, and we lack lexicostatistical data for Hungary and Finland, where Indoeuropean languages are not spoken. As a result, columns 1 and 5 present baseline regressions including geographic controls only, for comparison.

While our analysis provides a general macroeconomic framework to interpret our empirical findings, the study of the specific microeconomic mechanisms through which the effects operate is left for future research. An analysis of microeconomic data may shed light on the relations among genetic distance, vertical characteristics, imitation costs, and the spread of specific innovations.[73] Interestingly, we have not found that linguistic or religious distances, two culturally transmitted characteristics, are greatly responsible for the effect of genetic distance on income differences. This does not preclude a role for other slow-changing traits, such as norms or values.[74] These are inherently harder to measure, particularly within the long- term macroeconomic perspective that we have adopted, necessitating a more microeconomic approach. Another natural extension of our work would be to investigate whether and how genetic distance affects bilateral exchanges and interactions between different groups and societies, both peacefully (e.g., trade, foreign direct investment) and non-peacefully (conflict and wars).[75] Finally, it would be interesting to link our results to the vast literature on demography and economic growth, and explore the connections between genetic distance, intergenerationally-transmitted characteristics, and demographic patterns.[76]

A final consideration is about policy implications. A common concern with research documenting the importance of variables such as genetic distance or geography is pessimism about policy implications. What use is it to know that genetic distance explains income differences, if one cannot change genetic distance, at least in the short run? These concerns miss a bigger point: available policy variables may have a major impact not on genetic distance itself, but on the coefficients that measure the effect of genetic distance on income differences. Those coefficients have been changing over time, and can change further. If we are correct in interpreting our results as evidence for long-term barriers across different societies, significant reductions in income disparities could be obtained by encouraging policies that reduce such barriers, including efforts to translate and adapt technological and institutional innovations into different histories and traditions, and to fos-

[73] For instance, recent microeconomic analysis of international technological diffusion finds an important role for ethnic scientific communities, consistent with our interpretation (Kerr, 2007).

[74] Desmet et al. (2007) document strong relationships between responses to world-value-survey questions and genetic distance in a sample of European countries.

[75] This question is more closely related to recent work by Guiso et al. (2004) and Giuliano et al. (2006).

[76] For instance, Coale and Cotts Watkins (1986) documented the correlation between cultural similarity and the time paths of fertility across Europe (see also Richerson and Boyd, 2004, chapter 5). Galor (2005) provides an in-depth discussion of the economic literature on demographics and growth.

ter cross-societal exchanges and openness. More work is needed - at the micro as well as macro level - in order to understand the specific mechanisms, market forces, and policies that could facilitate the diffusion of development across countries with distinct long-term histories.

References

Acemoglu, Daron, Simon Johnson and James A. Robinson (2001), "The Colonial Origins of Comparative Development: An Empirical Investigation," *American Economic Review*, 91, 1369-1401.

Acemoglu, Daron, Simon Johnson and James A. Robinson (2002), "Reversal of Fortune: Geography and Institutions in the Making of the Modern World Income Distribution," *Quarterly Journal of Economics*, 117, 1231-1294.

Alcalá, Francisco and Antonio Ciccone (2004), "Trade and Productivity," *Quarterly Journal of Economics*, 119, 612-645.

Alesina, Alberto, Arnaud Devleeschauwer, William Easterly, Sergio Kurlat and Romain Wacziarg (2003), "Fractionalization," *Journal of Economic Growth*, 8, 55-194.

Barro, Robert J. and Xavier Sala-i-Martin (1997), "Technological Diffusion, Convergence and Growth," *Journal of Economic Growth*, 2, 1-26.

Bisin, Alberto and Thierry Verdier (2000), "Beyond the Melting Pot: Cultural Transmission, Marriage, and the Evolution of Ethnic and Religious Traits," *Quarterly Journal of Economics*, 105, 955-988.

Bisin, Alberto and Thierry Verdier (2001), "The Economics of Cultural Transmission and the Evolution of Preferences," *Journal of Economic Theory*, 97, 98-319.

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

Boyd, Robert and Joan B. Silk (2003), *How Humans Evolved* (New York: W. W. Norton & Company).

Brezis, Elise, Paul Krugman and Daniel Tsiddon (1993), "Leapfrogging in International Competition: A Theory of Cycles in National Technological Leadership," *American Economic Review*, 83, 1211–1219.

Brock, William A. and Anastasios Xepapadeas (2003), "Valuing Biodiversity from an Economic Perspective: A Unified Economic, Ecological, and Genetic Approach," *American Economic Review*, 93, 1597-1614.

Cameron, A. Colin, Jonah B. Gelbach and Douglas L. Miller (2006), "Robust Inference with Multi-Way Clustering", NBER Technical Working Paper #T0327, September.

Case, Anne (1991), "Spatial Patterns in Household Demand," *Econometrica*, 59, 953-965.

Caselli, Francesco and John Coleman (2006), "On the Theory of Ethnic Conflict", unpublished, LSE and Duke University.

Cavalli-Sforza, Luigi L., and Francesco Cavalli-Sforza (1995), *The Great Human Diasporas* (Reading, MA: Addison Wesley Publishing Company).

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

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

Clark, Gregory (2007), *A Farewell to Alms. A Brief Economic History of the World* (Princeton: Princeton University Press).

Clark, Gregory and Susan Wolcott (1999), "Why Nations Fail: Managerial Decisions and Performance in Indian Cotton Textiles, 1890-1938," *Journal of Economic History*, 59, 397-423.

Coale, Ansley and J. and Susan Cotts Watkins (1986), *The Decline of Fertility in Europe* (Princeton, NJ: Princeton University Press).

Dawkins, Richard (2004), *The Ancestor's Tale: A Pilgrimage to the Dawn of Evolution* (Boston, MA: Houghton Mifflin).

Desmet, Klaus, Michel Le Breton, Ignacio Ortuno Ortin and Shlomo Weber (2006), "Nation Formation and Genetic Diversity," CEPR Discussion Paper 5918, November.

Diamond, Jared (1992), *The Third Chimpanzee. The Evolution and Future of the Human Animal* (New York: Harper Collins).

Diamond, Jared (1997), *Guns, Germs, and Steel: The Fates of Human Societies* (New York: W. W. Norton & Company).

Dyen, Isidore, Joseph B. Kruskal and Paul Black (1992), "An Indoeuropean Classification: A Lexicostatistical Experiment," *Transactions of the American Philosophical Society*, 82, .1-132.

Easterly, William and Ross Levine (2003), "Tropics, Germs, and Crops: How Endowments Influence Economic Development," *Journal of Monetary Economics*, 50, 3-39.

Fearon, James (2003), "Ethnic and Cultural Diversity by Country," *Journal of Economic Growth,* 8, 195-222.

Fearon, James (2006), "Is Genetic Distance a Plausible Measure of Cultural Distance?" unpublished, Stanford University, June.

Gallup, John L., Andrew D. Mellinger and Jeffrey D. Sachs (1998), “Geography and Economic Development," NBER Working Paper No. 6849.

