Genetic Distance and International Migrant Selection*

Tim Krieger† Laura Renner‡ Jens Ruhose§

This version: July 2015

Abstract

This paper looks at the effect of the relatedness of two countries, measured by their genetic distance, on educational migrant selection. We exploit bilateral country-level education-specific migration stocks from 85 sending countries to the main 15 destination countries in 2000 and show that country pairs with higher genetic distances exhibit more selected migrant stocks compared to country pairs with lower genetic distances on average. The effect is driven by country pairs with genetic distances above the median genetic distance, suggesting that genetic distance must be sufficiently large to constitute a barrier to migration for low-skilled migrants. Results are robust to the inclusion of sending and destination country fixed effects, bilateral control variables, and an instrumental variables approach that exploits exogenous variation in genetic distances in 1500.

JEL-Code: F22, J61, Z1
Keywords: Genetic Distance, International Migration, Selection, Culture

*We are grateful to David Dorn, Benjamin Elsner, Oliver Falck, Gabriel Felbermayr, Paola Giuliano, Volker Grossmann, Gordon Hanson, Ingrid Kubin, Volker Nitsch, Jens Suedekum, and seminar and conference participants at the University of Freiburg, EGIT (Munich), ESPE (Braga), EEA (Toulouse), Verein für Socialpolitik (Hamburg), BevÖkA (Nuremberg) and EALE (Ljubljana) for their most helpful comments and discussions. We also thank Ingo Isphording and Sebastian Otten for sharing their language distance data with us. An earlier version of the paper circulated under the title "Culture, Selection, and International Migration".

†Department of Economics, University of Freiburg, Wilhelmstr. 1b, 79085 Freiburg i. Br., Germany; E-Mail: tim.krieger@vwl.uni-freiburg.de; Phone: +49 761 203 67651; and CESifo, Munich, Germany

‡Department of Economics, University of Freiburg, Wilhelmstr. 1b, 79085 Freiburg i. Br., Germany; E-mail: laura.renner@vwl.uni-freiburg.de; Phone: +49 761 203 67652

§Ifo Institute - Leibniz Institute for Economic Research at the University of Munich, Poschingerstr. 5, 81679 Munich, Germany; E-mail: ruhose@ifo.de; Phone: +49 89 9224 1388; and IZA, Bonn, Germany

1 Introduction

In 2011, about 5.38 million people migrated to OECD countries. This number has increased by 40 percent since 2000.[1] Importantly, international migration to these countries is dominated by individuals with higher skill levels as they are more likely to migrate than those with lower skills (Grogger and Hanson, 2011). Studying and understanding the determinants of international migrant selection is important because high-skilled migrants are essential for the economic development in countries that rely on innovation-driven economic growth (Nelson and Phelps, 1966; Coe and Helpman, 1995; Chambers et al., 1998).

Since Borjas (1987), a large literature has evolved that tries to explain migrant selection by differences in the returns to skills.[2] However, the most recent OECD International Migration Outlook (OECD, 2014) reports that only one third of all migrants can be seen as labor migrants that can be expected to migrate for purely economic reasons. Most other migration comes through, e.g., family reasons, humanitarian reasons, and by accompanying families of workers. Thus, it is likely that other factors than differences in earnings opportunities shape the selection of international migration as well. In the present paper, we argue that the genetic distance between two countries may serve well as a measure that is able to predict the migration behavior of a broader population group.

In recent papers, Spolaore and Wacziarg (2009, 2015) argue that innovation spreads less easily between societies that are genetically distant, as these societies find it more difficult to learn from each other. Human genetic distance is here seen as a *summary measure of very long-term divergence in intergenerationally transmitted traits across populations* (Spolaore and Wacziarg, 2009, p. 471). The closer societies are in terms of these traits, the easier they can interact, thereby facilitating the diffusion of knowledge and innovation across population boundaries.

We argue that the same kind of distance, or closeness, between populations ought to affect international migration as well. Assuming potential migrants to search for an optimal destination, expected migration costs rise at the individual level if the destination-country population is perceived as very different from the mates at home. Due to fewer advantages to cope with these differences the low-skilled ought to be less willing to move abroad than the high-skilled. The latter may, for instance, have advantages in information gathering and processing, providing them with a larger set of possible destinations than low-skilled individuals have. Hence, we should observe that the migration stock of country pairs with a high genetic distance is more positively selected than between country pairs with lower genetic distances.

[1] See OECD International Migration Database.

[2] Recently, see for example: Abramitzky (2009); Belot and Hatton (2012); Chiquiar and Hanson (2005); Fernández-Huertas Moraga (2011); Grogger and Hanson (2011); Stolz and Baten (2012); Kaestner and Malamud (2014); Gould and Moav (2014); Parey et al. (2015).

In contrast to the education-specific migration cost argument above, it is also possible that people migrate to other countries because they want to live in a different cultural environment, i.e., because of their pronounced intercultural interest or love of adventure (Krieger and Lange, 2010). This would mean that a higher genetic distance acts like a benefit in the migration decision. Because it is uncertain whether high-skilled individuals have a higher or lower propensity for 'lifestyle migration' (Benson and O'Reilly, 2009a,b) than low-skilled individuals, this sort of migration makes a clear prediction about selection difficult. That is, the overall effect of genetic distance on the selection of international migration is ambiguous, even though we would predict a priori that the migration cost mechanism is stronger than the lifestyle migration channel.

We show that the relatedness of countries, measured by their genetic distance, can explain international migrant selection. By looking at education-specific bilateral migrant stocks for the 15 main destination countries and 85 source countries (Docquier et al., 2007), we find evidence that, on average, migration is more skilled between country pairs that have a higher genetic distance than between countries with a lower genetic distance. The average effect, however, conceals important non-linearities. Sample splits and non-linear models show that the average effect is driven by country pairs with genetic distances above the median genetic distance. For country pairs below the median genetic distance, we do not observe that migration flows are selected. The findings suggest that genetic distance can be interpreted as education-specific migration costs at sufficiently high levels of genetic distance. However, at lower levels, genetic distances do not show up as substantial migration costs for neither of the two skill groups. The observed effects are robust to the inclusion of several control variables and to an instrumental variables approach, which uses exogenous variation in genetic distance in 1500 to correct for endogeneity bias that comes through past migration waves.

Why may genetic distance affect international migration patterns (including migrant selection)? *Dual inheritance theory* in social anthropology (Boyd and Richerson, 1985; Henrich and McElreath, 2003) argues that genes and culture develop together as time progresses. While genes are inherited, culture is learned and imitated from, for example, parents and teachers. Similar to the definition of genetic distance by Spolaore and Wacziarg (2009) above, Guiso et al. (2006, p. 23) define culture as *those customary beliefs and values that ethnic, religious, and social groups transmit fairly unchanged from generation to generation.* Hence, both genes and culture have in common that they are transmitted from generation to generation and change only very slowly.

A recent strand in the literature on migration shows that cultural traits affect migration flows, for instance, the size of these flows (Belot and Ederveen, 2011; Mayda, 2009; Falck et al., 2012, 2015). Furthermore, at least in case of inner-German migration, high-skilled individuals are more likely to cross cultural borders (Bauernschuster et al.,

2014).[3] The close relationship between culture and genes according to dual inheritance theory therefore suggests that genetic distance may serve as an appropriate measure of perceived differences between countries and may be responsible for migrant selection. In fact, genetic distance ought to be preferred over cultural distances due to the lack of consensus how to measure culture or cultural differences. Guiso et al. (2006) themselves rely on ethnic and religious differences, but Falck et al. (2012) use instead linguistic differences and Mayda (2009) a common language or a colonial history.

The problems of defining and measuring culture may be avoided securely by using data on genetic distance. Using respective data from Spolaore and Wacziarg (2009), we show that indeed genetic distance affects significantly migrant selection. Interestingly, there remains in our study an independent significant effect of genetic distance on migrant selection even after controlling for a number of variables typically used to measure cultural differences (e.g., linguistic distance, common language, religion, colonial history). Since genetic distance remains a significant predictor of migrant selection throughout, we argue that genetic distance is a proxy for normally unobserved cultural traits, habits, and norms that affect migration decisions.

The remainder of the paper is organized as follows. Section 2 introduces genetic distance and selection measures and describes the data. Section 3 provides the econometric setup and explains the identification strategy. In section 4, we provide the results of our analysis. Section 5 concludes.

2 Genetic Distance and Selection of International Migration: Concepts and Data

2.1 Genetic Distance

How Is Genetic Distance Measured?

In this paper, we use the genetic distance data from Spolaore and Wacziarg (2009), who, in turn, refer to the seminal work by Cavalli-Sforza et al. (1994). Cavalli-Sforza et al. (1994) assemble a matrix of bilateral genetic distances between populations on which they base their analysis of the timing of the emergence of the different populations across

[3] In recent years, the concept of culture has attracted the attention of many researchers in explaining economic outcomes (Ottaviano and Peri, 2005; Guiso et al., 2006; Tabellini, 2010; Ashraf and Galor, 2013; Burchardi and Hassan, 2013; Spolaore and Wacziarg, 2013). Cultural traits are especially successful in explaining the size and the direction of economic exchange, such as income differences between countries (Spolaore and Wacziarg, 2009), migration flows (Falck et al., 2012; Belot and Ederveen, 2011; Dahl and Sorenson, 2010; Mayda, 2009), the diffusion of technology (Comin et al., 2012; Spolaore and Wacziarg, 2012), trade patterns (Guiso et al., 2009; Felbermayr and Toubal, 2010), or investment behavior (Guiso et al., 2009). In a recent contribution, Spring and Grossmann (2015) show that bilateral trust might not predict economic exchange as well as Guiso et al. (2009) suggest. They use somatic distance as an instrument for trust. Thus, by using genetic distance we avoid the critique of Spring and Grossmann (2015) and capture, beside trust, also broader aspects of cultural differences between countries.

the world. Thus, intuitively, their measure is proportional to the time span since two populations have separated. This time span is what we want to exploit in this paper.[4]

The basis for the F_{ST} genetic distance, that we use in this paper, is the difference in the frequencies of alleles across populations. Alleles are different forms or variants of genes. While a gene determines a certain trait, e.g. the blood group, the allele, specifies which blood group an individual has (Cavalli-Sforza, 2001). The geneticists use data on 120 alleles of 42 world populations and calculate the frequencies for these alleles in all 42 populations. Specifically, the F_{ST} genetic distance between two populations is calculated for all genes available and then the distance values are averaged with the mean gene frequency.[5] If alleles are identically distributed across two populations, the F_{ST} genetic distance is zero. Consequently, this means that the populations have developed together or at least that they mix very frequently.

After calculating a matrix of genetic distances between population pairs, the next step is to connect the genetic distance to the timing of separation of the populations. By using the genetic distance, one can estimate the time that is elapsed–like a *molecular clock*–since the last common ancestor (Cavalli-Sforza et al., 1994). To apply this method, we have to assume that the evolution of genes is random, that is, differentiation of gene frequencies is by random mutation only (random drift). Geneticists take care of this assumption by looking only at genes that are considered as neutral and not at those that are best adapted in order to survive (survival of the fittest).

For the purpose of cross-country analysis, the matrix on genetic distances between populations by Cavalli-Sforza et al. (1994) needs to be assigned to countries within today's boundaries. Spolaore and Wacziarg (2009) provide a matched F_{ST} genetic distance that we also use in this paper, in which populations have weights according to the share of the respective populations in a country.[6]

Table 1, Panel A, shows summary statistics for the genetic distance data. One standard deviation in genetic distance is represented by 572 points, the mean is 716. Based on the genetic distance between the USA and Germany (352), one standard deviation indicates a shift to the genetic distance between the USA and Mexico (904), the USA and Thailand (920), or the USA and Turkey (927).[7] For the regression analysis, we divided genetic distance by its standard deviation, such that we can interpret the results for an increase of one standard deviation in genetic distance.

[4]Spolaore and Wacziarg (2009, p. 481) also argue that the time span since two populations shared a common ancestor stores information about the relatedness of populations.

[5]There are various ways to compute genetic distance measures. Cavalli-Sforza et al. (1994, p. 29) argue that the F_{ST} genetic distance has the most convenient properties and that the correlation between F_{ST} genetic distance and alternative measures, such as the *Nei modified genetic distance*, is high.

[6]Weights are calculated by Spolaore and Wacziarg (2009) based on ethnic composition data of countries by Alesina et al. (2003). We do not have information on genetic distance for the Czech Republic and therefore drop this country as a source country from the analysis.

[7]The F_{ST} genetic distance can take values between 0 and 1 in the data matrix provided by Cavalli-Sforza et al. (1994), which is multiplied by 10,000.

[Table 1 here]

However, some limitations need to be addressed: First, the matching from populations to countries might introduce some measurement error. This could be because population groups are hard to identify or that a higher within-country genetic diversity makes it harder to aggregate genetic diversity to the country level. Ashraf and Galor (2013) focus on the question whether such within-country genetic diversity has effects on the economic development of countries. However, the geneticists argue that within-country variation of genetic diversity is small compared to the variation between world populations (Cavalli-Sforza et al., 1994). Second, there might still be doubt that only random drifts affect genes. Geneticists argue that they use so many genes in the calculation of genetic distances that even if migration or natural selection has an impact on the flow of genes, it should not bias genetic distance measures (Cavalli-Sforza et al., 1994).

Spolaore and Wacziarg (2009) also provide a F_{ST} genetic distance based on populations in 1500. Since populations in 1500 are close to the world populations used by Cavalli-Sforza et al. (1994), this limits measurement error in the assignment of genetic distances to populations because populations at that time are largely unaffected by later mass migration flows. In their analysis, Spolaore and Wacziarg (2009) propose the genetic distance based on populations in 1500 as an instrument for genetic distance in the 1990s. We follow this proposition and use this instrument in our analysis too. As we show later, the genetic distance in 1500 is a good predictor for the distance in 1990. Notable exceptions are the United States and Australia, where native populations in 1500 are not at all influenced by later colonizations.

