NBER WORKING PAPER SERIES

MIGRANTS, ANCESTORS, AND INVESTMENTS

Konrad B. Burchardi
Thomas Chaney
Tarek A. Hassan

Working Paper 21847
http://www.nber.org/papers/w21847

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
January 2016

We are grateful to Richard Hornbeck, Emir Kamenica, and Nancy Qian for helpful discussions. We also thank seminar participants at the CEPR's summer trade meeting in Paris (ERWIT), Singapore Management University, National University of Singapore, Universitat Pompeu Fabra, IFN (Stockholm), Harvard and at the University of Chicago for their comments. Chaney is grateful for financial support from ERC grant N<337272FiNet. Hassan is grateful for financial support from the IGM and the Fama-Miller Center at the University of Chicago. All mistakes remain our own. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Migrants, Ancestors, and Investments
Konrad B. Burchardi, Thomas Chaney, and Tarek A. Hassan
NBER Working Paper No. 21847
January 2016
JEL No. F21,G15,J61,L14,N3,O11

ABSTRACT

We use 130 years of data on historical migrations to the United States to show a causal effect of the ancestry composition of US counties on foreign direct investment (FDI) sent and received by local firms. To isolate the causal effect of ancestry on FDI, we build a simple reduced-form model of migrations: migrations from a foreign country to a US county at a given time depend on (i) a push factor, causing emigration from that foreign country to the entire United States, and (ii) a pull factor, causing immigration from all origins into that US county. The interaction between time-series variation in country-specific push factors and county-specific pull factors generates quasi-random variation in the allocation of migrants across US counties. We find that a doubling of the number of residents with ancestry from a given foreign country relative to the mean increases by 4.2 percentage points the probability that at least one local firm invests in that country, and increases by 31% the number of employees at domestic recipients of FDI from that country. The size of these effects increases with the ethnic diversity of the local population, the geographic distance to the origin country, and the ethno-linguistic fractionalization of the origin country.

Konrad B. Burchardi
Institute for International Economic Studies
Stockholm University
SE-106 91 Stockholm
Sweden
konrad.burchardi@iies.su.se

Thomas Chaney
Department of Economics
University of Chicago
1126 East 59th Street
Chicago, IL 60637
thomas.chaney@gmail.com

Tarek A. Hassan
Booth School of Business
University of Chicago
5807 South Woodlawn Avenue
Chicago, IL 60637
and NBER
tarek.hassan@chicagobooth.edu

Over the last two centuries, the United States has attracted migrations from all corners of the earth, making it one of the most ethnically diverse nations on the planet. As of the 2010 census, 63% of US residents could trace their ancestry to a foreign country, including descendants of migrants from virtually all countries in the world. What are the long-term effects of immigration? In this paper, we quantify the causal impact of the ancestry composition of US counties, defined as the presence in the United States of descendants of foreign migrants, on their patterns of foreign direct investment (FDI).

We find a sizable causal impact of foreign ancestry on FDI. For an average US county, doubling the number of individuals with ancestry from a given foreign country increases by 4.2 percentage points (or 237% relative to the mean) the probability that at least one firm from that US county engages in FDI with that country. At the intensive margin, it increases by 10% the number of firms engaged in FDI, and increases by 31% the number of local jobs at subsidiaries of firms headquartered in the origin country. This effect appears strongly concave, where going from 1,000 to 10,000 descendants increases the probability of an FDI link by a factor of 3, suggesting higher diversity of origins may increase the overall level of FDI activity in a given destination.

To assess the *causal* impact of foreign ancestry on the patterns of FDI, we follow an instrumental variables (IV) strategy using the history of migrations into the United States. We isolate quasi-random variation in the allocation of migrants across destinations within the United States that results from the interaction of two facts: First, migrants from different origins tended to arrive in the United States at different times. Second, the set of destinations that are most economically attractive to the typical migrant arriving in the United States changed over time.

We motivate our approach using a simple reduced-form dynamic model of migrations. Migrations from a given foreign origin country o to a given destination county d in period t depend on the total number of migrants arriving in the United States from o (a push factor), the relative economic attractiveness of destination county d to migrants arriving at the time (a pull factor), and the size of the pre-existing local population of ancestry o, allowing for the fact that migrants tend to prefer settling near others of their own ethnicity (a recursive factor). Solving the model shows the number of residents today who are descendants of migrants from o is a function of simple and higher-order interactions of the sequence of pull and push factors. We then use these interactions to construct instruments for each county's present-day ancestry composition.