Galor, Oded and Omer Moav (2002), "Natural Selection and the Origin of Economic Growth," *Quarterly Journal of Economics*, 117, 1133-1191.

Galor, Oded (2005), "From Stagnation to Growth: Unified Growth Theory", in Philippe Aghion and Steven Durlauf, eds, *Handbook of Economic Growth (* Amsterdam: North-Holland).

Giuliano, Paola, Antonio Spilimbergo and Giovanni Tonon (2006), "Genetic, Cultural and Geographical Distances," unpublished, International Monetary Fund.

Glaeser, Edward, Rafael LaPorta, Florencio Lopez-de-Silanes and Andrei Shleifer (2004), "Do Institutions Cause Growth?", *Journal of Economic Growth*, 9, 271-303.

Guiso, Luigi, Paola Sapienza and Luigi Zingales (2004), "Cultural Biases in Economic Exchange," NBER Working Paper No. 11005, December.

Hall, Robert and Jones, Charles I. (1999), "Why Do Some Countries Produce so Much More Output Per Worker than Others?", *Quarterly Journal of Economics*, 114, 83-116.

Harrison, Lawrence E. and Samuel P. Huntington, eds. (2000), *Culture Matters: How Values Shape Human Progress* (New York: Basic Books).

Heston, Alan, Robert Summers and Bettina Aten (2002), "Penn World Table Version 6.1", *Center for International Comparisons at the University of Pennsylvania* (CICUP), October.

Howitt, Peter and David Mayer-Foulkes (2005), "R&D, Implementation, and Stagnation: A Schumpeterian Theory of Convergence Clubs," *Journal of Money, Credit and Banking*, 37, 147-77.

Hummels, David, and V. Lugovskyy (2006) “Usable Data? Matched Partner Trade Statistics as a Measure of Transportation Cost”, *Review of International Economics*, 14, 69-86.

Jobling, Mark A., Matthew E. Hurles and Chris Tyler-Smith (2004), *Human Evolutionary Genetics. Origins, Peoples & Diseases* (New York: Garland Science).

Kerr, William R. (2007), "Ethnic Scientific Communities and International Technology Diffusion," *Review of Economics and Statistics*, forthcoming.

Kimura, Motoo (1968), "Evolutionary Rate at the Molecular Level", *Nature*, 217, 624-626.

Kroeber, Alfred and Clyde Kluckhohn (1952), *Culture* (New York: Meridian Books).

Kremer, Michael (1993), "Population Growth and Technological Change: 1,000,000 B.C. to 1990", *Quarterly Journal of Economics*, 108,. 681-716

Limao, Nuno and Anthony Venables (2001), “Infrastructure, Geographical Disadvantage, Transport Costs and Trade”, *World Bank Economic Review*, 15, 451-79.

Masters, William A and Margaret S. McMillan (2001), "Climate and Scale in Economic Growth," *Journal of Economic Growth*, 6, .167-186.

Maddison, Angus (2003), *The World Economy: Historical Statistics* (Paris: OECD Development Center).

Mecham, R. Quinn, James Fearon and David Laitin (2006), "Religious Classification and Data on Shares of Major World Religions," *unpublished*, Stanford University.

McNeill, John Robert and William H. McNeill (2003), *The Human Web: A Bird’s Eye View of Human History* (New York: W. W. Norton & Co).

Olsson, Ola and Douglas A. Hibbs Jr. (2005), "Biogeography and Long-Run Economic Development," *European Economic Review*, 49, 909-938.

Parente, Stephen L. and Edward C. Prescott (1994), "Barriers to Technology Adoption and Development", *Journal of Political Economy*, 102, 298-321.

Parente, Stephen L. and Edward C. Prescott (2002), *Barriers to Riches*, Cambridge: MIT Press

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

Rogers, Everett M. (1962), *Diffusion of Innovations*, (New York: The Free Press), first edition (fifth edition: 2003).

Sachs, Jeffrey (2001), "Tropical Underdevelopment," *NBER Working Paper* 8119.

Shennan, Stephen (2002), *Genes, Memes and Human History. Darwinian Archaeology and Cultural Evolution* (London: Thames and Hudson).

Tabellini, Guido (2005), "Culture and Institutions: Economic Development in the Regions of Europe", *IGIER Working Paper* #292, June.

Wang, E.T., G. Kodama, P. Baldi, and R.K. Moyzis (2006), "Global Landscape of Recent Inferred Darwinian Selection for Homo Sapiens", *Procedeeings of the Natural Academy of Sciences*, January 3, 135-140.

Weitzman, Martin (1992), "On Diversity," *Quarterly Journal of Economics*, 107, 363-405.

Whitehead, Alfred North (1926), *Science and the Modern World* (Cambridge: Cambridge University Press).

Appendix A: Definition of FST

In this Appendix we illustrate the construction of F_{ST} for the simple case of two populations (a and b) of equal size, and one locus (that is, only one "genetic characteristic") with two variants (allele 1 and allele 2). That is, populations can differ genetically only one way, Let p_a and $q_a = 1-p_a$ be the gene frequency of allele 1 and allele 2, respectively, in population a.[77] The probability that two randomly selected alleles at a given locus are *identical* within the population (homozygosity) is $p_a^2 + q_a^2$, and the probability that they are different (heterozygosity) is:

$$h_a = 1 - \left(p_a^2 + q_a^2\right) = 2p_a q_a \tag{13}$$

By the same token, heterozygosity in population b is:

$$h_b = 1 - \left(p_b^2 + q_b^2\right) = 2p_b q_b \tag{14}$$

where p_b and $q_b = 1 - p_b$ are the gene frequency of allele 1 and allele 2, respectively, in population b. The average gene frequencies of allele 1 and 2 in the two populations are, respectively:

$$\overline{p} = \frac{p_a + p_b}{2} \tag{15}$$

and:

$$\overline{q} = \frac{q_a + q_b}{2} = 1 - \overline{p} \tag{16}$$

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

$$h = 1 - \left(\overline{p}^2 + \overline{q}^2\right) = 2\overline{pq} \tag{17}$$

Average heterozygosity is measured by:

$$h_m = \frac{h_a + h_b}{2} \tag{18}$$

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

$$F_{ST} = 1 - \frac{h_m}{h} = 1 - \frac{p_a q_a + p_b q_b}{2\overline{pq}} = (1/4)\frac{(p_a - p_b)^2}{\overline{p}(1-\overline{p})} \tag{19}$$

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 $p_a = 0$

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

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 et al. (1994, pp. 26-27).

Appendix B: The Presence of Spatial Correlation

Consider three countries, 1, 2 and 3. Observations on the dependent variable $|\log y_1 - \log y_2|$ and $|\log y_1 - \log y_3|$ will be correlated by virtue of the presence of country 1 in both observations. Conditioning on the right-hand side variables (which are bilateral in nature) should reduce cross-sectional dependence in the errors ε_{12} and ε_{13}, but we are unwilling to assume that observations on the dependent variable are independent conditional on the regressors.[78] In other words, simple least squares standard errors will lead to misleading inferences due to spatial correlation.