What Does Genetic Distance Measure?

What do we measure with genetic distances between countries? We follow the interpretation of Spolaore and Wacziarg (2015) who argue that genetic distance represents *a summary statistic for a wide array of cultural traits transmitted intergenerationally.* In other papers, Spolaore and Wacziarg (2009, 2013) use the same genetic distance as we use as a measure for the relatedness of two countries.

An important theoretical basis for using genetic distance as a proxy variable for differences in cultural traits comes from the *dual inheritance theory* in social anthropology. This theory points specifically to the parallels between genes and culture. Boyd and Richerson (1985) and Henrich and McElreath (2003) argue that culture is a system of inheritance, following evolutionary developments as genes do. In addition, geological and ecological barriers strengthen the differentiation between groups and therefore can affect genes and culture in the same way. Finally, cultural differences and genetic differences enforce each other. One example for this is that marriage appears mostly within the same ethnic or religious group (Falck et al., 2012).

Genetic differences and cultural differences are similar in the sense that they are both

transmitted from generation to generation and are both changing rather slowly. The longer two populations develop separately, the more time for development in different directions and the greater the distance in genes and culture (Cavalli-Sforza et al., 1994). This does not assume that genes determine culture or that culture determines genes, but it indicates important parallels in the development of genes and culture. More specifically, the main idea, given by Cavalli-Sforza et al. (1994, p. 23 and pp. 380-382), is that both genome and culture follow the same history of fissions, that is, split-ups of populations. Most importantly, genome and culture develop over similar channels: Both consist of information which is accumulated and given on from generation to generation. While genes are inherited, culture is learned and imitated from, for example, parents and teachers. A longer time span since the last fission implies more time for the accumulation of differentiated information. Like genes, deeply rooted beliefs and behaviors (e.g. family structures), which are already imitated and learned from early ages on, are probably also changing very slowly.[8]

2.2 Selection of International Migrants

To investigate the relationship between selection of migrants and genetic differences, we need bilateral migration data by skill level between countries. In this paper, we use the 2000 cross-sectional bilateral dataset from Docquier et al. (2007). Their data provides information on emigrant stocks and residents by source and destination countries, including education level (primary, secondary, and tertiary). As in Grogger and Hanson (2011), we restrict our analysis to the 15 main immigrant destination countries: Australia, Austria, Canada, Denmark, Finland, France, Germany, Ireland, the Netherlands, New Zealand, Norway, Spain, Sweden, the UK and the US. Due to data availability, the sample of source countries is restricted to 85.[9]

Departing from utility maximization and assuming that the error structure follows an i.i.d. extreme value distribution, it can be shown that the log odds of migrating to destination country d versus staying in source country s is equal to the log of the share of the population of skill level $j \in \{H(igh), L(ow)\}$ from s that has migrated to d, that is E^j_{sd}, over the population with skill level j in s that remains in s, that is E^j_s (McFadden, 1974). Hence, $\ln \frac{E^j_{sd}}{E^j_s}$ gives the log of the share of the migrants in d of skill group j from country s. A larger fraction signals a larger scale of migrants from country s residing in country d (by skill level).

[8]Several studies examine the persistence of culture and their results point to the existence of deeply-rooted beliefs, which are changing only very slowly. Alesina et al. (2013), for example, show that the use of the plough in pre-industrial times leads to stricter gender roles regarding work behavior of women today. Voigtländer and Voth (2012), as another example, shows that regions within Germany that had pogroms in the 14th century against Jewish people, who were blamed for the black death, voted more for the Nazi party in 1928.

[9]See Appendix Table A-2 for the list of source countries.

Figure 1 plots the log odds of emigration for tertiary educated versus the log odds of emigration for primary educated for each source country in our sample. All log odds for the primary educated migrants are below zero which indicates that the low-skilled migrant population is always smaller than the low-skilled population left behind. Indicated by positive log odds, the figure reveals that for countries such as Trinidad and Tobago (TTO), Jamaica (JAM), and Guyana (GUY), the tertiary-educated population living abroad is larger than the tertiary-educated population left behind. The 45°-line in Figure 1 describes equal log odds of migration between the two skill groups. Almost all countries show a higher propensity of tertiary-educated than primary-educated migration. The USA is a notable exception.

[Figure 1 here]

The question of this paper is how genetic distance (as a possible approximation of cultural distance) influences migrant selection. Our hypothesis is that a higher genetic distance between source country s and destination country d leads, relatively, to a larger high-skilled (tertiary-educated) migrant population than to a low-skilled (primary-educated) migrant population from s in d.

Figure 2 gives a first impression of the scale of high- and low-skilled migration between pairs of countries, depending on the extent of the genetic distance between the two. To ease interpretation, we use non-parametric binned scatter plots. Instead of showing all country pairs, we bin genetic distance into 20 equal sized bins and then plot the means of the log odds of emigration within each bin for each skill level. The relationship between genetic distance and the log odds of primary-educated emigration is negative; indicating that a higher genetic distance is associated with lower migration of low-skilled individuals. The relationship between genetic distance and the log odds of tertiary-educated migrants is not that clear cut. However, if at all, the relationship is slightly positive but not statistically significant.

[Figure 2 here]

We can combine the measures for the scale of emigration by skill level and use the ratio of the two as a measure of the migrant skill mix that destination country d receives from source country s, that is $\left(\ln \frac{E^H_{sd}}{E^L_{sd}} - \ln \frac{E^H_s}{E^L_s}\right)$. Figure 3 shows the relationship between the skill mix of migrants and genetic distance. Again, the figure is a non-parametric binned scatter plot as described above. As expected from Figure 2, the relationship is strongly positive. A higher genetic distance is again associated with more high-skilled migrants than low-skilled migrants. The figure also reveals that at very low genetic distances, we can expect that the migrant skill mix is close to 1 or even below 1. Thus, for country pairs that are genetically similar, we predict a balanced inflow of tertiary-educated versus primary-educated migrants.

[Figure 3 here]

Table 1 provides in Panel B summary statistics for emigration shares by skill level and the migrant skill mix. The emigration share of the primary-educated population has a mean of 0.003. That means that, on average, 0.3 percent of the source country low-skilled population lives abroad. That share is equal to 2.9 percent for the high-skilled. Again, this reveals the positive selection in international migration (cf. Figure 1).

The migrant skill mix, $\left(\ln \frac{E^H_{sd}}{E^L_{sd}} - \ln \frac{E^H_s}{E^L_s}\right)$, is the outcome of interest in this paper. Migrants are positively selected when the share of migrants from country s is disproportionally high-skilled, that is, when the scale of high-skilled migrants is larger then the scale of low-skilled migrants, $\ln \frac{E^H_{sd}}{E^L_{sd}} > \ln \frac{E^H_s}{E^L_s}$. Table 1 shows the sample mean of the migrant skill mix. The log odds interpretation indicates that it is, on average, 81 percent more likely to see high-skilled emigration versus low-skilled emigration.

An empirical problem is that we do not observe the migrant stock for 158 out of potentially 1,260 country pairs (about 13 percent of the sample).[10] The destination countries with the largest missing information is Ireland (40 source countries), followed by Austria (35), Sweden (25), and Spain (19). Altogether, these four destinations make up 75 percent of all missing migrant stock observations. In a robustness check, we exclude all four destination countries from the sample and do not observe that the results are different from the baseline result using all destination countries.

2.3 Other Variables influencing Migrant Selection

To control for important confounding factors, we consider several variables which could drive migrant selection and might be correlated with genetic distance. Summary statistics for all variables are documented in Panel C of Table 1.

First of all, geographic barriers between two countries influence the flow of migrants as they increase transportation and adaptation costs. Geographical barriers could also be a reason for the observed genetic distance as populations developed along those barriers.[11] For example, Giuliano et al. (2014) show the importance of geographical barriers in the relationship between genetic distance and trade. Therefore, our regressions take care of the (log) geographic distance (in km) and whether the destination and the source country share a common border (contiguity). The data comes from Head et al. (2010). To capture non-linearities in geographic distance, we include the difference in absolute longitude, the difference in absolute latitude of the two countries and the differences in

[10] In fact, it is not clear whether the migrant stocks are zero or are missing because bilateral migration propensities are so small that the countries' survey does not report migrants from particular countries. We cannot simply impute zero values because we use logged migrant stocks as our dependent variable later on.

[11] Appendix Table A-3 shows that the correlation between genetic distance and log geographic distance is 0.43.

average temperature and average precipitation (Ashraf and Galor, 2013).[12]

The next set of variables is concerned with language barriers. Adsera and Pytlikova (2014) show that the language distance between the source and the destination country is a major obstacle for migration flows. Learning a new language or being proficient in a foreign language might be easier for high-skilled people; thereby affecting migrant selectivity. However, language differences are also a part of cultural differences. Thus, we want to test whether the effect of genetic distance comes purely through language differences.

To capture language differences sufficiently well, we use several indicators. Isphording and Otten (2013) construct a language distance indicator that is conceptually closely related to genetic distance. The language distance statistic relies on the Levensthein distance and compares the pronunciation of a set of words with the same meaning across languages and can be understood as the number of cognates, that is common ancestries, between two languages. The final Levensthein distance is achieved by averaging over the set of words and gives a percentage measure of dissimilarity.[13] The closer the languages of source and destination country are, the smaller the Levensthein distance. The smallest language distance in our sample is between Finland and Estonia, while Denmark and Jordan have the maximum value. Furthermore, since English is widely taught in schools, we introduce a dummy for anglophone destination countries. Finally, we control for a shared official language, which is the case when at least 9 percent of the population speak the same language (Head et al., 2010).[14]

Another important factor for migrant selection is the presence of a diaspora or migrant network in the destination country. Existing networks (Diasporas) increase information access and offer a surrounding close to that in the home country. This reduction in migration costs result in increased migration flows with relatively low average education levels from sending regions with larger migrant networks compared to sending regions with smaller networks (Beine et al., 2011). The calculation of migrant networks follows the procedure of Belot and Hatton (2012) who calculate the share of migrants (of all education levels) from a source country in the destination country relative to all residents in the source country. Arguably, like differences in the language, migrant networks are a product of genetic distances. Thus, introducing migrant networks potentially explains some of the effects of genetic distance.

We use wage data from Grogger and Hanson (2011) to capture the influence of skill premia, which are key factors in the selection of migrants (Borjas, 1987). Grogger and

[12]For robustness, we also run regressions by including geographic distance linearly and introducing geographic distance, squared and cubic.

[13]When languages do not even have random similarities the value can be above 100 percent, e.g. Vietnamese to English (104,06).

[14]Appendix Table A-3 shows that the language distance and sharing a common language is positively correlated with the genetic distance.

Hanson (2011) provide comparable wage measures for the 80th and 20th income percentile for each source and destination country in our sample. We use the difference between the destination and the source country in the 80th/20th wage ratio, to proxy for monetary incentives of selective migration. The underlying data is compiled by using wage data from the World Development Indicators and from the WIDER World Income Inequality Database. However, differences in 80th/20th income ratio might also be a function of genetic distance. Spolaore and Wacziarg (2009) indeed show that income differences across countries are converging in genetic proximity.

Arguably, political and legal barriers as well as the general openness of a destination country also contribute to migration costs, which might be more easily carried by high-skilled migrants. Visa restrictions are a major source for countries such as the United States, Canada, or Australia to control for the quality of migrants. Therefore, we control for visa restrictions by using a dummy which is 1 if the destination country has imposed a visa restriction on the source country (Neumayer, 2006). We also use dummies for country pairs that are signatories of the Schengen agreement and for country pairs that were in a colonial relationship (Head et al., 2010). To measure the general openness of a country toward immigration, we include the log of the aggregate inflow of foreigners and the log number of asylum-seekers into the country, both retrieved from the International Migration Dataset of the OECD.

Our baseline model is completed by measures of the general country skill level because countries with a more similar skill mix are more likely to interact. Thus, we use the difference between the destination country and the source country in years of schooling and in the share of people with completed tertiary schooling, both taken from Barro and Lee (2013).[15]

3 Econometric Setup

3.1 Estimation

As mentioned above, the aim of this study is to explain the migrant skill mix in destination country d from source country s, that is, $\ln \frac{E^H_{sd}}{E^L_{sd}} - \ln \frac{E^H_s}{E^L_s}$, through variations in the genetic distance. Whenever $\ln \frac{E^H_{sd}}{E^L_{sd}} > \ln \frac{E^H_s}{E^L_s}$, then migrants are positively selected from the source country population. Equation (1) sets out the regression model that we are using for estimating the correlation of genetic distance on the skill mix of migrants. Grogger and Hanson (2011) derive this equation formally based on individual utility maximization.

$$\ln \frac{E^H_{sd}}{E^L_{sd}} - \ln \frac{E^H_s}{E^L_s} = \beta_0 + \beta_1 \, Genetic\ distance_{sd} + \mathbf{X'_{sd}}\phi + \mu_{sd} \tag{1}$$

[15]Including these measures mean that we have do drop 13 source countries from the analysis. However, because the omitted countries are not important source countries, the results are unchanged when including them in models without the education variables.