To prevent omitted variables that affect both migrations and FDI from driving our results, we measure the pull factor of each US destination for migrants from o, using the number migrants arriving in d at the same time from a continent other than o's continent of origin. That is, we predict a migrant's choice of destination within the United States using the revealed behavior

of the average migrant arriving at the same time but from a different continent. Similarly, we measure the push factor using the total number of migrants arriving in the United States from o at time t, excluding those who settle in the vicinity (the census region) of d. Interacting these measures of pull and push factors for each vintage of census data since 1880, we are able to construct a set of instruments that isolate quasi-random variation in today's ancestry composition of US counties that derives solely from the interaction of the staggered arrival of migrants from different origins with time-series variations in the relative attractiveness of different destinations.

To illustrate our approach, consider the examples of migrations from Ireland and Germany. The peak of Irish migrations to the United States occurred in the mid-19th century, when the Great Potato Famine triggered massive migrations from Ireland. At that time, the eastern half of the United States was attracting large numbers of foreign migrants, including non-Europeans. We predict a large population with Irish ancestry in the eastern half of the United States today, as subsequent waves of Irish migrants were more likely to settle where previous Irish migrants had gone. The peak of German migrations occurred several decades later, at the end of the 19th century. At that time, the Midwest was booming and attracting large numbers of migrants. We predict a large population with German ancestry in the Midwest, which persists to this day. The difference in the timing of migrations from Ireland and Germany, together with the time variation in the spatial distribution of economic development within the United States, generates quasi-random variation in the predicted distribution of ancestry across the United States.

Our main identifying assumption for a causal interpretation of our results is that omitted variables making a given location within the United States differentially more attractive for migrants from a specific origin for both settlement and FDI do not affect the location choices of the average migrant originating from other continents and *simultaneously* have large effects outside of the surrounding states of the destination in question.

Flexibly applying our instrumentation strategy to the entire set of origins and destinations allows us to corroborate our identifying assumption in various ways and to guard against a range of potentially confounding factors using fixed effects. We are able to simultaneously control for both origin and destination fixed effects, thus controlling for all origin and destination-specific factors, such as differences in size, market access, and productivity. Moreover, regional interactions of the origin and destination fixed effects also allow us to address a wide range of threats to our identifying assumption. In addition, we conduct a range of falsification exercises and robustness checks. For example, we obtain quantitatively similar effects of ancestry on FDI if we condition only on migrations from former communist countries that occurred during a period of economic isolation between the origin country and the United States.

Our approach delivers the statistical power and a sufficiently large number of instruments to reliably estimate heterogeneous effects of ancestry on FDI across origins, destinations, and sectors. Across origin countries, the effect is almost always positive and significant and of similar size. Among the five largest origin countries, the estimates range from 0.174 (s.e.=0.011) for Mexico to 0.265 (s.e.=0.009) for the UK. The effect increases with the geographic distance and the judicial quality of the origin country, and with its ethnolinguistic fractionalization. Once these characteristics are accounted for, other measures of genetic, linguistic, and religious distance do not significantly affect the size of the effect of ancestry on FDI. The effect is similar across destination counties, but increases with the degree of ethnic diversity of the local population. In addition, the effect is heterogeneous across sectors, being about five times larger for manufacturing than for natural resources.

We also find evidence of negative spillovers. The effect of ancestry on FDI falls with the population of migrants from the same origin in neighboring counties, and from neighboring origins. For example, a large Polish contingent in a given county has a lower effect on FDI if nearby counties also host large contingents from Poland, or if this county hosts a large contingent from the Czech Republic.

Because we can instrument separately for each wave of immigration, we are also able to distinguish the effect of first-generation immigrants from the effect of their descendants. We find a significantly smaller effect for the first generation, implying the full effect of ancestry on FDI is long lasting and takes multiple generations to fully unfold.

By aggregating our data set to the state level, we study the effect of ancestry on trade flows between US states and foreign countries. Although our results with respect to FDI remain largely stable at the state level, and in contrast to previous studies, we find no systematic causal impact of ancestry on the patterns of international trade once we control for state fixed effects.

We view our results as convincing evidence that migration, and the ethnic diversity resulting from it, affect the pattern of international investment in a quantitatively important way. This effect appears to unfold and persist over several generations, and to increase in size even after the first generation of immigrants has passed. Our evidence on heterogeneous effects, negative spillovers, and the concavity of the effect is consistent with a positive impact of regional diversity on FDI: the more diverse and less ethnically homogeneous the local population, the larger the total effect of ancestry on FDI. It is also consistent with the "strength of weak ties" (Granovetter, 1973), where a small minority from a distant part of the world that is not otherwise represented in the local ethnic mix has the largest marginal impact on FDI.

To illustrate the quantitative implications of our results, we conduct two thought experiments.

In the first, we calculate the effect of Chinese exclusion – the effective ban on Chinese immigration between 1882 and 1965 – on the extensive margin of FDI between the United States and China today. In the second, we calculate the effect of a hypothetical "L.A. gold rush" – an early population growth in Los Angeles before 1880 – for the intensive margin of FDI links between the city and various foreign countries.