Before proceeding, we note the following: with N countries, there are $N(N-1)/2$ distinct pairs. Denote the observation on absolute value income differences between country i and country j as dy_{ij}. Pairs are ordered so that country 1 appears in position i and is matched with all countries from $2...N$ appearing in position j. Then country 2 is in position i and is matched with $3...N$ appearing in position j, and so on. The last observation has country $N-1$ in position i and country N in position j. We denote the non-zero off-diagonal elements of the residual covariance matrix by σ_m where m is the country common to each pair.

A simple example when the number of countries is $N = 4$ is illustrative. In this case, under our maintained assumption that the error covariances among pairs containing a common country m

[78] Another feature that reduces the dependence across pairs is the fact that the dependent variable involves the *absolute value* of log income differences. Simple simulations show that under i.i.d. Normal income draws with moments equal to those observed in our sample, the correlation between absolute value differences in income for any two pair containing the same country will be about 0.22. Without taking absolute values, it is straightforward to notice that the correlation would be exactly 0.5. Hence, taking absolute values reduces the cross-sectional dependence induced by the construction of the dependent variable..

are equal to a common value σ_m, the covariance matrix of the vector of residuals ε is of the form:

$$\Omega = cov \begin{pmatrix} \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{14} \\ \varepsilon_{23} \\ \varepsilon_{24} \\ \varepsilon_{34} \end{pmatrix} = \begin{pmatrix} \sigma_\varepsilon^2 & & & & & \\ \sigma_1 & \sigma_\varepsilon^2 & & & & \\ \sigma_1 & \sigma_1 & \sigma_\varepsilon^2 & & & \\ \sigma_2 & \sigma_3 & 0 & \sigma_\varepsilon^2 & & \\ \sigma_2 & 0 & \sigma_4 & \sigma_2 & \sigma_\varepsilon^2 & \\ 0 & \sigma_3 & \sigma_4 & \sigma_3 & \sigma_4 & \sigma_\varepsilon^2 \end{pmatrix}$$

This clearly demonstrates the presence of cross-sectional (spatial) correlation. It is important to note however that our data are not linearly dependent, i.e. there is additional information brought in by considering the bilateral approach. One major reason is that the dependent variable is the absolute difference in log income, not just the difference in log income. It is easy to show that taking absolute values greatly reduces spatial dependence in the dependent variable. Another major reason is that we are conditioning on right hand side variables (such as geodesic distance, genetic distance, etc.) that are truly bilateral in nature, i.e. our empirical model is *not* the result of simply differencing a "level" specification across cross-sectional units.

Table 1 - Income level regressions, World dataset
Dependent variable: Log income per capita 1995

	(1)	**(2)**	**(3)**
	Univariate	**Add geographic distance**	**Add cultural distance**
Fst genetic distance **to the USA, weighted**	**-12.906** **(1.383)****	**-12.523** **(1.558)****	**-10.245** **(1.567)****
Absolute difference in latitudes with the USA		1.970 (0.868)**	1.518 (0.827)*
Absolute difference in longitudes with the USA		0.438 (0.454)	0.786 (0.401)*
Geodesic distance from the USA (1000s of km)		-0.179 (0.075)**	-0.191 (0.071)**
1 for contiguity with the USA		1.055 (0.300)**	0.452 (0.390)
=1 if the country is an island		0.505 (0.397)	0.362 (0.483)
=1 if the country is landlocked		-0.384 (0.206)*	-0.410 (0.198)**
=1 if the country shares at least one sea or ocean with the USA		-0.201 (0.197)	-0.080 (0.171)
Freight rate to Northeastern USA (surface transport)		3.460 (2.507)	5.794 (2.816)**
Linguistic Distance to the USA, weighted			-0.520 (0.648)
Religious Distance to the USA, weighted			-2.875 (0.591)**
Constant	9.421	8.876	10.499
	(0.149)**	(0.536)**	(0.751)**
Observations	137	137	137
Adjusted R-squared	0.39	0.46	0.53

Robust standard errors in parentheses; * significant at 10%; ** significant at 5%

Table 2 – Summary statistics for the main variables (World dataset)

Panel a. Simple correlations among genetic and economic distance measures

	F_{ST} Gen. Dist. weighted	FST Gen. Dist., weighted, Relative to USA	F_{ST} Gen. Dist., 1500 match	Nei Gen. Dist., weighted	Abs. log income diff. 1995	Abs. log income difference, 1870[a]
F_{ST} Genetic Distance, weighted, relative to USA	0.634	1				
F_{ST} Genetic Distance, 1500 match	0.827	0.544	1			
Nei Genetic Distance, weighted	0.939	0.742	0.782	1		
Abs. log income difference, 1995	0.197	0.337	0.203	0.231	1	
Abs. log income difference, 1870[a]	0.007	0.200	0.058	0.072	0.596	1
Abs. log income difference, 1500[b]	-0.088	-0.030	0.198	-0.068	-0.051	0.060

(number of observations: 9,316, except [a]: 1,326 and [b]: 325)

Panel b. Means and standard deviations

Variable	# of obs.	Mean	Std. dev.	min	max
F_{ST} Genetic Distance weighted	9,316	0.111	0.071	0.000	0.344
F_{ST} Genetic Distance, weighted, relative to USA	9,316	0.062	0.048	0.000	0.213
F_{ST} Genetic Distance, 1500 match	9,316	0.127	0.082	0.000	0.356
Nei Genetic Distance, weighted	9,316	0.018	0.013	0.000	0.059
Abs. log income difference, 1995	9,316	1.290	0.912	0.000	4.133
Abs. log income difference, 1870[b]	1,326	0.658	0.488	0.000	2.110
Abs. log income difference, 1500[a]	325	0.327	0.237	0.000	1.012

Table 3 - Univariate regressions (two-way clustered standard errors)
Dependent variable: absolute value of log income differences, 1995

	(1)	(2)	(3)	(4)	(5)	(6)
	FST Gen. Dist.	**FST Gen Dist, relative to US**	**Weighted FST Gen. Dist.**	**Weighted FST Gen. Dist., relative to US**	**Weighted Nei Gen. Dist.**	**Weighted regression**
Fst Genetic Distance	1.853 (0.508)**					2.214 (0.533)**
Fst genetic distance relative to the USA		3.541 (0.654)**				
Weighted Fst Genetic Distance			2.516 (0.630)**			
Fst gen. dist. relative to the USA, weighted				6.357 (0.996)**		
Weighted Nei Genetic Distance					16.868 (3.792)**	
Constant	1.079 (0.051)**	0.977 (0.049)**	1.010 (0.059)**	0.893 (0.052)**	0.986 (0.057)**	1.044 (0.050)**
Standardized Beta (%)	16.79%	26.98%	19.71%	33.65%	27.01%	20.07%
R-Squared	0.03	0.07	0.04	0.11	0.05	0.03

Two-way clustered standard errors in parentheses; * significant at 10%; ** significant at 5%
9,316 observations from 137 countries.