In our baseline specifications, we stepwise include the set of control variables explained above. These variables contain the log geographic distance and other geographic controls, language distance and variables capturing the difficulties to communicate, the difference in the 80/20 wage ratio, migrant networks, visa restrictions, the inflow of foreigners and asylum seekers as well as the difference in the population skill mix. The error term ϵ_{sd} of Equation (1) is clustered at the destination country level to allow for arbitrary correlation within destination countries.[16]

The coefficient of interest in Equation (1) is the coefficient on genetic distance, β_1. The coefficient would reveal a causal effect of genetic distance on the migrant skill mix if and only if genetic distance is not correlated with the error term. This identifying assumption is unlikely to hold. Subsection 3.2 discusses why this might be the case and explains our identification strategy.

3.2 Identification

The main concern in the current cross-sectional framework is that persistent, unobserved factors, which have shaped the genetic distance in 1990, also cause migrants to select into different destination countries. This can be migrant networks that go beyond the simple measure that we are using in the analysis. It could also be that exactly the (unobserved) cultural traits and habits, that we are trying to identify, have driven genetic distances in the past and are also causing migrant selection today. More complex migrant networks and persistent cultural traits and habits cause an upward bias in the OLS regression; meaning that the true effect of genetic distance on the migrant skill mix is lower than β_1 from Equation (1). Furthermore, genetic distance is measured with more or less precision for different countries. For example, genetic distance can be expected to be measured more accurately, when the genetic variation within both countries is lower. The measurement error that is introduced through the imprecise measurement of genetic distance causes a bias in β_1 toward zero. Thus, the true effect in this case should be higher. Therefore, the bias in β_1 from Equation (1) could go either way.

To break the omitted variable problem and mitigate the measurement error issue, we use exogenous variation in genetic distance that is reported before major migration waves have happened. As proposed by Spolaore and Wacziarg (2009), we use the genetic distance in 1500 as an instrument for the genetic distance in 1990. The identifying assumption is that the genetic distance in 1500 has an effect on the migrant skill mix in 2000 only through the genetic distance in 1990 (see Spolaore and Wacziarg (2009) for a detailed discussion of the validity of the instrumental variables approach).

Empirically, we estimate the model in two steps. In the first step, we predict the genetic distance in 1990 by using the variation in genetic distance from 1500; controlling

[16] Clustering at the destination × source country level or using two-way clustering (Cameron et al., 2011) at the destination and the source country level do not affect the results.

for the full set of control variables. Equation (2) gives the first stage regression of the two stage-least-squares procedure.

$$Genetic\ distance_{sd} = \lambda_0 + \lambda_1\ Genetic\ distance_{sd}^{1500} + \mathbf{X'_{sd}}\omega + \nu_{sd} \tag{2}$$

Once we have predicted the genetic distance from the first stage, we can include the fitted values into the second stage of the second stage regression (Equation (3)). In this step, we use only the variation in genetic distance that is triggered by the variation in 1500.

$$\ln \frac{E_{sd}^H}{E_{sd}^L} - \ln \frac{E_s^H}{E_s^L} = \beta_0 + \beta_1\ \widehat{Genetic\ distance}_{sd} + \mathbf{X'_{sd}}\phi + \mu_{sd} \tag{3}$$

Note that we still cluster the standard errors at the destination country level and that we estimate the first and the second stage within the same routine to account for the predicted values in the second stage, which is important to receive correct standard errors.

4 Results

4.1 Explaining Migrant Selection

Figure 3 provides a graphical illustration of the relationship between genetic distance and the migrant skill mix. Table 2 shows the results of the OLS regressions. This exercise should give a first impression on which variables are important for explaining migrant selection. We deal with causality more seriously in the next subsection. There, we also discuss the use of destination and source country fixed effects.

Column (1) of Table 2 shows the unconditional correlation between genetic distance and migrant selection. We observe that the coefficient on genetic distance is positive and highly significant. This indicates that a higher genetic distance of a country pair is associated with a higher migrant skill mix. We discuss effect sizes later, but note that we have standardized genetic distance by dividing the variable through the own standard deviation. We do the same with log geographic distance and language distance. This has the advantage that the coefficient between these important variables are directly comparable. In addition, the interpretation of the effect sizes are now in terms of standard deviations.

[Table 2 here]

In Columns (2) and (3), we add geographic variables to the model. Column (2) reveals that country pairs that are geographically farther away exhibit a more selective migration. At the same time, the coefficient on genetic distance drops substantially from 0.808 to 0.660. As expected, genetic distance is to some degree determined by geographic distance because gene pools that are further apart mix less often. Introducing a dummy for contiguous countries, the difference in the absolute latitude and longitude, the difference in

temperature, and the difference in precipitation reduces the coefficient on genetic distance further (Column (3)). This specification also shows that contiguity and the difference in the absolute latitude explain migrant selection better than the geographic distance alone. Contiguity is negatively related to the migrant skill mix as it should be much easier for low-skilled migrants to gather information on and to move to neighboring countries than to countries that are farther away. The reason for the finding that the difference in the absolute latitude matters more is that most of our destination countries are in the Northern hemisphere. Thus, the latitude is a better predictor than the longitude. The difference in the climate variables do not play a role. Overall, geographic distance is indeed a strong predictor of the migrant skill mix, but the coefficient on genetic distance is still positive and highly significant, meaning that genetic distance does not simply proxy geographic features between countries.

Language is closely related to culture. Therefore, it is a major concern that genetic distance is only a proxy for language differences. Column (4) of Table 2 adds the language distance of Isphording and Otten (2013), a dummy for an anglophone speaking destination, and whether the two countries have a common language to the model. In this specification, language distance is also positively correlated with migrant selection. However, this correlation disappears once we control for other variables, for example migrant networks. Interestingly, conditional on language distance, anglophone destinations and country pairs that share a common language show a higher migrant selection. However, the coefficient on genetic distance is not much affected. Thus, genetic distance measures also more than just differences in the language.

Migrant networks are the next candidate which should be heavily influenced by genetic distance – we would expect that migrant networks are larger between countries with a lower genetic distance – but migrant networks should also drive down migration cost and therefore lead to a lower migrant selection. The coefficient on migrant networks has the expected negative sign and shows up as a highly significant predictor of migrant selection. However, the coefficient on genetic distance is largely unaffected by the introduction of this network variable (Column (5)).

The next column, Column (6) of Table 2, introduces the difference in the 80/20 income ratios. Like Grogger and Hanson (2011), we find that this measure is positively correlated with migrant selection.

Legal restrictions on immigration are nowadays widely common. They might have also been evolved over the years according to the cultural distance between countries. Therefore, it could be that these restrictions correlate with genetic distance and migrant selection. Adding a dummy for whether there is a visa restriction in place enters highly significant and positive. Adding further a dummy for a Schengen country pair and a dummy for a former colony do not show up significantly. The introduction of the variables in Column (7) reduces the coefficient on genetic distance again only slightly.

In Column (8) of Table 2, we introduce the inflow of foreigners in the destination country as a measure of how open the country is in general. We also include the inflow of asylum-seekers. The literature on illegal migration uses this indicator as a proxy for illegal migration. However, the coefficient on genetic distance is not affected as both variables enter insignificantly.

The last column, Column (9), shows our fully specified model, which we use in all other applications in the paper. Here, we introduce the difference in the years of schooling and the difference in share of tertiary educated. We see that the difference in the years of schooling is significantly positive and the coefficient on genetic distance is reduced again but remains highly significant.

To sum up, through the introduction of all these variables, we are able to reduced the coefficient of genetic distance by 56 percent from 0.808 to 0.356. However, the coefficient on genetic distance in the full model is still significant, which, according to the discussion above, suggests that genetic distance capture cultural differences over and above those that we can observe. The next section exploits further the robustness of the OLS result with regard to potential endogeneity biases.

4.2 Dealing with Endogeneity

The OLS result in Column (9) of Table 2 describes only a causal effect of genetic distance on migrant selection when genetic distance is uncorrelated with the error term in Equation (3). Following the discussion in Section 3.2, one concern is omitted variable bias in the relationship between genetic distance (measured in 1990) and the migrant skill mix (measured in 2000). Specifically, persistent (selected) migration flows could have led to the genetic distance that we observe in 1990. Not accounting, for example, for persistent migration flows would lead to an upward bias in the coefficient on genetic distance. This is the case because the OLS regression would describe an effect of genetic distance that is mediated through a third variable, which we can not capture entirely. Another problem is measurement error in the genetic distance variable. In fact, because of a substantial, but unmeasured genetic diversity within a country (Ashraf and Galor, 2013), our country-wide (average) measure of genetic distance can only approximate the 'true' genetic distance between two countries. Using a noisy measure for genetic distance leads to a downward bias in the coefficient on genetic distance in the OLS regression. Thus, because of omitted variable bias and measurement error, the overall effect of the bias is unknown in advance.

To address a potential bias, we use the instrumental variables (IV) approach suggested by Spolaore and Wacziarg (2009). As explained in detail in Section 3.2, we exploit the variation in genetic distance in 1500 to purge out the part of genetic distance, which is endogenous. We can see the corresponding IV results in Table 3. The first column replicates the OLS results for comparison. Column (2) shows that the first stage is

very strong with a Kleibergen-Paap F statistic of 271.6. The reduced form shows up highly significant and has the expected positive sign. This reduced form effect already indicates that there is a causal impact of genetic distance on the selection of migrants (Column (3)). Column (4) shows the IV estimation results. We observe that the coefficient is substantially larger than the coefficient in the OLS model, increasing by 48 percent to 0.527. This could be explained by measurement error in the genetic distance variable that leads to a downward biased coefficient in the OLS regression.

[Table 3 here]

However, the absolute size of the coefficient is rather uninformative. Therefore, we perform the following effect size calculation: Recall that we have standardized genetic distance such that the coefficient gives the effect on migrant selection for a one standard deviation increase in genetic distance. Evaluating the increase of the migrant skill mix for the mean country pair (1.805), we see that migrant skill mix increases by 29.2 percent (= 0.527/1.805). The ratio of tertiary to primary educated migrants is 8.6 (= 0.0289/0.003). Thus, increasing the migrant skill mix by 29.2 percent would mean to increase the ratio of tertiary to primary educated migrants by 2.5 tertiary-educated migrants for each primary-educated migrant. The OLS results would only imply an increase of 1.7 tertiary-educated migrants for each primary-educated migrant.

The next three columns show the estimation of a more demanding IV model which includes destination and source fixed effects. However, conceptually, it is questionable whether one would like to use country fixed effects when measuring the extent of migrant selection between country pairs. Source country fixed effects lead to an estimation approach that compares within source countries the extent of migrant selection to the 15 different destination countries. In that sense, the regression answers more the question about sorting into different destinations and not about selection in general (Grogger and Hanson, 2011). Destination fixed effects are more justifiable because the purpose of the paper is to explain the extent of migrant selection in these countries.

Column (5) shows the results with destination fixed effects which could capture, for example, the strictness of immigration policies much better than the dummy for visa restrictions does. The coefficient is lower than the baseline coefficient but still larger than the OLS coefficient.

Column (6) uses source country fixed effects. This leads to a substantial drop in the F statistic on the excluded instrument. As mentioned already, the reason is that we take out the main variation in genetic distance that comes over the variation between source countries (and not between destination countries). The variation in genetic distance between destination countries is not very large as we are only dealing with 15 developed countries; most of them located in Europe. Nevertheless, the F statistic is at least 9.1. Even though the coefficient in this model is comparable to the baseline model without

fixed effects, the coefficient on genetic distance identifies a parameter for the extent of migrant *sorting* due to differences in genetic distance and not for migrant *selection*.

Including both, destination and source fixed effects, we obtain our most restrictive model in Column (7) of Table 3. Note that the F statistic on the excluded instrument is reduced further down to 8.3. The coefficient on genetic distance is much larger than the coefficient without fixed effects. Due to the low F statistic, we might run into a weak instrumental variable problem, which could bias the coefficient on genetic distance.

Although using fixed effects in this application is a questionable strategy and is not supported by the theoretical model derivations of Grogger and Hanson (2011), the exercise rules out a lot of unobservable explanations between country pairs that could drive the relationship between genetic distance and migrant selection.

Hence, at this point, we can conclude that larger genetic distances can be interpreted as education-specific migration costs that are more relevant for low-skilled migrants and much less so for high-skilled migrants. The next section exploits the possibility that the marginal effect of increasing genetic distance is not the same for each level of genetic distance.

4.3 Non-Linearities in Genetic Distance

The IV model above assumes that the effect of genetic distance on migrant selection is linear. This assumption might be wrong when genetic distance does not play a role at very low levels of genetic distance and is increasingly important for larger genetic distances. We explore this issue in two ways: First, we split the sample above and below the median genetic distance. Second, we estimate non-linear IV models by including a squared genetic distance term in the regression model.

Table 4 shows the results for splitting the sample above and below the median genetic distance. Columns (2) and (3) reveal that the baseline effect (see Column (1)) is mainly driven by country pairs above the median genetic distance. We do not find a significant effect for country pairs below the median genetic distance. Columns (4) to (7) show who is reacting to genetic distance, low- (primary educated) or high-skilled (tertiary educated) migrants. In these specifications, we regress the scale of migration by skill level, that is, $\ln \frac{E^j_{sd}}{E^j_s}$ for skill level $j = \{Low, High\}$, on the same model for migrant selection as outlined in Section 3.2. Genetic distances above the median prevent low-skilled migrants from migrating but leave high-skilled migrants largely unaffected (Columns (4) and (6)). This pattern generates the selection result observed for country pairs with a genetic distance above the median.