Table 4 - Controlling for geographic distance (two-way clustered standard errors)
Dependent variable: absolute value of log income differences, 1995

	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**
	Baseline	**Distance Metrics**	**Add micro-geography controls**	**Add transport costs**	**Continent dummies**	**Climatic difference control**	**Tropical difference control**
Fst gen. dist. relative to the USA, weighted	**6.357** **(0.996)****	**6.387** **(0.994)****	**6.273** **(0.989)****	**6.312** **(0.988)****	**4.134** **(1.046)****	**6.067** **(0.960)****	**6.368** **(1.003)****
Absolute difference in latitudes		0.523 (0.241)**	0.494 (0.238)**	0.494 (0.237)**	-0.228 (0.217)	0.254 (0.221)	0.497 (0.237)**
Absolute difference in longitudes		0.387 (0.235)*	0.391 (0.226)*	0.376 (0.224)*	0.084 (0.162)	0.257 (0.224)	0.380 (0.224)*
Geodesic Distance (1000s of km)		-0.050 (0.028)*	-0.057 (0.026)**	-0.081 (0.039)**	-0.008 (0.036)	-0.062 (0.038)*	-0.081 (0.039)**
1 for contiguity			-0.456 (0.064)**	-0.462 (0.064)**	-0.284 (0.060)**	-0.328 (0.061)**	-0.464 (0.064)**
=1 if either country is an island			0.178 (0.094)*	0.180 (0.094)*	0.119 (0.090)	0.162 (0.102)	0.181 (0.092)*
=1 if either country is landlocked			0.071 (0.076)	0.078 (0.076)	0.110 (0.071)	0.084 (0.076)	0.075 (0.075)
=1 if pair shares at least one sea or ocean			-0.029 (0.062)	-0.024 (0.062)	0.030 (0.050)	0.044 (0.059)	-0.024 (0.062)
Freight rate (surface transport)				1.282 (1.568)	-0.197 (1.517)	1.160 (1.490)	1.286 (1.583)
Climatic difference of land areas, by 12 KG zones						0.032 (0.007)**	
Difference in % land area in KG tropical climates							-0.033 (0.083)
Constant	0.893 (0.052)**	0.866 (0.066)**	0.889 (0.078)**	0.675 (0.263)**	1.919 (0.407)**	0.284 (0.263)	0.681 (0.264)**
Standardized Beta (%)	33.65%	33.81%	33.20%	33.41%	21.88%	32.12%	33.70%
R-Squared	0.11	0.12	0.13	0.13	0.22	0.15	0.13

Two-way clustered standard errors in parentheses; * significant at 10%; ** significant at 5%
9,316 observations from 137 countries in all columns
Column (5) includes two set of continent dummies (estimates not reported): a set of dummies each equal to 1 if both countries in a pair are on the same given continent; and a set of dummies each equal to one if exactly one country belongs to a given continent, and the other not. Continents are defined as Europe, Africa, Latin America, North America, Asia and Oceania.

Table 5 - Controlling for geographic distance relative to the USA (two-way clustered standard errors)
(Dependent variable: absolute value of log income differences, 1995)

	(1)	**(2)**	**(3)**	**(4)**
	Baseline	**Add micro-geography controls**	**Add transport costs**	**Continent dummies**
Fst gen. dist. relative to the USA, weighted	**6.357** **(0.996)****	**6.518** **(0.986)****	**6.533** **(0.982)****	**4.371** **(1.051)****
Latitude difference, relative to USA		-0.606 (0.178)**	-0.578 (0.180)**	-0.528 (0.180)**
Longitude difference, relative to USA		-0.140 (0.064)**	-0.168 (0.061)**	0.062 (0.113)
Geodesic Distance, relative to USA		0.020 (0.014)	0.004 (0.018)	-0.008 (0.016)
Freight cost (surface transport), relative to the USA			1.035 (0.623)*	0.581 (0.526)
Constant	0.893 (0.052)**	0.960 (0.084)**	0.936 (0.084)**	1.982 (0.268)**
Standardized Beta (%)	33.65%	34.50%	34.58%	23.14%
R-Squared	0.11	0.14	0.14	0.23

Two-way clustered standard errors in parentheses; * significant at 10%; ** significant at 5%
9,316 observations from 137 countries in all columns
All regressions include the following additional controls (estimates not reported): dummy for contiguity, dummy if either country is an island, dummy if either country is landlocked, dummy for common sea or ocean.
Column (5) includes two set of continent dummies (estimates not reported): a set of dummies each equal to 1 if both countries in a pair are on the same given continent; and a set of dummies each equal to one if exactly one country belongs to a given continent, and the other not. Continents are defined as Europe, Africa, Latin America, North America, Asia and Oceania.

Table 6 - Endogeneity of genetic distance and diamond gap (two-way clustered standard errors)
Dependent variable: absolute value of log income differences, 1995 (columns 1-3) or 1500 (column 4)

	(1) **2SLS with 1500 GD**	**(2)** **Without New World**	**(3)** **Diamond Gap, w/o New World**	**(4)** **Income 1500, Diamond Gap**
Fst genetic distance relative to the USA, weighted	**9.400** **(1.665)****	**4.428** **(1.252)****	**2.815** **(1.347)****	
Fst genetic distance relative to the English, 1500 match				**1.737** **(0.427)****
Absolute difference in latitudes	0.402 (0.293)	0.901 (0.420)**	1.078 (0.471)**	0.152 (0.138)
Absolute difference in longitudes	0.601 (0.246)**	0.349 (0.258)	0.781 (0.333)**	-0.007 (0.070)
Geodesic Distance (1000s of km)	-0.114 (0.039)**	-0.087 (0.051)*	-0.155 (0.055)**	-0.016 (0.022)
1 for contiguity	-0.381 (0.063)**	-0.471 (0.069)**	-0.461 (0.067)**	-0.048 (0.040)
=1 if either country is an island	0.209 (0.094)**	0.134 (0.113)	0.176 (0.115)	0.004 (0.053)
=1 if either country is landlocked	0.052 (0.076)	0.016 (0.081)	0.022 (0.076)	-0.059 (0.034)*
=1 if pair shares at least one sea or ocean	-0.043 (0.077)	-0.060 (0.087)	-0.071 (0.085)	-0.068 (0.047)
Freight rate (surface transport)	1.700 (1.341)	1.627 (1.809)	1.508 (1.781)	-0.263 (0.847)
Diamond gap			0.472 (0.137)**	0.164 (0.059)**
Constant	0.488 (0.241)**	0.701 (0.312)**	0.760 (0.309)**	0.338 (0.144)**
# of observations	9,316	6,105	6,105	325
# of countries	137	111	111	26
Standardized Beta (%)	49.75%	23.56%	14.98%	37.96%
R-Squared	0.10	0.11	0.13	0.22

Two-way clustered standard errors in parentheses; * significant at 10%; ** significant at 5%
The Diamond gap is a dummy variable that takes on a value of 1 if one and only one of the countries in each pair is located on the Eurasian landmass, and 0 otherwise.