In contrast, for genetic distances below the median, we observe that both, low- and high-skilled migrants are attracted by a higher genetic distance (Columns (5) and (7)). The effect is slightly stronger for low-skilled migrants. This result is not in line with the interpretation of genetic distance as education-specific migration costs. It seems that for

the group of migrants who look for a destination that is not too far away genetically – or, arguably, culturally – and conditional on geographic controls, language differences, migration networks, wage differentials, and so on, higher genetic distances are a benefit on average. We conclude that at low levels of genetic distance, genetic distance should not be considered as a substantial migration cost. One possible explanation, even though speculative, could be that these migrants are open for a different, although not too distant (cultural) environment compared to what they had back in their home country. This type of migrants might be more in search of a different lifestyle in a different country (Benson and O'Reilly, 2009a,b). A similar explanation could be that migrants are attracted by intercultural interest and therefore prefer to move to countries that are culturally more distant (Krieger and Lange, 2010).

Another way of looking at non-linearities in genetic distance is to take the model in Equation (3) and add a squared term of genetic distance. The instrumental variable vector in this model contains the linear term of genetic distance in 1500 and the same variable squared. The following regression shows the second stage of this model:

$$\ln \frac{E^H_{sd}}{E^L_{sd}} - \ln \frac{E^H_s}{E^L_s} = \beta_0 + \beta \, \widehat{Genetic\ distance}_{sd} + \delta \, \widehat{Genetic\ distance^2_{sd}},\ squared + \mathbf{X'_{sd}}\phi + \mu_{sd} \tag{4}$$

The results of the model are depicted in Table 5. The main term for genetic distance in Column (2) is negative significant and the squared term is positive significant. This indicates that there are indeed significant non-linearities. The non-linear model reveals a u-shape relationship between genetic distance and the migrant skill mix, indicating that the marginal effect is increasing in genetic distance. Columns (3) and (4) are mirroring this result by showing an inverse u-shape relationship between genetic distance and the scale of low- and high-skilled migration. The pattern indicates that marginal effects are positive at low levels of genetic distance and approach zero or eventually become negative at higher levels.

[Table 5 here]

However, interpreting marginal effects in non-linear models is not innocuous because the marginal effect depends on the level of the variable. This can be seen by the first derivative of the model with respect to genetic distance (holding all other variables constant): $\beta_{genetic\ distance} + 2 \cdot \delta_{genetic\ distance,\ squared} \cdot genetic\ distance$. Column (2) computes the linear combinations for three percentile positions (10th, 50th, and 90th) of the genetic distance distribution. For very low percentiles, we see that the marginal effect is negative, indicating that for low levels of genetic distance, relatively higher genetic distances attract relatively more low-skilled than high-skilled migrants. This is confirmed by the scale regressions in Columns (3) and (4). Both groups are attracted by a higher genetic

distance at these low levels. At the country pair with the median genetic distance, the relationship changes and higher genetic distances lead to a more selected migrant skill mix. However, the scale regression still reveal that both groups are attracted by higher genetic distances. Eventually, at the 90th percentile, migrant selection is highly positive with respect to genetic distance. Low-skilled migrants are underrepresented in country pairs with a high genetic distance, whereas high-skilled migrants do not respond to genetic distances at the 90th percentile.

To see the complete picture of the non-linearity, we evaluate the model for each (standardized) genetic distance from 0 to 4.710 in 0.1 steps (see Table 1). The result can be seen in Figure 4. The marginal effects increase with genetic distance and are negative for small distances. This means that for country pairs at low levels of genetic distance, genetic distance does not hinder low-skilled migrants more from moving than high-skilled migrants. This finding is in line with Figure 3 that shows that at low levels of genetic distance, the migrant skill mix is expected to be balanced (or even slightly in favor of primary-educated migrants).

[Figure 4 here]

Based on the parameters from the non-linear model, we calculate that point estimates of marginal effects are positive for (standardized) genetic distances above 0.9. But marginal effects are statistically indistinguishable from zero for genetic distances between 0.33 and 1.24. Because the median of the standardized genetic distance is 1.209, marginal effects are indistinguishable from zero (or slightly negative) for more than 50 percent of the country pairs in the sample. As already revealed by the results above, genetic distances are only a barrier to migration when they are sufficiently high.

We explain that pattern with migration costs that increase disproportionately in genetic distance for low-skilled migrants. Given their willingness to migrate, low-skilled people should move disproportionately more often to countries with low genetic distance. The reason might be that information on cultural differences, proxied by genetic differences, and how to cope with them are more easily available for those countries. But with larger genetic distances, the cultural barriers increase substantially and the migrant flow becomes more and more skilled.

Figure 5 sheds more light on the non-linear selection effect by looking at the scale of migration by education. Marginal effects are estimated by non-linear models, regressing the log odds of emigration for each skill group on the baseline model, including genetic distance and genetic distance squared. Both figures show that the marginal effect is decreasing with genetic distance, which means that an increasing genetic distance leads to lower migrant flows at a decreasing rate in both skill groups. However, the marginal effect line is much steeper for primary-educated than for tertiary-educated migrants. This finding supports our conclusion that genetic distance is an important barrier for low-skilled

migrants at sufficiently high levels of genetic distance. High-skilled migrants do not react at all to genetic distance when its level is above some threshold.

[Figure 5 here]

At lower levels of genetic distance, low-skilled migrants are overrepresented because marginal effects for primary educated migrants are larger than for tertiary educated migrants. This explains why we find negative marginal effects of the migrant skill mix for low genetic distances in Figure 4.

4.4 Extensions and Robustness Checks

In Table 6, we perform several robustness checks. In Columns (2) and (3), we look at two kinds of constraints in the source country that could increase positive selection– even though the source fixed effects models in Table 3 do not show that source-specific unobserved variables are a major problem. On the one hand, poverty constraints could be hindering low-skilled migrants, hence, we include the average predicted poverty rate in the source country. The construction follows Belot and Hatton (2012) and uses data from the World Bank Development Indicators (Column (2)). On the other hand, the political freedom in the source country could be important for high-skilled people. Limited political freedom might push them out of the country. The Freedom House Index for source countries is included in Column (3). The average poverty rate enters significantly with the correct sign whereas the political freedom index is not significant. Confirming Belot and Hatton (2012), poverty in the source country is an important driver of the migrant skill mix and also explains some of the average genetic distance effect. However, genetic distance is still highly significant.

[Table 6 here]

Next, we look in Columns (4) and (5) at two variable sets that might influence the likelihood of interaction between two countries. First, we include differences in religious orientations because we speculate that people migrate more to countries with a similar religious orientation. We control for heterogeneity in religions across a country pair by including differences in the share of Protestants, Muslims and other religions in the population. The data is taken from Ashraf and Galor (2013). Second, we use differences in the economic structure of destination and source countries. Having a similar structure might also stem from a closer culture and might induce more migration flows. For that reason, differences between the destination and the source country in the value added of the agricultural, industrial, and service sector is included. Data is from the World Development Indicators. Including both sets reduces the coefficient on genetic distance down to 0.344 or 0.325, respectively. However, it is still significant at the five percent level.

Column (6) should check the robustness of the results regarding the precision in the measurement of genetic distances between countries. We already argued above that we measure genetic distance between two countries less precisely when there is large within-country genetic diversity. Ashraf and Galor (2013) use the distance to Addis Ababa in East Africa, arguably the cradle of human mankind, to explain genetic and language diversity within a country. A larger migratory distance from a country's capital to Addis Ababa has an adverse effect on within-country genetic diversity because settlers further away from Africa carried only a subset of the overall genetic diversity with them.[17] Therefore, genetic distance should be measured more precisely for country pairs with a high migratory distance to Addis Abeba because of less within-country genetic diversity. For country pairs with small distances to Addis Abeba, however, high within-country genetic diversity makes the measurement of genetic distance between these two countries less precise. We include the migratory difference in the distance to Addis Abeba between the sending and the destination country as a control to capture the difference in the general within-country genetic diversity.[18] Adding this variable, the coefficient on genetic distance is unaffected, however the significance is reduced because of an increased standard error. The difference in the distance to Addis Ababa is not significant.

Column (6) in Table 6 reveals that the average effect of genetic distance is reduced from 0.527 to 0.333 once we include further control variables. However, and more importantly, the non-linear nature of the relationship is hardly affected. Splitting the sample according to above and below the median genetic distance reveals a very similar pattern as before for each of the models. Including a squared term in Column (7) and comparing the results to Column (2) in Table 5 shows also no significant differences. In the extended model, the marginal effect of genetic distance is above zero for genetic distances above 1.1 instead of 0.9. The marginal effects for different positions in the genetic distance distribution show very similar values to.

An important assumption of the model is the assumption of irrelevant alternatives. That means that the estimates should not be influenced by the presence of an alternative destination. We can check the stability of the parameters by piecewise omitting one destination country. Comparing the coefficients and significance levels in Table 7 over the different samples shows, that they are all in the same ballpark.

[Table 7 here]

However, from Table 7, we can also infer that especially excluding countries that have a selective immigration policy in place, like Australia and the USA, reduce the coefficient on genetic distance. The worry is therefore that these countries–despite including a range

[17] Ashraf and Galor (2013) consider intercontinental waypoints, such that migration took place mainly on land and only if not otherwise possible over sea.

[18] Appendix Table A-3 provides correlation results with the genetic distance and other variables.

of variables that should take care of anglophone destination and selective immigration policies–drive the results entirely. Column (1) of Table 8 excludes together four countries (Australia, Canada, USA and UK) known to have particularly selective immigration policies relative to the rest of the destinations in our sample. The coefficient drops substantially but remains positive and highly significant. The effect is also reduced, when we limit the sample to EU member countries only in Column (2). The coefficient increases considerably, when looking at non-EU destinations in Column (3).[19]

Following the argument of Grogger and Hanson (2011), Table 7 also indicates whether zero migration cells are a problem for the estimation. In our sample, Ireland (Column (8)), Sweden (Column (13)), Austria (Column (2)), and Spain (Column (12)) are the countries that account for 75 percent of all zero migration cells. Omitting one destination country after another, we do not observe that the coefficient of interest changes significantly. Omitting all countries together results in an coefficient on genetic distance that is identical to the coefficient in the baseline regression in Column (4) of Table 3.[20] Therefore, we conclude (like in Grogger and Hanson (2011)) that zero migration cells are not an important problem in our analysis.

[Table 8 here]

Table 9 shows that genetic distance cannot be explained by simple non-linearities in geographic distance. Adding geographic distance unlogged, squared or cubic does not affect the results for genetic distance.

[Table 9 here]

We argue in the introduction that the evolution of genetic distances is a persistent process that has established over decades and does not change easily. Thus, genetic distance should only be able to explain the cross-sectional variation in migrant selectivity between countries and not the variation within countries over time. Table 10 shows regressions where we control for the migrant skill mix in 1990, that is, controlling for the lagged dependent variable. We see that this variable is a highly significant predictor of the skill mix in 2000. Over the different specifications, we observe that the coefficient on genetic distance becomes very small and insignificant. This confirms that genetic distance is only able to capture long-run differences and is not able to explain short-run changes in migrant selectivity.

[19]Looking within the EU only to ensure that there are no major legal barriers to migration is not meaningful. The reason is that the variation among European countries in genetic distance is rather low. We checked a more detailed genetic distance matrix for European populations (also from Cavalli-Sforza et al. (1994)), which can be more easily matched to countries. However, the median genetic distance among European countries is smaller than the 10th percentile of the world sample. Thus, the analysis of non-linearities in genetic distance implies that we should not see a large effect from genetic distance on migrant selection for EU member states. By using the European data, we can confirm our prior that genetic distance is not able to explain selective migration within the EU.

[20]Results are available from the authors upon request.

[Table 10 here]

5 Conclusion

This paper provides evidence on the role of country pair relatedness, measured by bilateral genetic distance, for the selectivity of international migration. We show that country pairs with a higher genetic distance experience more selected migration on average. Dual inheritance theory in social anthropology suggests that genetic distance may proxy differences in deeply rooted cultural traits, norms and beliefs of societies. Hence, our finding of migrant selection can be explained through greater difficulties for low-skilled workers to overcome cultural differences.

The size of the estimated effects is substantial: Increasing genetic distance by one standard deviation (which corresponds, e.g., to the change in genetic distance when switching from the USA-Germany to USA-Mexico country pair) would mean to increase the ratio of tertiary to primary educated migrants by 2.5 tertiary-educated migrants for each primary-educated migrant. This conclusion is robust to the introduction of several control variables as well as to an instrumental variables approach that exploits variation in genetic distance before large migration waves in 1500.

In addition, digging deeper into the data reveals important non-linearities in genetic distance. The average result is driven by country pairs above the median genetic distance. Country pairs below the median genetic distance do not show selected migration stocks. Thus, genetic distance has to be sufficiently large to constitute education-specific migration costs. Non-linear IV models confirm this pattern. The effect is driven by low-skilled migrants who avoid countries that are genetically too distant. High-skilled migrants do not respond to genetic distances at levels above the median. Interestingly, we find that both groups are attracted by higher genetic distances at genetic distance levels below the median at almost similar proportions. This finding is compatible with some sort of lifestyle migration and that these migrants are open for a different, although not too distant (cultural) environment compared to what they had back in their home country.

The paper remains speculative on the specific mechanisms that ultimately explain why migrants choose specific destinations according to genetic distance. Additional fixed effects specifications exclude factors that are destination and/or source country specific. We also control for several potential bilateral transmission channels, such as, language differences, migrant networks, religious similarities, industry similarities, and immigration policies. None of these channels can explain the effect of genetic distance on the migrant skill mix entirely. At the same time, several of these factors have previously been named as cultural differences between countries (e.g., language or religion). Arguably, there are still some cultural factors that the researcher cannot observe or measure and that are explained by genetic distance. This can be (more complex) informal networks or cultural

norms, traits, and habits, which are not explained by the observables. Therefore, genetic distance has been shown to be a good proxy for a very comprehensive concept of cultural differences and to be able to predict the migration behavior of broader population groups.