Table 7 –Summary statistics for genetic distance and measures of cultural distance

Panel a: Correlations between genetic distance and various measures of cultural distance

	Weighted Fst genetic distance	**Weighted linguistic distance**	**Weighted religious distance**	**Weighted Fst gen. dist. relative to USA**	**Weighted ling. dist. relative to USA**	**Weighted religious distance relative to USA**	**1-% cognate, relative to USA, weighted [a]**
Weighted linguistic distance	0.227	1					
Weighted religious distance	0.171	0.438	1				
Weighted Fst genetic distance relative to USA	0.634	0.210	0.126	1			
Weighted linguistic distance relative to USA	-0.020	0.058	0.026	0.062	1		
Weighted religious distance relative to USA	0.052	0.143	0.343	0.061	0.459	1	
1 - % cognate, weighted [a]	-0.102	0.810	0.545	-0.026	0.143	0.190	1
1-% cognate, relative to USA, weighted [a]	0.122	0.376	0.337	0.198	0.697	0.312	0.473

(# of observations=9,316, except [a]: # of obs=903)

Panel b: Summary statistics for genetic distance and various measures of cultural distance

Variable	**Mean**	**Standard deviation**	**Minimum**	**Maximum**
Weighted Fst genetic distance	0.111	0.071	0.000	0.344
Weighted linguistic distance	0.968	0.106	0.000	1.000
Weighted religious distance	0.841	0.151	0.089	1.000
Weighted Fst genetic distance relative to USA	0.062	0.048	0.000	0.213
Weighted linguistic distance relative to USA	0.088	0.169	0.000	1.000
Weighted religious distance relative to USA	0.149	0.134	0.000	0.999

(# of observations=9,316)

Table 8 - Controlling for common history and cultural distance (two-way clustered standard errors)
Dependent variable: absolute value of log income differences, 1995

	(1)	**(2)**	**(3)**	**(4)**	**(5)**
	Baseline	**Colonial history controls**	**Linguistic distance, weighted**	**Religious distance, weighted**	**Religious + Linguistics, weighted**
Fst gen. dist. relative to the USA, weighted	**6.312** **(0.988)****	**6.283** **(0.988)****	**5.827** **(0.944)****	**5.702** **(0.950)****	**5.557** **(0.940)****
1 if countries were or are the same country		-0.217 (0.088)**	-0.223 (0.087)**	-0.209 (0.087)**	-0.214 (0.087)**
1 for pairs ever in colonial relationship		0.304 (0.131)**	0.255 (0.134)*	0.307 (0.101)**	0.282 (0.106)**
1 for common colonizer post 1945		-0.226 (0.066)**	-0.214 (0.063)**	-0.135 (0.060)**	-0.142 (0.059)**
1 for pairs currently in colonial relationship		-1.033 (0.193)**	-0.823 (0.200)**	-0.969 (0.167)**	-0.873 (0.176)**
Linguistic distance index, relative to USA, weighted			0.815 (0.204)**		0.409 (0.292)
Religious distance index, relative to USA, weighted				1.373 (0.266)**	1.172 (0.317)**
Constant	0.675 (0.263)**	0.740 (0.257)**	0.849 (0.246)**	0.703 (0.250)**	0.763 (0.246)**
Standardized Beta (%)	33.41%	33.26%	30.84%	30.18%	29.42%
R-Squared	0.13	0.14	0.16	0.17	0.18

Two-way clustered standard errors in parentheses; * significant at 10%; ** significant at 5%
9,316 observations from 137 countries.
All columns include geographic controls, i.e. absolute difference in latitudes, absolute difference in longitudes, geodesic distance, dummy for contiguity, dummy=1 if either country is an island, dummy=1 if either country is landlocked, dummy=1 if pair shares at least one sea or ocean, freight rate for surface transport (estimates not reported).

Table 9 - Controlling for lexicostatistical distance measures (two-way clustered standard errors)
Dependent variable: absolute value of log income differences, 1995

	(1)	(2)	(3)	(4)
	Baseline	**% cognate, weighted**	**Baseline**	**% cognate, dominant**
Fst gen. dist. relative to the USA, weighted	7.869 (2.521)**	7.591 (2.548)**	6.853 (2.116)**	5.995 (2.109)**
1 if countries were or are the same country	-0.187 (0.127)	-0.111 (0.122)	-0.262 (0.084)**	-0.196 (0.090)**
1 for pairs ever in colonial relationship	0.002 (0.126)	0.070 (0.147)	0.109 (0.112)	0.167 (0.119)
1 for common colonizer post 1945	-0.145 (0.198)	-0.089 (0.187)	-0.046 (0.111)	-0.038 (0.082)
1 for pairs currently in colonial relationship	a	a	-0.720 (0.162)**	-0.444 (0.179)**
1-% cognate, relative to USA, weighted		0.515 (0.235)**		
1-% cognate, relative to USA, plurality				0.631 (0.185)**
Constant	0.903 (0.286)**	0.767 (0.275)**	1.053 (0.179)**	0.928 (0.188)**
# observations	903	903	1,830	1,830
# countries	43	43	61	61
Standardized Beta (%)	30.91%	29.82%	23.58%	20.63%
R-Squared	0.16	0.19	0.14	0.19

Two-way clustered standard errors in parentheses; * significant at 10%; ** significant at 5%
All columns include geographic controls, i.e. absolute difference in latitudes, absolute difference in longitudes, geodesic distance, dummy for contiguity, dummy=1 if either country is an island, dummy=1 if either country is landlocked, dummy=1 if pair shares at least one sea or ocean, freight rate for surface transport (estimates not reported).

[a]: Dropped due to singularity (no observations with current colonial relationships in the subsample).

Table 10 - Regressions using historical income data (two-way clustered standard errors)
Dependent variable: difference in per capita income, in dates specified in row 2.

	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**
	Income 1500	**Income 1700**	**Income 1820**	**Income 1870**	**Income 1913**	**Income 1960**	**Income 1995**
Relative Fst genetic distance to the English, 1500 match	**2.059** **(0.479)****	**2.788** **(0.515)****					
Fst genetic distance relative to the English, weighted			**0.671** **(0.338)****	**1.684** **(0.846)****	**1.967** **(0.924)****	**3.503** **(0.784)****	**4.948** **(0.785)****
Absolute difference in latitudes	0.233 (0.124)*	0.690 (0.180)**	1.034 (0.199)**	1.188 (0.274)**	1.286 (0.263)**	1.201 (0.244)**	0.527 (0.245)**
Absolute difference in longitudes	0.040 (0.075)	0.164 (0.085)*	0.525 (0.144)**	0.692 (0.267)**	0.950 (0.261)**	0.742 (0.248)**	0.420 (0.239)*
Geodesic Distance (1000s of km)	-0.015 (0.023)	-0.082 (0.025)**	-0.096 (0.033)**	-0.124 (0.064)*	-0.175 (0.084)**	-0.171 (0.065)**	-0.087 (0.040)**
1 for contiguity	-0.051 (0.047)	-0.168 (0.054)**	-0.226 (0.053)**	-0.257 (0.048)**	-0.272 (0.049)**	-0.102 (0.064)	-0.466 (0.064)**
=1 if either country is an island	-0.069 (0.031)**	0.003 (0.044)	-0.067 (0.029)**	0.070 (0.099)	0.062 (0.088)	-0.010 (0.073)	0.177 (0.095)*
=1 if either country is landlocked	-0.042 (0.032)	-0.018 (0.063)	0.136 (0.037)**	0.173 (0.078)**	0.217 (0.081)**	0.125 (0.089)	0.083 (0.075)
=1 if pair shares at least one sea or ocean	-0.027 (0.045)	-0.045 (0.067)	-0.009 (0.042)	0.070 (0.049)	0.118 (0.047)**	0.082 (0.061)	-0.011 (0.067)
Freight rate (surface transport)	-0.005 (0.863)	1.727 (1.187)	1.098 (1.045)	1.668 (2.859)	3.072 (4.374)	4.600 (3.331)	1.362 (1.584)
Constant	0.249 (0.149)*	0.051 (0.170)	0.139 (0.192)	0.100 (0.471)	-0.066 (0.689)	-0.250 (0.526)	0.663 (0.267)**
# observations	325	406	1,035	1,485	1,653	4,753	9,316
# countries	26	29	46	55	58	98	137
Standardized Beta (%)	45.01%	40.48%	8.74%	14.95%	14.89%	29.07%	32.84%
Standardized Beta (%) (common sample of 325 obs.)	45.01%	40.58%	10.20%	32.63%	29.71%	26.24%	25.35%
R-Squared	0.18	0.24	0.23	0.16	0.17	0.17	0.13

Two-way clustered standard errors in parentheses; * significant at 10%; ** significant at 5%

Table 11 – Summary statistics for the European dataset

a. Correlations between the main variables

	Abs. log income diff., 1995	FST Genetic Distance	FST Genetic distance, rel. to the English	Geodesic Distance	Freight cost (surface transport)	Linguistic dist., rel. to the English language
FST Genetic Distance	0.328	1				
FST Genetic Distance, relative to the English population	0.409	0.647	1			
Geodesic Distance	0.076	0.433	0.260	1		
Freight cost (surface transport)	0.119	0.463	0.303	0.968	1	
Linguistic distance, relative to the English language	-0.068	-0.123	0.066	-0.032	-0.030	1
Religious distance, relative to the Lutheran religion	0.030	-0.053	0.002	0.037	0.059	0.047

300 observations from 25 countries

b. Summary statistics

Variable	Mean	Std. Dev.	Min	Max
Abs. log income difference, 1995	0.671	0.579	0.004	2.592
FST Genetic Distance	0.009	0.006	0.000	0.029
FST Genetic Distance, relative to the English population	0.006	0.005	0.000	0.020
Geodesic Distance	1.309	0.689	0.060	3.913
Freight cost (surface transport)	0.183	0.014	0.159	0.237
Linguistic distance, relative to the English language	0.108	0.251	0.000	1.000
Religious distance, relative to the Lutheran religion	0.210	0.206	0.000	1.000

300 observations from 25 countries

Table 12 - Results for the European dataset (two-way clustered standard errors)
Dependent variable: Difference in log per capita income across pairs (in 1995 for columns 1-4, in 1870 for column 5)

	(1)	**(2)**	**(3)**	**(4)**	**(5)**
	No controls, simple GD	**No controls, relative GD**	**Add distance metrics**	**Add micro-geography**	**1870 Income data**
Fst genetic distance in Europe	**28.134 (14.605)***				
Genetic distance, relative to the English		**45.222 (22.193)****	**46.685 (21.359)****	**39.333 (18.708)****	**39.842 (11.052)****
Absolute difference in latitudes			-0.596 (0.616)	-0.800 (0.723)	0.467 (1.397)
Absolute difference in longitudes			0.268 (0.138)*	0.233 (0.129)*	1.022 (1.171)
Geodesic Distance (1000s of km)			-0.026 (0.077)	-0.344 (0.306)	-0.150 (0.177)
1 for contiguity				-0.136 (0.073)*	-0.204 (0.063)**
=1 if either country is an island				-0.078 (0.087)	0.039 (0.086)
=1 if pair shares at least one sea or ocean				-0.159 (0.137)	-0.063 (0.070)
Average elevation between countries				-0.028 (0.223)	-0.049 (0.141)
Freight rate (surface transport)				16.532 (14.387)	-4.004 (5.938)
=1 if either country is landlocked				0.074 (0.178)	[a]
Constant	0.378 (0.099)**	0.382 (0.084)**	0.413 (0.113)**	-2.079 (2.293)	1.130 (0.947)
# of observations	325	325	325	325	171
# of countries	26	26	26	26	19
Standardized beta (%)	31.69%	39.80%	41.08%	34.61%	59.28%
R-Squared	0.10	0.16	0.17	0.21	0.39

Two-way clustered standard errors in parentheses; * significant at 10%; ** significant at 5%

[a]: dropped due to singularity

Table 13: Controlling for cultural distance in the Europe dataset (two-way clustered standard errors)
Dependent variable: absolute value of log income differences, 1995

	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
	Baseline	**Linguistic distance**	**Religious distance**	**Both measures**	**% cognate, plurality**	**% cognate, plurality**
Genetic distance, relative to the English	**41.691** **(18.875)****	**42.766** **(19.223)****	**41.424** **(18.963)****	**42.485** **(19.310)****	**44.252** **(20.209)****	**44.096** **(20.288)****
Linguistic distance, plurality, relative to English		-0.221 (0.114)*		-0.224 (0.109)**		
Religious distance, plurality, relative to Lutherans			0.096 (0.164)	0.107 (0.171)		
1-% cognate, plurality, relative to English						0.045 (0.186)
Constant	-1.537 (2.025)	-1.466 (1.975)	-1.519 (2.027)	-1.445 (1.973)	-1.339 (2.161)	-1.314 (2.231)
# of observations	300	300	300	300	276	276
# of countries	25	25	25	25	24	24
Standardized beta	37.55%	38.51%	37.31%	38.26%	39.52%	39.38%
R-Squared	0.21	0.22	0.21	0.22	0.25	0.25

Two-way clustered standard errors in parentheses; * significant at 10%; ** significant at 5%
All columns include the following controls (estimates not reported): absolute difference in latitudes, absolute difference in longitudes, geodesic distance, dummy for contiguity, dummy=1 if either country is landlocked, dummy=1 if pair shares at least one sea or ocean, average elevation between countries, freight rate (surface transport).
Compared to Table 12, in columns (1)-(4) Iceland is dropped due to missing data on linguistic and religious distance from Fearon. In columns (5) and (6) Hungary and Finland are dropped because their languages are not Indo-European, and thus not part of the lexicostatistical dataset.