Hence, there are hard to detect differences between countries beyond purely economic factors that influence individual migration decisions, thereby causing specific selection patterns. Since these patterns are deeply rooted in the populations' norms and beliefs systems, they tend to shape migration flows in a 'natural' way. If innovation-driven societies aim to attract relatively more high-skilled than low-skilled workers, they need to acknowledge that genetic or, rather, cultural differences in a broad sense may foster or impede the inflow of immigrants at different skill levels from specific countries. Simply opening the country for immigrant workers will not necessarily lead to an inflow of the most desirable workers in terms of skill composition. Immigration policies to attract high-skilled migrants might fail if they do not acknowledge and factor in cultural differences between countries.

References

Abramitzky, R. (2009). The Effect of Redistribution on Migration: Evidence from the Israeli Kibbutz. *Journal of Public Economics*, 93(3-4):498–511.

Adsera, A. and Pytlikova, M. (2014). The Role of Language in Shaping International Migration: Evidence from OECD Countries 1985-2006. *Economic Journal*, forthcoming.

Alesina, A., Devleeschauwer, A., Easterly, W., Kurlat, S., and Wacziarg, R. (2003). Fractionalization. *Jornal of Economic Growth*, 8(2):155–194.

Alesina, A., Guiliano, P., and Nunn, N. (2013). On the Origins of Gender Roles: Women and the Plough. *Quarterly Journal of Economics*, 128(2):469–530.

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

Barro, R. and Lee, J.-W. (2013). A new data set of educational attainment in the world, 1950-2010. *Journal of Development Economics*, 104:184–198.

Bauernschuster, S., Falck, O., Heblich, S., Suedekum, J., and Lameli, A. (2014). Why Are Educated and Risk-Loving Persons More Mobile Across Regions? *Journal of Economic Behavior & Organization*, 98:56–69.

Beine, M., Docquier, F., and Özden, Ç. (2011). Diasporas. *Journal of Development Economics*, 95(1):30–41.

Belot, M. and Ederveen, S. (2011). Cultural Barriers in Migration between OECD Countries. *Journal of Population Economics*, 25(3):1077–1105.

Belot, M. and Hatton, T. J. (2012). Immigrant Selection in the OECD. *Scandinavian Journal of Economics*, 114(4):1105–1128.

Benson, M. and O'Reilly, K., editors (2009a). *Lifestyle Migration. Expectations, Aspirations and Experiences*. Ashgate.

Benson, M. and O'Reilly, K. (2009b). Migration and the Search for a Better Way of Life: A Critical Exploration of Lifestyle Migration. *Sociological Review*, 57(4):608–625.

Borjas, G. J. (1987). Self-Selection and the Earnings of Immigrants. *American Economic Review*, 77(4):531–553.

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

Burchardi, K. B. and Hassan, T. A. (2013). The Economic Impact of Social Ties: Evidence from German Reunification. *Quarterly Journal of Economics*, 128(3):1219–1271.

Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2011). Robust Inference With Multiway Clustering. *Journal of Business & Economic Statistics*, 29(2):238–249.

Cavalli-Sforza, L. L. (2001). *Genes, Peoples and Languages*. Penguin Group.

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

Chambers, E. G., Foulon, M., Handfield-Jones, H., Hankin, S. M., and Michaels, E. G. (1998). The War for Talent. *McKinsey Quarterly*, 3:44–57.

Chiquiar, D. and Hanson, G. H. (2005). International Migration, Self-Selection, and the Distribution of Wages: Evidence from Mexico and the United States. *Journal of Political Economy*, 113(2):239–281.

Coe, D. T. and Helpman, E. (1995). International R&D Spillovers. *European Economic Review*, 39(5):859–887.

Comin, D. A., Dmitriev, M., and Rossi-Hansberg, E. (2012). The Spatial Diffusion of Technology. NBER Working Paper 18534.

Dahl, M. S. and Sorenson, O. (2010). The Migration of Technical Workers. *Journal of Urban Economics*, 67(1):33–45.

Docquier, F., Lowell, B. L., and Marfouk, A. (2007). A Gendered Assessment of the Brain Drain. Mimeo.

Falck, O., Heblich, S., Lameli, A., and Südekum, J. (2012). Dialects, Cultural Identity, and Economic Exchange. *Journal of Urban Economics*, 72(2-3):225–239.

Falck, O., Lameli, A., and Ruhose, J. (2015). Cultural Biases in Migration: Estimating Non-Monetary Migration Costs. IZA Discussion Paper No. 8922.

Felbermayr, G. J. and Toubal, F. (2010). Cultural Proximity and Trade. *European Economic Review*, 54(2):279–293.

Fernández-Huertas Moraga, J. (2011). New Evidence on Emigrant Selection. *Review of Economics and Statistics*, 93(1):72–96.

Giuliano, P., Spilimbergo, A., and Tonon, G. (2014). Genetic Distance, Transportation Costs, and Trade. *Journal of Economic Geography*, 14(1):179–198.

Gould, E. D. and Moav, O. (2014). Does High Inequality Attract High Skilled Immigrants? *Economic Journal*, forthcoming.

Grogger, J. and Hanson, G. H. (2011). Income Maximization and the Selection and Sorting of International Migrants. *Journal of Development Economics*, 95(1):42–57.

Guiso, L., Sapienza, P., and Zingales, L. (2006). Does Culture Affect Economic Outcomes? *Journal of Economic Perspectives*, 20(2):23–48.

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

Head, K., Mayer, T., and Ries, J. (2010). The Erosion of Colonlia Trade Linkages after Independence. *Journal of International Economics*, 81(1):1–14.

Henrich, J. and McElreath, R. (2003). The Evolution of Cultural Evolution. *Evolutionary Anthropology*, 12(3):123–135.

Isphording, I. E. and Otten, S. (2013). The Costs of Babylon: Linguistic Distance in Applied Economics. *Review of International Economics*, 21(2):354–369.

Kaestner, R. and Malamud, O. (2014). Self-Selection and International Migration: New Evidence from Mexico. *Review of Economics and Statistics*, 96(1):78–91.

Krieger, T. and Lange, T. (2010). Education Policy and Tax Competition with Imperfect Student and Labor Mobility. *International Tax and Public Finance*, 17(6):587–606.

Mayda, A. M. (2009). International Migration: A Panel Data Analysis of the Determinants of Bilateral Flows. *Journal of Population Economics*, 23(4):1249–1274.

McFadden, D. (1974). The Measurement of Urban Travel Demand. *Journal of Public Economics*, 3(4):303–328.

Nelson, R. R. and Phelps, E. S. (1966). Investment in Humans, Technological Diffusion, and Economic Growth. *American Economic Review*, 56(1/2):69–75.

Neumayer, E. (2006). Unequal Access to Foreign Spaces: How States Use Visa Restrictions to Regulate Mobility in a Globalized World. *Transactions of the British Institute of*

Geographers, 31(3):72–84.

OECD (2014). International Migration Outlook 2014. OECD Publishing, http://dx.doi.org/10.1787/migr_outlook-2014-en.

Ottaviano, G. I. and Peri, G. (2005). The Economic Value of Cultural Diversity: Evidence from US Cities. *Journal of Economic Geography*, 6(1):9–44.

Parey, M., Ruhose, J., Waldinger, F., and Netz, N. (2015). The Selection of High-Skilled Migrants. IZA Discussion Paper No. 9164.

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

Spolaore, E. and Wacziarg, R. (2012). Long-Term Barriers to the International Diffusion of Innovations. In Frankel, J. and Pissarides, C., editors, *NBER International Seminar on Macroeconomics*, chapter 1, pages 11–46. University of Chicago Press.

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

Spolaore, E. and Wacziarg, R. (2015). Ancestry, language, and culture. CESifo Working Paper No. 5388.

Spring, E. and Grossmann, V. (2015). Does Bilateral Trust Across Countries Really Affect International Trade and Factor Mobility? *Empirical Economics*, forthcoming.

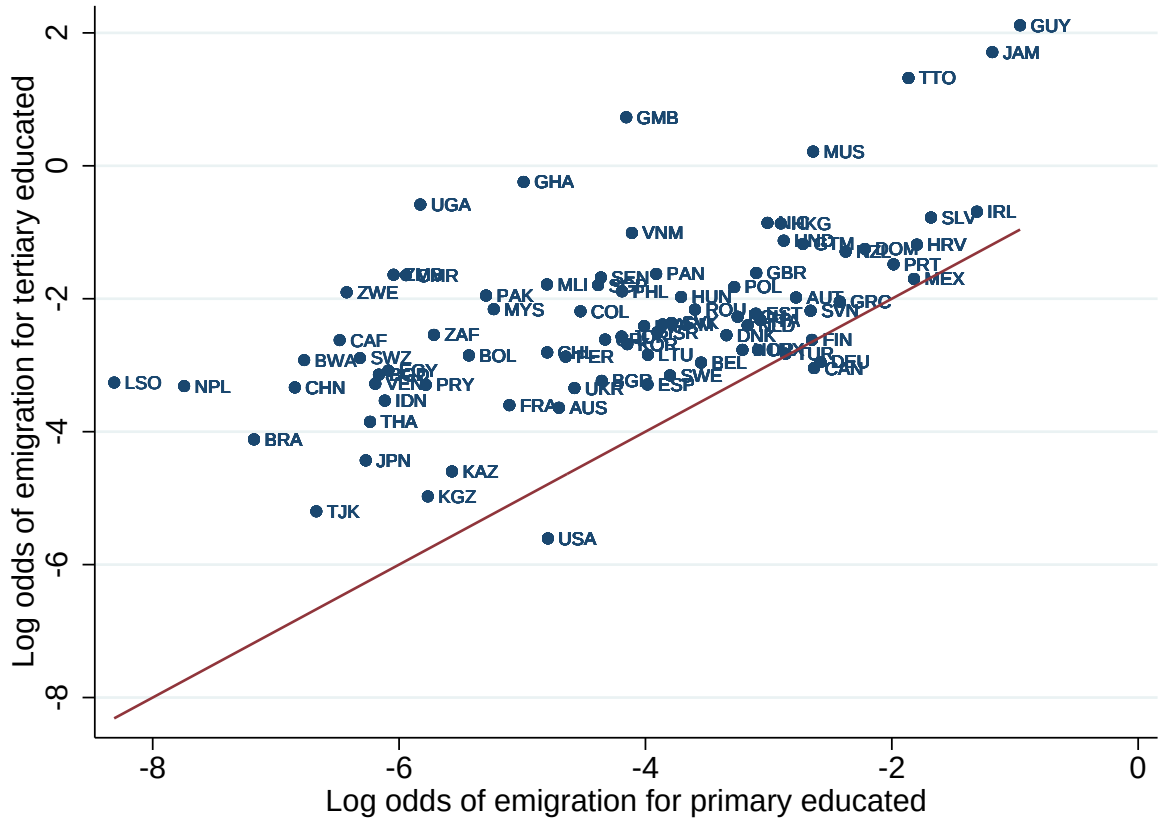
Stolz, Y. and Baten, J. (2012). Brain Drain in the Age of Mass Migration: Does Relative Inequality Explain Migrant Selectivity? *Explorations in Economic History*, 49(2):205–220.

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

Voigtländer, N. and Voth, H.-J. (2012). Persecution Perpetuated: The Medieval Origins of Anti-Semitic Violence in Nazi Germany. *Quarterly Journal of Economics*, 127:1229–1392.

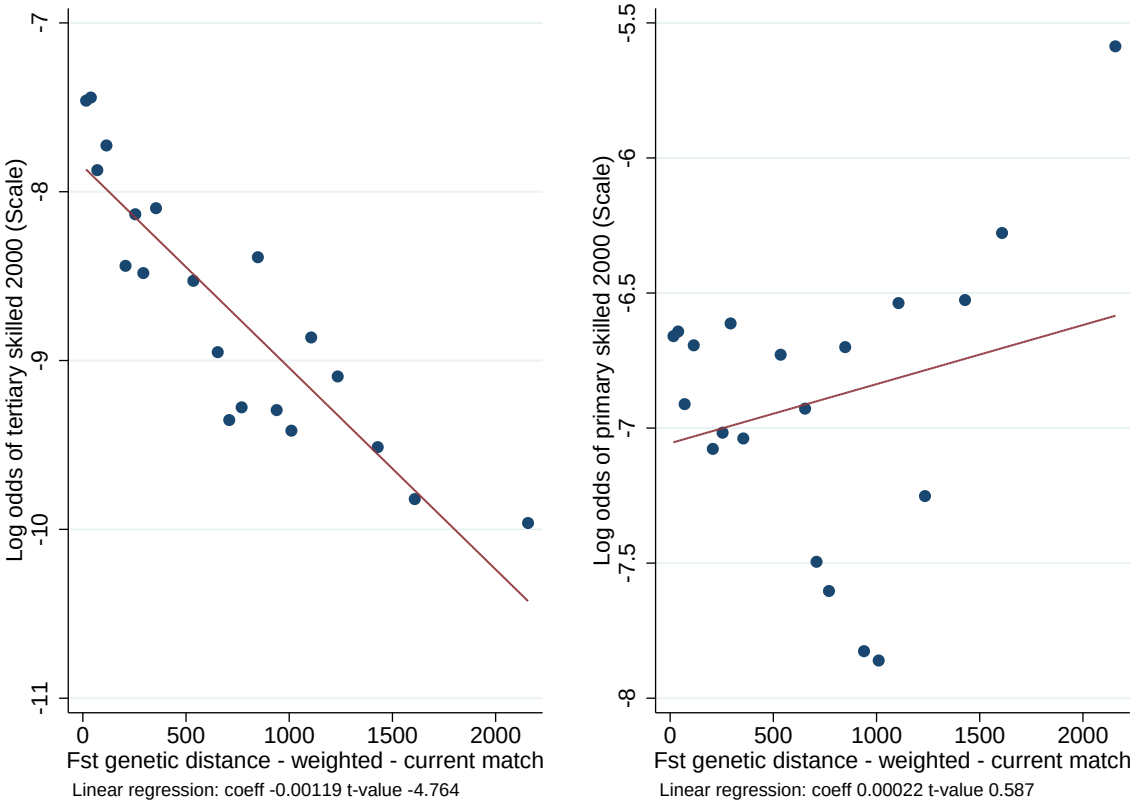
Figures and Tables

Figure 1: ***Emigration odds (primary and tertiary-educated) by source country, 2000***