Figure 1 - Population Tree

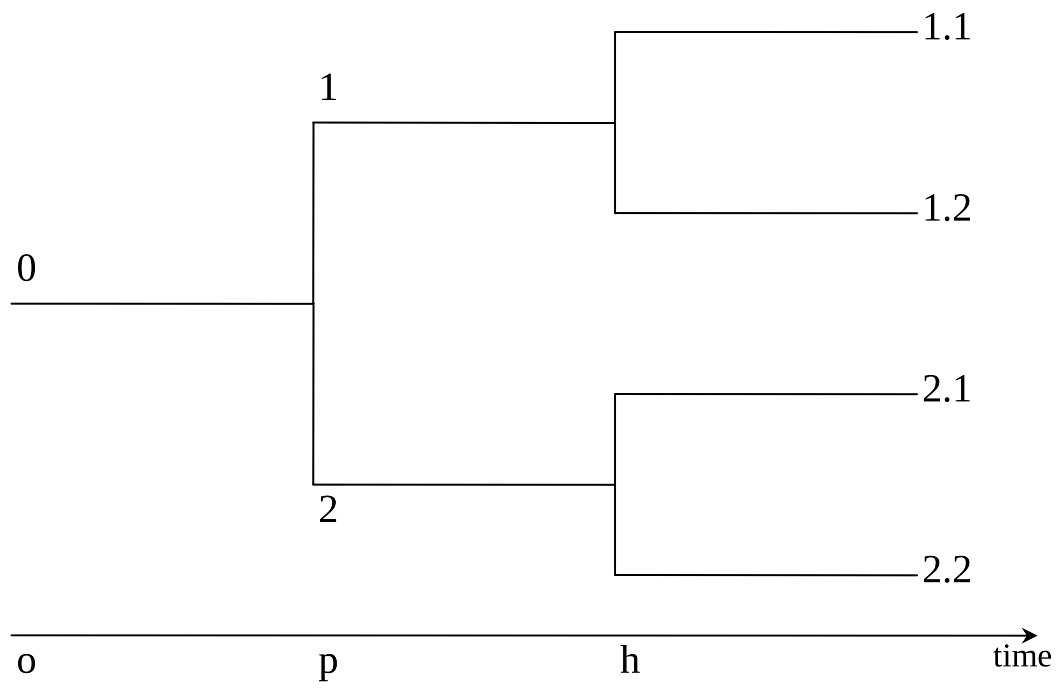


Figure 2 - Genetic distance among 42 populations.
Source: Cavalli-Sforza et al., 1994.