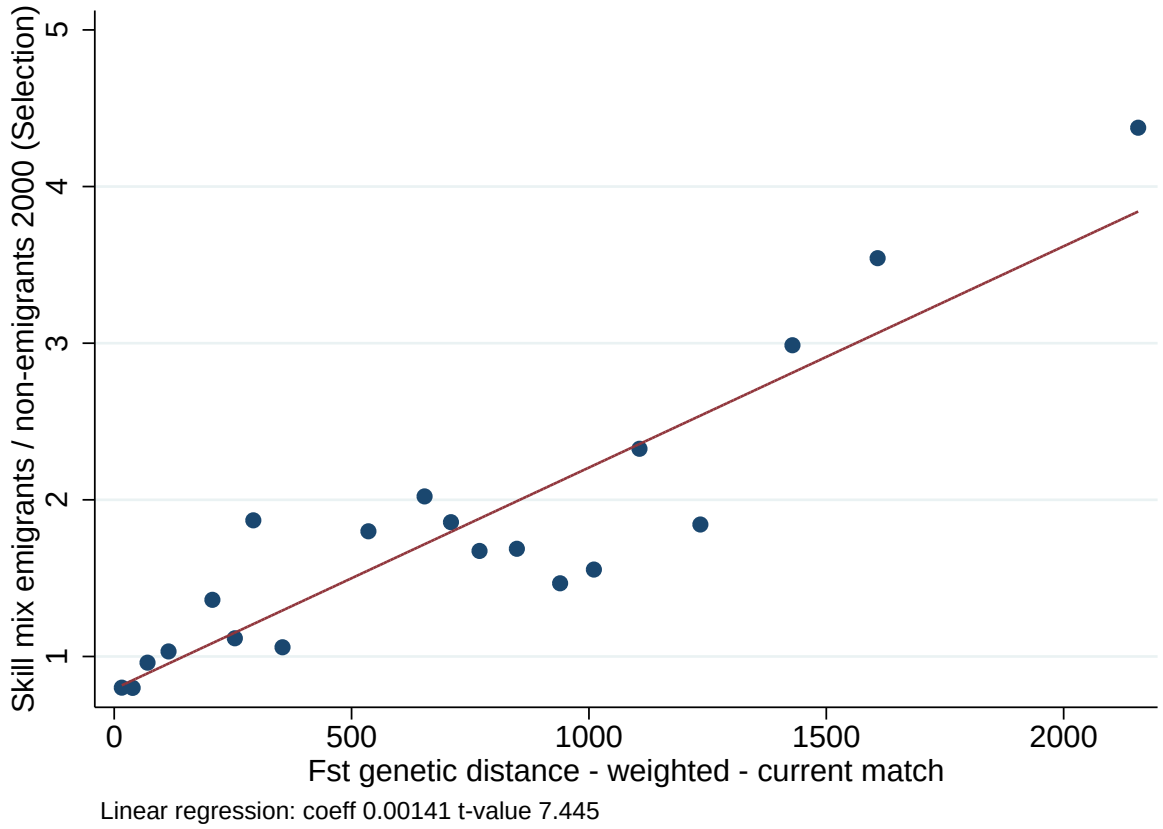
Notes: The figure shows the log odds of emigration for primary and tertiary educated migrants. Source: Docquier et al. (2007). The graph is replicated from Grogger and Hanson (2011).

Figure 2: ***Genetic distance and emigration odds by skill level, 2000***



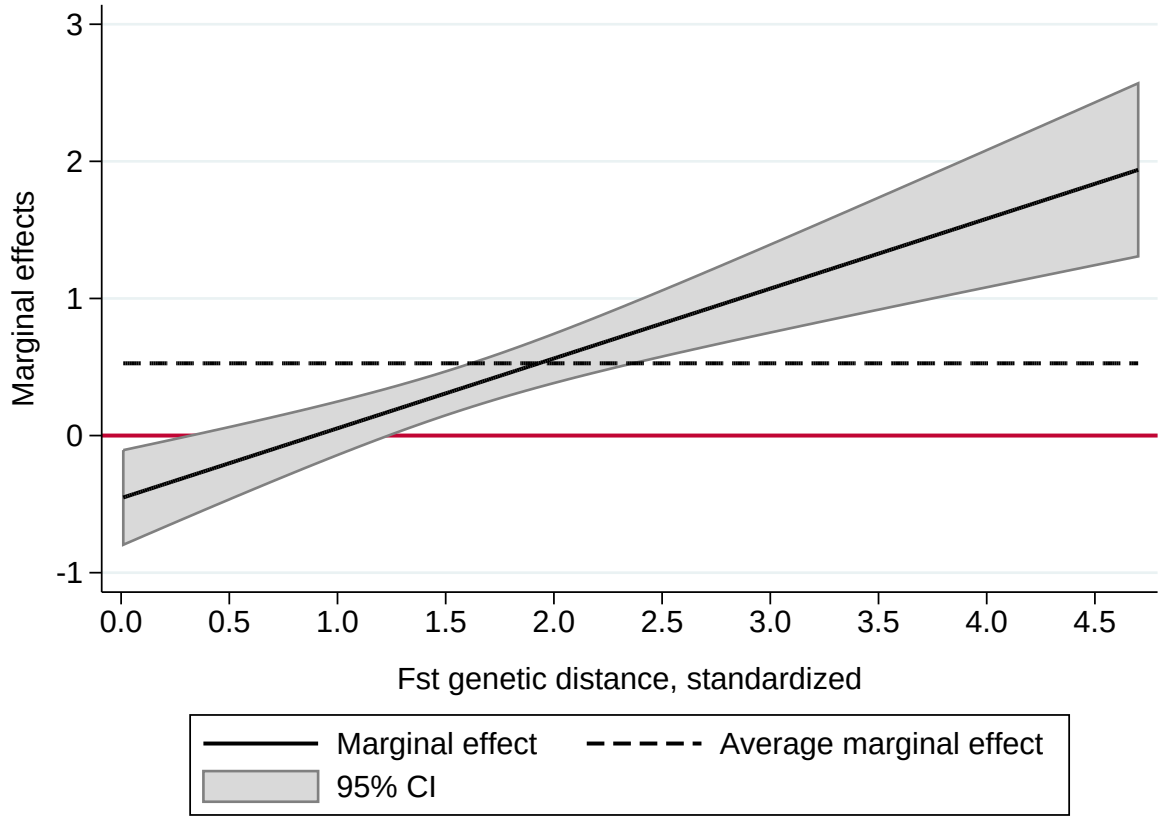
Notes: These non-parametric binned scatter plots show the relationship between the log odds of emigration and genetic distance. The figure on the left shows the relationship between the log odds of emigration for tertiary educated migrants and genetic distance and the figure on the right shows the relationship between the log odds of emigration for primary educated migrants and genetic distance. For both figures, the coefficients and t-statistics are from OLS regressions with the microdata. We bin genetic distance into 20 bins of equal size and then the means of genetic distance and log emigration odds are obtained within each bin. Data of migrant stocks by skill level is from Docquier et al. (2007) and the genetic distance data is from Spolaore and Wacziarg (2009).

Figure 3: ***Genetic distance and emigration selection, 2000***



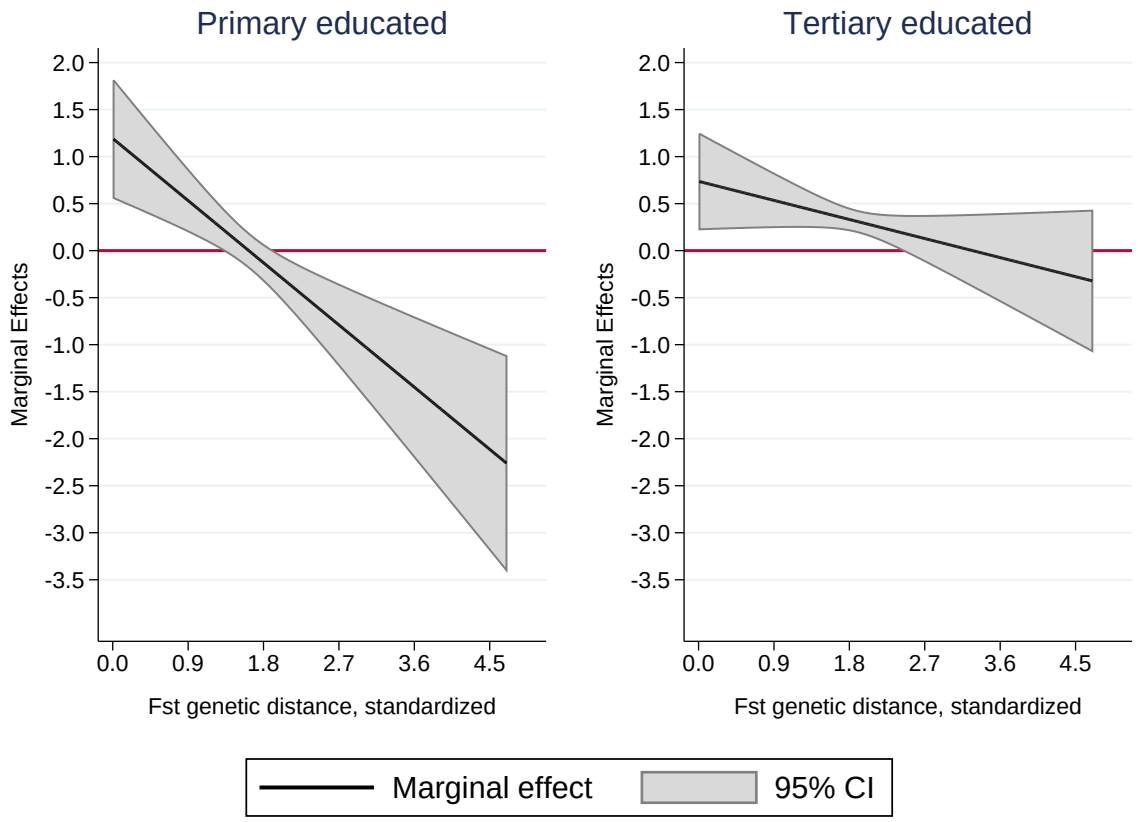
Notes: The non-parametric binned scatter plot shows the relationship between the skill mix of migrants and genetic distance. For the figure, the coefficients and t-statistics are from OLS regressions with the microdata. We bin genetic distance into 20 bins of equal size and then the means of genetic distance and log emigration odds are obtained within each bin. Data of migrant stocks by skill level is from Docquier et al. (2007) and the genetic distance data is from Spolaore and Wacziarg (2009).

Figure 4: ***Non-linear effects between genetic distance and the migrant skill mix***



Notes: The figure shows the relationship between the level of genetic distance and the marginal effect on the selection of migrants. Genetic distance is standardized, which means that it is divided by the standard deviation of genetic distance. The dashed line represents the average marginal effect which is obtained from the baseline model. For the other coefficients, we have estimated a non-linear model, including genetic distance and genetic distance, squared. This model is then evaluated at each level of genetic distance. Marginal effects are computed by $\beta_{genetic\ distance} + 2 \cdot \delta_{genetic\ distance,\ squared} \cdot genetic\ distance$.

Figure 5: ***Non-linear effects between genetic distance and the log odds of emigration***



Notes: The figure shows the relationship between the level of genetic distance and the marginal effect on the scale of migrants by skill level. Genetic distance is standardized, which means that it is divided by the standard deviation of genetic distance. Marginal effects are estimated by non-linear models, regressing the log odds of emigration for each skill group on the baseline model, including genetic distance and genetic distance, squared. This model is then evaluated at each level of genetic distance. Marginal effects are computed by $\beta_{genetic\ distance} + 2 \cdot \delta_{genetic\ distance,\ squared} \cdot genetic\ distance$.