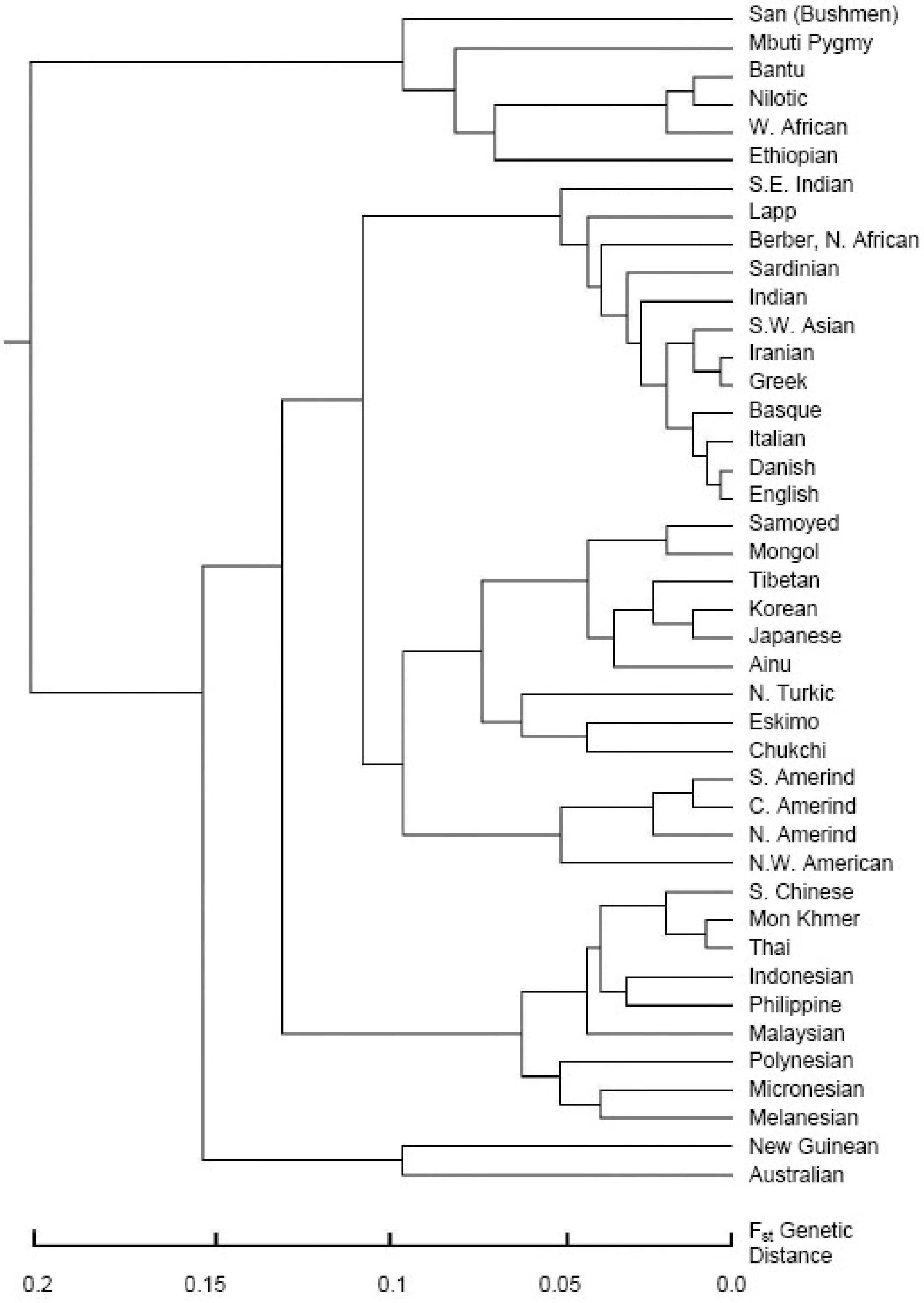


Figure 3 – Log Income in 1995 and Genetic Distance to the USA

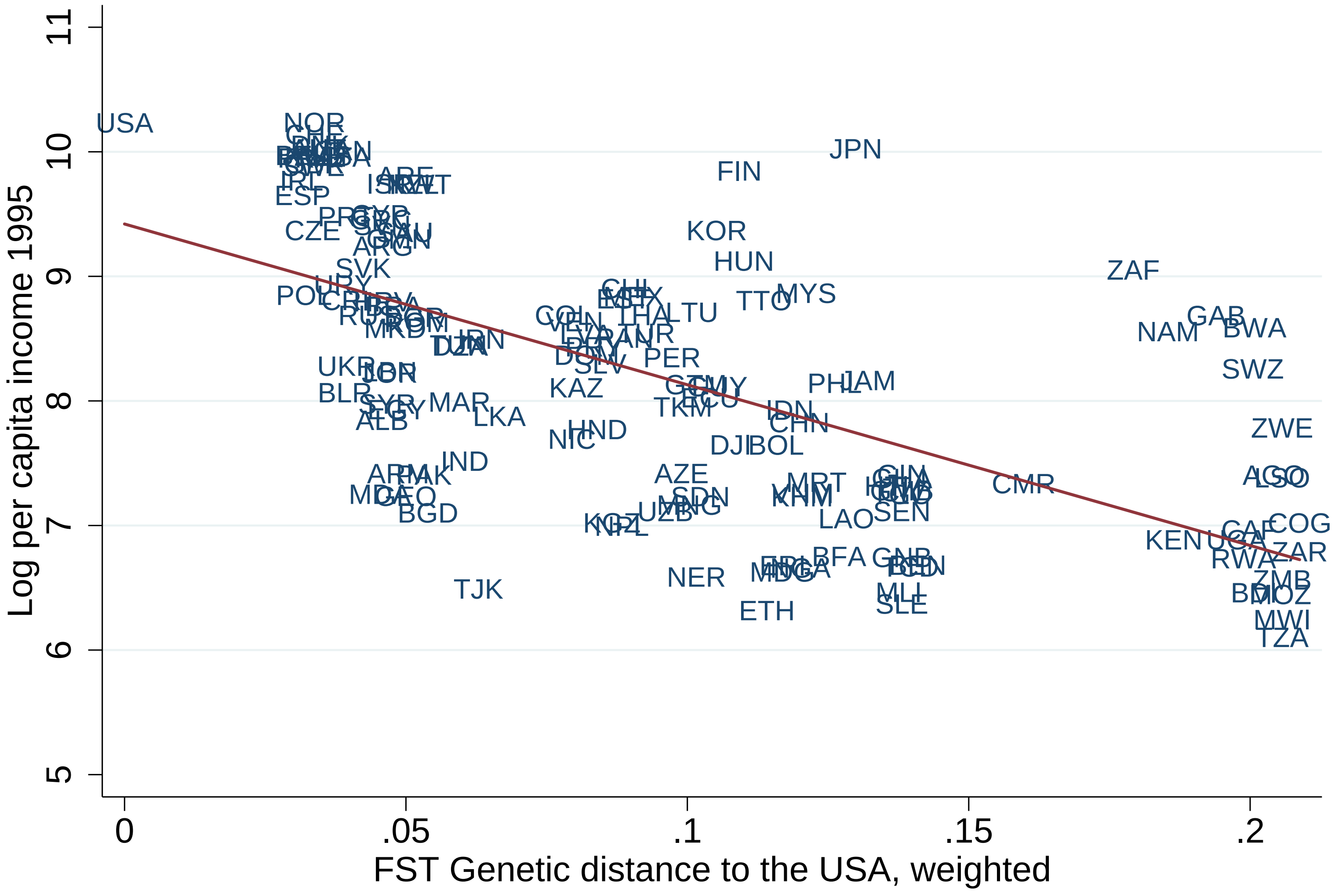
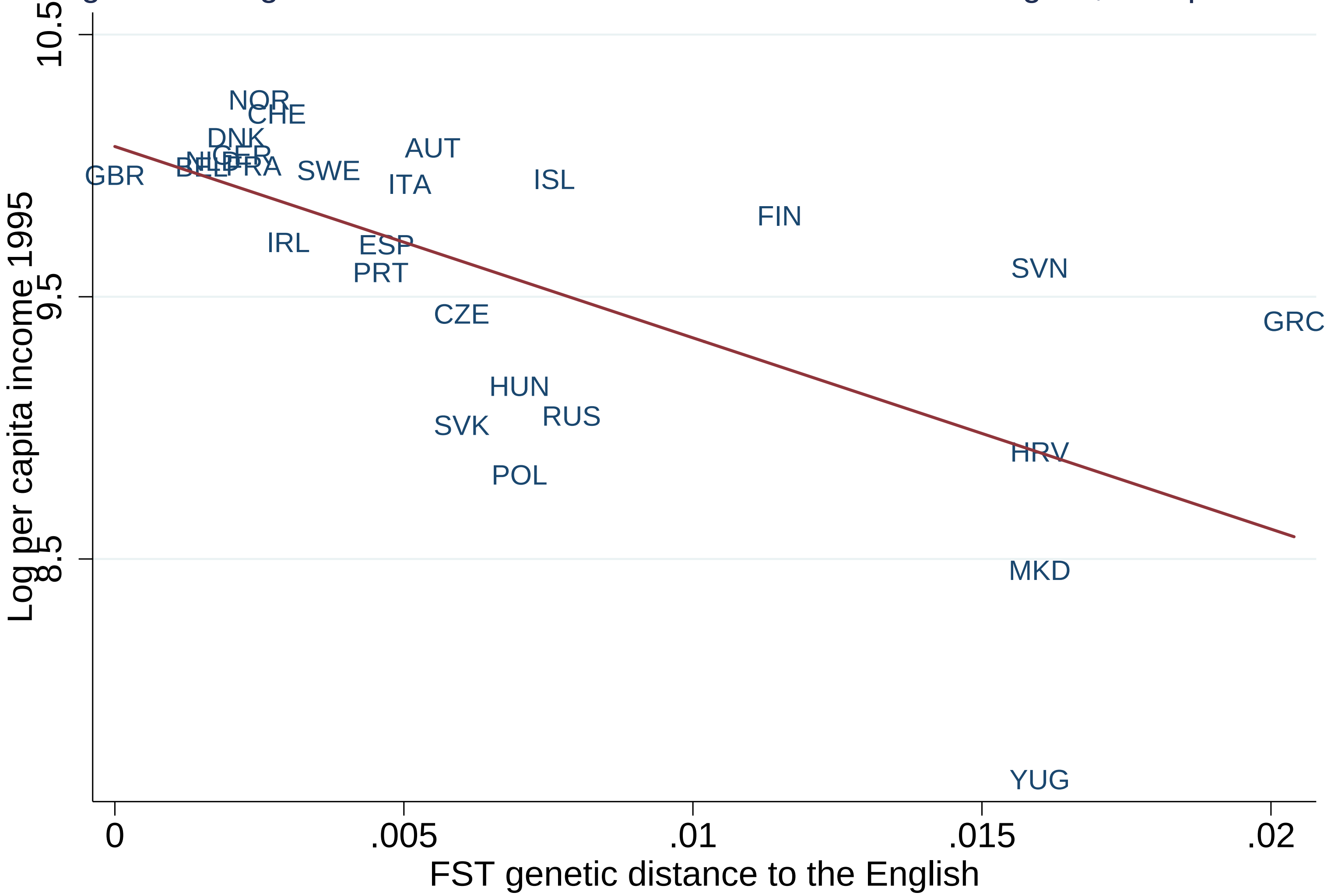


Figure 4 − Log Income in 1995 and Genetic Distance to the English, Europe Dataset


Log per capita income 1995
FST genetic distance to the English
NOR
CHE
DNK
GER
NLD
BEL
FRA
GBR
SWE
AUT
ITA
ISL
FIN
IRL
ESP
PRT
SVN
CZE
GRC
HUN
RUS
SVK
HRV
POL
MKD
YUG