Table 1: ***Summary Statistics***

Variable	Mean	Std. Dev.	Min.	Max.	Obs.
Panel A: Genetic Distance Data					
F_{ST} genetic distance	716	572	0	2,695	1,102
F_{ST} genetic distance, standardized	1.252	1	0	4.710	1,102
F_{ST} genetic distance, 1500	989	650	0	3,557	1,102
Panel B: Migrant Selection Measures					
Primary-educated emigration share, $\left(E^L_{sd}/E^L_s\right)$	0.003	0.016	0	0.252	1,102
Tertiary-educated emigration share, $\left(E^H_{sd}/E^H_s\right)$	0.029	0.228	0	4.798	1,102
Migrant skill mix, $\left(\ln \frac{E^H_{sd}}{E^L_{sd}} - \ln \frac{E^H_s}{E^L_s}\right)$	1.805	1.567	-3.041	8.077	1,102
Panel C: Controls					
Log distance	1.567	1.040	0.035	4.229	1,102
Contiguous	0.030	–	0	1	1,102
Δ absolute latitude	-9.274	68.585	-172	172	1,102
Δ absolute longitude	18.738	20.728	-39	62.633	1,102
Δ temperature	-8.546	10.787	-36.513	29.264	1,102
Δ precipitation	-27.232	64.335	-202.626	135.989	1,102
Language distance	86.732	23.890	0	105.270	1,102
Anglophone destination	0.406	–	0	1	1,102
Common language	0.166	–	0	1	1,102
Migrant networks	0.004	0.018	0	0.263	1,102
Δ 80/20 wage ratio	22.632	10.276	-7.934	47.999	1,102
Visa restriction	0.477	–	0	1	1,102
Schengen pair	0.159	–	0	1	1,102
Colony	0.064	–	0	1	1,102
Log inflow foreigners	11.286	1.236	8.979	13.421	1,102
Log inflow asylum-seekers	9.670	1.117	7.332	11.463	1,102
Δ years of schooling	2.637	2.975	-4.571	11.984	1,102
Δ share tertiary	6.116	8.842	-22.836	30.437	1,102
Political Freedom	2.798	1.651	1	7	1,082
Average poverty rate, predicted	25.507	18.2450	4.760	75.395	1,029
Δ share Agriculture	-8.154	10.880	-54.953	8.688	1,061
Δ share Industry	-1.779	9.352	-29.719	29.580	1,050
Δ share Service	9.859	13.521	-31.587	47.960	1,050
Δ share Catholics	-4.750	51.066	-96.800	96.900	1,102
Δ share Protestants	28.590	40.874	-97.7	97.800	1,087
Δ share Muslims	-12.8633	28.431	-99.2	3	1,102
Δ share Other Religion	-11.3012	38.134	-98.2	68.7	1,087
Δ distance Addis Ababa	-0.061	1.004	-2.277	1.994	1,102

Notes: Δ represents the simple difference between destination and source country, that is, $\Delta X = X_d - X_s$. When they are used in the regression models, F_{ST} genetic distance, F_{ST} genetic distance, 1500, language distance, and geographic distance are standardized such that they have a standard deviation of 1 over all country pairs in the sample. Standard deviations are not reported for dummy variables. See Appendix Table A-1 for variable definitions and data sources.

Table 2: ***OLS Results of Migrant Selection***

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
F_{ST} genetic distance$_{sd}$	0.808***	0.660***	0.507***	0.468***	0.481***	0.448***	0.411***	0.408***	0.356***
	(0.109)	(0.114)	(0.125)	(0.095)	(0.098)	(0.084)	(0.089)	(0.072)	(0.060)
Log geographic distance$_{sd}$		0.356***	0.189	0.025	–0.006	–0.008	0.000	–0.001	–0.063
		(0.079)	(0.116)	(0.096)	(0.090)	(0.078)	(0.083)	(0.110)	(0.091)
Contiguous$_{sd}$			–0.720***	–0.809***	–0.762***	–0.611**	–0.653**	–0.726***	–0.888***
			(0.215)	(0.206)	(0.206)	(0.234)	(0.227)	(0.212)	(0.190)
Δ absolute latitude$_{sd}$			0.008***	0.005**	0.005**	0.005***	0.005***	0.005**	0.002
			(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)
Δ absolute longitude$_{sd}$			0.011	0.019**	0.017*	0.016**	0.015*	0.023***	0.012**
			(0.009)	(0.008)	(0.009)	(0.008)	(0.007)	(0.006)	(0.004)
Δ temperature$_{sd}$			–0.025	–0.012	–0.016	–0.009	–0.010	–0.002	0.003
			(0.017)	(0.016)	(0.017)	(0.012)	(0.012)	(0.013)	(0.010)
Δ precipitation$_{sd}$			–0.000	–0.000	–0.000	–0.001	–0.001	–0.000	0.001
			(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
Language distance$_{sd}$				0.107*	0.077	0.001	–0.025	–0.035	–0.086
				(0.051)	(0.055)	(0.054)	(0.055)	(0.055)	(0.064)
Anglophone destination$_{d}$				0.769**	0.817**	0.688**	0.751**	0.614	0.817***
				(0.294)	(0.291)	(0.278)	(0.273)	(0.364)	(0.213)
Common language$_{sd}$				0.440**	0.503***	0.485**	0.465**	0.414**	0.381**
				(0.167)	(0.152)	(0.174)	(0.184)	(0.159)	(0.171)
Migrant networks$_{sd}$					–8.425***	–11.980***	–12.140***	–13.325***	–11.851***
					(2.371)	(2.556)	(2.545)	(3.238)	(3.095)
Δ 80/20 wage ratio$_{sd}$						0.033**	0.028**	0.026**	0.007
						(0.012)	(0.011)	(0.011)	(0.011)
Visa restriction$_{sd}$							0.374***	0.314***	0.054
							(0.115)	(0.099)	(0.103)
Schengen pair$_{sd}$							0.222	0.247	0.341**
							(0.153)	(0.164)	(0.151)
Colony$_{sd}$							–0.035	–0.091	–0.081
							(0.136)	(0.132)	(0.138)
Log inflow foreigners$_{d}$								0.197	0.036
								(0.229)	(0.237)
Log inflow asylum-seekers$_{d}$								–0.024	0.111
								(0.233)	(0.268)
Δ years of schooling$_{sd}$									0.235***
									(0.038)
Δ share tertiary$_{sd}$									0.008
									(0.013)
R^2	0.265	0.305	0.441	0.488	0.496	0.531	0.539	0.552	0.643
Observations	1,102	1,102	1,102	1,102	1,102	1,102	1,102	1,102	1,102
Cluster	15	15	15	15	15	15	15	15	15

Notes: The dependent variable is the migrant skill mix in 2000, i.e. $\ln\left(E^H_{sd}/E^L_{sd}\right) - \ln\left(E^H_s/E^L_s\right)$. F_{ST} genetic distance, geographic distance, and language distance are standardized such that they have a standard deviation of 1 over all country pairs in the sample. Δ represents the simple difference between destination and source country, that is, $\Delta X = X_d - X_s$. Robust standard errors in parentheses clustered at the destination country level. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

Table 3: ***IV Results***

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Selection	Genetic Distance	Selection	Selection	Fixed effects		
	OLS	FS	RF	IV	destination	source	d & s
F_{ST} genetic distance$_{sd}$	0.356***			0.527***	0.413***	0.579**	0.796**
	(0.060)			(0.113)	(0.093)	(0.284)	(0.352)
F_{ST} genetic distance$_{sd}$, 1500		0.783***	0.412***				
		(0.048)	(0.076)				
Control variables	YES	YES	YES	YES	YES	YES	YES
R^2	0.643	0.755	0.647	0.637	0.709	0.803	0.811
Observations	1,102	1,102	1,102	1,102	1,102	1,102	1,102
Cluster	15	15	15	15	15	15	15
Kleibergen-Paap F statistic				271.6	361.2	9.1	8.3

Notes: The dependent variable is the migrant skill mix in 2000, i.e. $\ln\left(E^H_{sd}/E^L_{sd}\right) - \ln\left(E^H_s/E^L_s\right)$. Control variables: Log geographic distance, contiguous, Δ absolute latitude, Δ absolute longitude, Δ temperature, Δ precipitation, language distance, anglophone destination, common language, migrant networks, Δ 80/20 wage ratio, visa restriction, schengen pair, colony, log inflow foreigners, log inflow asylum-seekers, Δ years of schooling, Δ share tertiary. F_{ST} genetic is standardized such that it has a standard deviation of 1 over all country pairs in the sample. Robust standard errors in parentheses clustered at the destination country level. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

Table 4: ***Effect Heterogeneities by Genetic Distance***

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Selection	Selection		Scale Primary		Scale Tertiary	
	baseline	above	below	above	below	above	below
F_{ST} genetic distance$_{sd}$	0.527***	0.960***	–0.240	–0.986***	1.405**	–0.026	1.165***
	(0.113)	(0.159)	(0.313)	(0.182)	(0.549)	(0.147)	(0.423)
Control variables	YES	YES	YES	YES	YES	YES	YES
R^2	0.637	0.677	0.626	0.667	0.662	0.790	0.709
Observations	1,102	551	551	551	551	551	551
Cluster	15	15	15	15	15	15	15
Kleibergen-Paap F statistic	271.6	100.7	74.1	100.7	74.1	100.7	74.1

Notes: The dependent variable *Selection* is the migrant skill mix in 2000, i.e. $\ln\left(E^H_{sd}/E^L_{sd}\right) - \ln\left(E^H_s/E^L_s\right)$. The dependent variables *Scale Primary* and *Scale Tertiary* are the log odds of emigration for low-skilled or high-skilled migrants, respectively, i.e. $\ln\left(E^L_{sd}/E^L_s\right)$ or $\ln\left(E^H_{sd}/E^H_s\right)$, respectively. *Above* and *below* indicate sample splits by above and below the median genetic distance. Control variables: Log geographic distance, contiguous, Δ absolute latitude, Δ absolute longitude, Δ temperature, Δ precipitation, language distance, anglophone destination, common language, migrant networks, Δ 80/20 wage ratio, visa restriction, schengen pair, colony, log inflow foreigners, log inflow asylum-seekers, Δ years of schooling, Δ share tertiary. F_{ST} genetic is standardized such that it has a standard deviation of 1 over all country pairs in the sample. Robust standard errors in parentheses clustered at the destination country level. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

Table 5: ***Non-Linearities in Genetic Distance***

	(1)	(2)	(3)	(4)
	Selection	Selection	Scale Primary	Scale Tertiary
F_{ST} genetic distance$_{sd}$	0.527***	–0.456***	1.194***	0.737***
	(0.113)	(0.177)	(0.321)	(0.261)
F_{ST} genetic distance$_{sd}$, squared		0.255***	–0.367***	–0.113*
		(0.050)	(0.094)	(0.067)
Control variables	YES	YES	YES	YES
R^2	0.637	0.658	0.670	0.746
Observations	1,102	1,102	1,102	1,102
Cluster	15	15	15	15
Kleibergen-Paap F statistic	271.6	47.4	47.4	47.4
F_{ST} genetic distance at				
10th percentile		-0.412**	1.129***	0.717***
		(0.169)	(0.306)	(0.249)
50th percentile		0.159*	0.305**	0.465***
		(0.089)	(0.123)	(0.108)
90th percentile		0.855***	-0.698***	0.157
		(0.129)	(0.089)	(0.108)

Notes: The dependent variable *Selection* is the migrant skill mix in 2000, i.e. $\ln\left(E^H_{sd}/E^L_{sd}\right) - \ln\left(E^H_s/E^L_s\right)$. The dependent variables *Scale Primary* and *Scale Tertiary* are the log odds of emigration for low-skilled or high-skilled migrants, respectively, i.e. $\ln\left(E^L_{sd}/E^L_s\right)$ or $\ln\left(E^H_{sd}/E^H_s\right)$, respectively. Control variables: Log geographic distance, contiguous, Δ absolute latitude, Δ absolute longitude, Δ temperature, Δ precipitation, language distance, anglophone destination, common language, migrant networks, Δ 80/20 wage ratio, visa restriction, schengen pair, colony, log inflow foreigners, log inflow asylum-seekers, Δ years of schooling, Δ share tertiary. F_{ST} genetic is standardized such that it has a standard deviation of 1 over all country pairs in the sample. Robust standard errors in parentheses clustered at the destination country level. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

Table 6: ***Further Mechanisms and Robustness Checks***

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
F_{ST} genetic distance$_{sd}$	0.527***	0.424***	0.421***	0.344**	0.325**	0.333*	–0.638***
	(0.113)	(0.124)	(0.123)	(0.158)	(0.163)	(0.171)	(0.185)
F_{ST} genetic distance$_{sd}$, squared							0.289***
							(0.057)
Average poverty rate$_s$, predicted		0.014***	0.012***	0.012***	0.009**	0.010**	0.007
		(0.003)	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)
Political freedom$_s$			0.040	0.035	0.074***	0.074***	0.078***
			(0.027)	(0.024)	(0.019)	(0.019)	(0.023)
Δ share protestants$_{sd}$				–0.007***	–0.007***	–0.007***	–0.006***
				(0.002)	(0.002)	(0.003)	(0.002)
Δ share muslims$_{sd}$				–0.001	0.000	–0.001	–0.003
				(0.003)	(0.002)	(0.003)	(0.003)
Δ share other religion$_{sd}$				–0.000	–0.000	–0.001	–0.003**
				(0.001)	(0.002)	(0.001)	(0.002)
Δ share industry$_{sd}$					0.032***	0.030***	0.030***
					(0.008)	(0.008)	(0.007)
Δ share service$_{sd}$					0.011**	0.010*	0.017***
					(0.006)	(0.006)	(0.005)
Δ distance Addis Ababa$_{sd}$						–0.100	–0.272**
						(0.131)	(0.134)
Control variables	YES	YES	YES	YES	YES	YES	YES
R^2	0.637	0.613	0.613	0.634	0.643	0.643	0.656
Observations	1,102	1,029	1,009	994	942	942	942
Cluster	15	15	15	15	15	15	15
Kleibergen-Paap F statistic	271.6	197.2	202.9	136.7	121.6	114.0	57.7
F_{ST} genetic distance							
above median genetic distance	0.960***	0.957***	0.959***	0.847***	0.899***	1.004***	
	(0.159)	(0.203)	(0.206)	(0.199)	(0.176)	(0.215)	
below median genetic distance	–0.240	-0.139	-0.260	-0.481	-0.457	-0.437	
	(0.313)	(0.347)	(0.327)	(0.380)	(0.355)	(0.354)	
F_{ST} genetic distance at							
10th percentile							-0.587***
							(0.177)
50th percentile							0.061
							(0.124)
90th percentile							0.851***
							(0.201)

Notes: The dependent variable is the migrant skill mix in 2000, i.e. $\ln\left(E^H_{sd}/E^L_{sd}\right) - \ln\left(E^H_s/E^L_s\right)$. Control variables: Log geographic distance, contiguous, Δ absolute latitude, Δ absolute longitude, Δ temperature, Δ precipitation, language distance, anglophone destination, common language, migrant networks, Δ 80/20 wage ratio, visa restriction, schengen pair, colony, log inflow foreigners, log inflow asylum-seekers, Δ years of schooling, Δ share tertiary. F_{ST} genetic is standardized such that it has a standard deviation of 1 over all country pairs in the sample. Δ represents the simple difference between destination and source country, that is, $\Delta X = X_d - X_s$. Robust standard errors in parentheses clustered at the destination country level. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

Table 7: ***Omitting Destination Countries***

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
	AUS	AUT	CAN	DNK	FIN	FRA	DEU	IRL	NLD	NZL	NOR	ESP	SWE	USA	UK
F_{ST} genetic distance$_{sd}$	0.443***	0.532***	0.501***	0.558***	0.573***	0.539***	0.552***	0.543***	0.494***	0.495***	0.548***	0.495***	0.556***	0.381***	0.524***
	(0.104)	(0.114)	(0.118)	(0.114)	(0.121)	(0.126)	(0.117)	(0.116)	(0.108)	(0.114)	(0.117)	(0.108)	(0.113)	(0.066)	(0.124)
Control variables	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
R^2	0.623	0.635	0.623	0.633	0.648	0.646	0.633	0.650	0.642	0.639	0.641	0.678	0.640	0.659	0.633
Observations	1,020	1,053	1,021	1,021	1,027	1,018	1,025	1,058	1,018	1,030	1,021	1,037	1,043	1,018	1,018
Cluster	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14
Kleibergen-Paap F statistic	327.0	260.2	239.3	256.5	225.5	295.6	252.6	249.8	293.6	286.8	254.5	290.9	254.0	308.1	281.2

Notes: The country above each column is omitted as a destination country. The dependent variable is the migrant skill mix in 2000, i.e. $\ln\left(E^H_{sd}/E^L_{sd}\right) - \ln\left(E^H_s/E^L_s\right)$. Control variables: Log geographic distance, contiguous, Δ absolute latitude, Δ absolute longitude, Δ temperature, Δ precipitation, language distance, anglophone destination, common language, migrant networks, Δ 80/20 wage ratio, visa restriction, schengen pair, colony, log inflow foreigners, log inflow asylum-seekers, Δ years of schooling, Δ share tertiary. F_{ST} genetic is standardized such that it has a standard deviation of 1 over all country pairs in the sample. Robust standard errors in parentheses clustered at the destination country level. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

Table 8: ***Migrant Selection toward Selected Country Groups***

	(1)	(2)	(3)
	Excl. AUS, CAN, USA, UK	EU Destinations	Non-EU Destinations
F_{ST} genetic distance$_{sd}$	0.265***	0.211***	0.690***
	(0.061)	(0.060)	(0.139)
Control variables	YES	YES	YES
R^2	0.591	0.590	0.725
Observations	771	702	400
Cluster	11	10	5
Kleibergen-Paap F statistic	819.1	939.8	661.3

Notes: The dependent variable is the migrant skill mix in 2000, i.e. $\ln\left(E^H_{sd}/E^L_{sd}\right) - \ln\left(E^H_s/E^L_s\right)$. Control variables: Log geographic distance, contiguous, Δ absolute latitude, Δ absolute longitude, Δ temperature, Δ precipitation, language distance, anglophone destination, common language, migrant networks, Δ 80/20 wage ratio, visa restriction, schengen pair, colony, log inflow foreigners, log inflow asylum-seekers, Δ years of schooling, Δ share tertiary. F_{ST} genetic is standardized such that it has a standard deviation of 1 over all country pairs in the sample. Robust standard errors in parentheses clustered at the destination country level. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

Table 9: ***Non-Linearities in Geographic Distance***

	(1)	(2)	(3)	(4)
F_{ST} genetic distance$_{sd}$	0.527***	0.478***	0.529***	0.528***
	(0.113)	(0.105)	(0.105)	(0.103)
Log geographic distance$_{sd}$	–0.108			
	(0.082)			
Geographic distance$_{sd}$		–0.026	–0.452**	–0.837**
		(0.075)	(0.220)	(0.423)
Geographic distance$_{sd}$, squared			0.111***	0.360
			(0.043)	(0.242)
Geographic distance$_{sd}$, cubic				–0.043
				(0.040)
Control variables	YES	YES	YES	YES
R^2	0.637	0.639	0.638	0.639
Observations	1,102	1,102	1,102	1,102
Cluster	15	15	15	15
Kleibergen-Paap F statistic	271.6	260.2	269.1	265.5

Notes: The dependent variable is the migrant skill mix in 2000, i.e. $\ln\left(E^H_{sd}/E^L_{sd}\right) - \ln\left(E^H_s/E^L_s\right)$. Control variables: Log geographic distance, contiguous, Δ absolute latitude, Δ absolute longitude, Δ temperature, Δ precipitation, language distance, anglophone destination, common language, migrant networks, Δ 80/20 wage ratio, visa restriction, schengen pair, colony, log inflow foreigners, log inflow asylum-seekers, Δ years of schooling, Δ share tertiary. F_{ST} genetic and geographic distance are standardized such that they have a standard deviation of 1 over all country pairs in the sample. Robust standard errors in parentheses clustered at the destination country level. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

Table 10: ***Controlling for Migrant Selection in 1990***

	(1)	(2)	(3)
	OLS		IV
F_{ST} genetic distance$_{sd}$	0.042	–0.004	0.064
	(0.042)	(0.046)	(0.064)
Migrant skill mix$_{sd}$, 1990	0.797***	0.738***	0.724***
	(0.031)	(0.035)	(0.035)
Control variables	–	YES	YES
R^2	0.845	0.878	0.877
Observations	1,053	1,053	1,053
Cluster	15	15	15
Kleibergen-Paap F statistic			199.1

Notes: The dependent variable is the migrant skill mix in 2000, i.e. $\ln\left(E^H_{sd}/E^L_{sd}\right) - \ln\left(E^H_s/E^L_s\right)$. Control variables: Log geographic distance, contiguous, Δ absolute latitude, Δ absolute longitude, Δ temperature, Δ precipitation, language distance, anglophone destination, common language, migrant networks, Δ 80/20 wage ratio, visa restriction, schengen pair, colony, log inflow foreigners, log inflow asylum-seekers, Δ years of schooling, Δ share tertiary. F_{ST} genetic is standardized such that it has a standard deviation of 1 over all country pairs in the sample. Robust standard errors in parentheses clustered at the destination country level. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

A Appendix

Table A-1: ***Variable Definitions and Sources***

Variable	Definition	Source
F_{ST} genetic distance	Average variation in the frequencies of 120 alleles between two populations matched to countries according to majority populations.	Spolaore and Wacziarg (2009)
F_{ST} genetic distance, 1500	Average variation in the frequencies of 120 alleles between two populations matched to countries, with majority populations as of 1500.	Spolaore and Wacziarg (2009)
Migration data	Data on migrant stocks (aged 25 and above) in 2000 with skill levels.	Docquier et al. (2007)
Log geographic distance	Population-weighted great circle distance between large cities of the two countries.	Head et al. (2010)
Contiguous	Dummy equal to 1 if both countries share a common border.	Head et al. (2010)
Δ absolute latitude	Absolute value of the latitude of a country's approximate geodesic centroid.	Ashraf and Galor (2013)
Δ absolute longitude	Absolute value of the longitude of a country's approximate geodesic centroid.	Ashraf and Galor (2013)
Δ temperature	Difference in average monthly temperatures between source and destination country, measured in degree Celsius from 1961-1990.	Ashraf and Galor (2013)
Δ precipitation	Difference in average monthly precipitation between source and destination country, measured in degree Celsius from 1961-1990.	Ashraf and Galor (2013)
Language distance	Global percentage of dissimilarity in the pronunciation of words with the same meaning in two languages, the value is averaged over 40 words.	Isphording and Otten (2013)
Anglophone destination	Dummy equal to 1 if English is the first official language.	Own research
Common language	Dummy equal to 1 if destination and source country share a language that is spoken by at least 9 per cent of the population.	Head et al. (2010)
Migrant networks	Ration of the stock of migrants from a source country summed over all education levels to the residents in the source country summed over all education levels	Own calculations with the data from Docquier et al. (2007)
Δ 80/20 wage ratio	Difference in wage differences, i.e. between high- and low-skilled wages.	Grogger and Hanson (2011)
Visa restriction	Dummy equal to 1 if there are visa restrictions imposed by the destination on a source country.	Neumayer (2006)
Schengen pair	Dummy equal to 1 if both countries are signatories of the Schengen agreement.	Own research
Colony	Dummy equal to 1 if the countries have ever been in a colonial relationship.	Head et al. (2010)
Inflow foreigners	Inflow of foreign population from 216 source countries in 1999, measured in 1000	Own calculation, data from the International Migration Dataset, OECD
Inflow asylum seekers	Inflow of asylum seekers from 216 source countries in 1999, based on data provided by the United Nations High Commission for Refugees	Own calculation, data from the International Migration Dataset, OECD
Δ years of schooling	Difference in the average years of schooling attained	Barro and Lee (2013)
Δ share tertiary	Difference in the percentage of complete tertiary education in population	Barro and Lee (2013)
Political freedom	Index between 1 and 7 measuring the degree of political freedom in the source country. 1 is free, 7 is not free.	Freedom House Index 1999-2000
Average poverty rate, predicted	Prediction of the average poverty rate in source countries by the regression of the average poverty rate (the share of population living with less than two Dollars per day) on the average share of employees in the agricultural sector. Both are averaged over the years 1980-2000. Calculation as by Belot and Hatton (2012).	Data from World Bank Development Indicators
Δ share agriculture	Difference in the value added as percentage of GDP of the agricultural sector between destination and source country.	Own calculation, data from the WDI
Δ share industry	Difference in the value added as percentage of GDP of the industrial sector between destination and source country.	Own calculation, data from the WDI
Δ share service	Difference in the value added as percentage of GDP of the service sector between destination and source country.	Own calculation, data from the WDI
Δ share protestants	Difference in the percentage of the population being Protestant.	Ashraf and Galor (2013)
Δ share catholics	Difference in the percentage of the population being Catholic.	Ashraf and Galor (2013)
Δ share muslims	Difference in the percentage of the population being Muslim.	Ashraf and Galor (2013)
Δ share other religion	Difference in the percentage of the population belonging to any other religion than Catholic, Protestant or Muslim.	Ashraf and Galor (2013)
Δ distance Addis Ababa	Difference in the migratory distance to East Africa. Calculated as the great circle distance from Addis Ababa in East Africa, Ethopia, to the capital of each country as long as possible along land and following specified waypoints. Measured in 1000 km.	Ashraf and Galor (2013)

Notes: Δ represents the simple difference between destination and source country, that is, $\Delta X = X_d - X_s$.

Table A-2: ***Source Countries by World Region***

North America 12	South America 9	Asia 20	Europe 27	Africa 15	Oceania 2
Canada	Bolivia	Armenia	Austria	Botswana	Australia
Costa Rica	Brazil	Bangladesh	Belgium	Cameroon	New Zealand
Dominican Republic	Chile	China	Bulgaria	Central Afr. Rep.	
El Salvador	Colombia	Hong Kong	Croatia	Egypt	
Guatemala	Ecuador	Indonesia	Denmark	Gambia	
Honduras	Guyana	Israel	Estonia	Ghana	
Jamaica	Paraguay	Japan	Finland	Lesotho	
Mexico	Peru	Jordan	France	Mali	
Nicaragua	Venezuela	Kazakhstan	Germany	Mauritius	
Panama		Korea	Greece	Senegal	
Trinidad and Tobago		Kyrgyzstan	Hungary	South Africa	
USA		Malaysia	Ireland	Swaziland	
		Nepal	Italy	Uganda	
		Pakistan	Latvia	Zambia	
		Philippines	Lithuania	Zimbabwe	
		Singapore	Luxembourg		
		Tajikistan	Netherlands		
		Thailand	Norway		
		Turkey	Poland		
		Vietnam	Portugal		
			Romania		
			Slovakia		
			Slovenia		
			Spain		
			Sweden		
			Ukraine		
			UK		

Table A-3: ***Correlations for Selected Variables***

Variables	F_{ST} genetic distance	Language distance	Common language	Log distance	Contiguity	Colony	Δ share catholics	Δ share protes-tants	Δ share muslims	Δ share other religion	Δ distance Addis Ababa
F_{ST} genetic distance	1										
Language distance	0.2499	1									
Common language	0.1861	-0.4397	1								
Log distance	0.4291	0.0617	0.1815	1							
Contiguity	-0.1582	-0.1656	0.0647	-0.3870	1						
Colony	0.0455	-0.3300	0.3994	0.0068	0.1057	1					
Δ share catholics	0.0724	0.0184	0.0685	-0.0247	-0.0079	0.0488	1				
Δ share protestants	0.0745	0.2405	-0.1876	0.0635	-0.0947	-0.1749	-0.5649	1			
Δ share muslims	-0.0740	-0.2045	-0.0177	-0.0845	0.0811	0.0157	-0.3177	-0.1549	1		
Δ share other religion	-0.1214	-0.1326	0.1257	0.0391	0.0528	0.1121	-0.4973	-0.1992	-0.1607	1	
Δ distance Addis Ababa	0.0259	0.1169	0.1742	-0.0297	0.0106	-0.0955	0.2728	-0.1502	-0.2022	-0.0612	1.0000

Notes: Δ represents the simple difference between destination and source country, that is, $\Delta X = X_d - X_s$.