Finally, we note two important limitations to our analysis. First, our results rely purely on cross-sectional variation in FDI within the United States. Although we believe that, in light of our results, the ethnic diversity of the United States likely also raises the extent of FDI for the United States as a whole, we cannot exclude the possibility that increases in FDI in one county are partially or fully offset by decreases in FDI elsewhere. Second, our results provide only indirect evidence on the nature of the mechanisms linking ancestry composition to FDI. Directly quantifying the relative roles of underlying microeconomic mechanisms, such as the transmission of information via co-ethnic networks versus the provision of social collateral, and enforceability of contracts, requires a structural model or microeconomic data at a higher level of disaggregation than we have available. Some of our earlier work speaks precisely to these issues. Burchardi and Hassan (2013) present evidence that social ties have a causal impact on economic development and FDI. Chaney (2014a) structurally estimates a model in which the percolation of information about foreign trading partners drives firm-level exports.

Existing literature. A large body of literature shows that measures of affinity between regions, such as common ancestry, social ties, trust, and telephone volume, correlate strongly with aggregate economic outcomes, such as foreign direct investment (Guiso et al., 2009), international asset flows (Portes and Rey, 2005), and trade flows (Gould, 1994; Rauch and Trindade, 2002).[1] How much of this association should be interpreted as causal, however, remains an open question because these measures of affinity, and ancestry in particular, are likely to be nonrandom across regions. Two recent papers use historical decisions by the US government on the location of Japanese internment camps during World War II (Cohen et al., 2015) and the placement of Vietnamese refugees after the Vietnam War (Parsons and Vezina, 2014) to identify a causal effect of concentrations of descendants of these migrants on contemporary trade flows between locations within the United States and Japan and Vietnam, respectively. Burchardi and Hassan (2013) use variation in wartime destruction across West German regions in 1945, a time when millions

[1] Combes et al. (2005) and Garmendia et al. (2012) study the impact of migrations on trade using variation only within France and Spain, respectively. Head and Ries (1998) study the impact of migrations on trade of Canadian provinces. Aleksynska and Peri (2014) distinguish the impact of migrations on trade by migrants' occupation. Docquier et al. (2014) distinguish between potential and actual migrations.

of refugees were arriving from East Germany, as an instrument for the share of the population with social ties to the East, and show evidence of a causal effect of these social ties on GDP growth, entrepreneurial activity, and FDI in East Germany after the fall of the Berlin Wall.[2]

We contribute to this literature in several ways. First, we identify a causal effect of ancestry on FDI in a setting with a high degree of external validity directly relevant for assessing, for example, the long-term effects of immigration policy. Second, because our identification strategy can be flexibly applied to the entire set of origin countries, we are able to guard against a wide range of possible confounding factors and to relate directly to the previous literature by employing a gravity equation that features both destination and origin fixed effects. Third, the increased statistical power allows us to identify the determinants of heterogeneity in the effect of ancestry on FDI across origins,[3] destinations, and sectors, and show evidence of negative geographic spillovers. Fourth, our results suggest the causal effects of ancestry on FDI and trade flows may be very different, although they appear similar in ordinary least squares (OLS) regressions.

Our paper also contributes to the debate on the costs and benefits of immigration. Much of the existing literature has focused on the effects of migration on local labor markets.[4] A more recent literature focuses on the effect of cultural, ethnic, and birthplace diversity on economic development and growth.[5] Most closely related is Fulford et al. (2015) who study the effect of ancestry composition of US counties on GDP growth. We add to this literature by examining the effect of migration on the pattern of economic exchange and employment. By looking at the composition of US residents with *foreign ancestry*, as opposed to just *foreign-born* residents, we are able to separate the short-term and long-term effects of migration. In this sense, our results show a long-term effect of migration on the comparative advantage in FDI of different regions that operates across multiple generations and may explain part of the association between diversity and long-term growth found in other studies.

On a methodological level, our paper relates to a wider literature that has employed leave-out shift-share instruments in the identification of causal effects (Bartik, 1991; Katz and Murphy, 1992). Our instrumentation strategy, based on a simple recursive model of migrations, can easily be replicated for other countries, other time periods, or variables other than migrations where cumulated flows matter, without the need for a rare or even unique historical accident.

[2] See Fuchs-Schuendeln and Hassan (2015) and Chaney (2014b) for surveys of this literature.

[3] For instance, using the measure of fractionalization from Alesina et al. (2003), we find ancestry matters more for FDI to and from countries with a higher degree of fractionalization.

[4] See for example Card (1990), Card and Di Nardo (2000), Friedberg (2001), Borjas (2003), and Cortes (2008). Borjas (1994) provides an early survey.

[5] Ottaviano and Peri (2006), Putterman and Weil (2010), Peri (2012), Ashraf and Galor (2013), Ager and Brückner (2013), Alesina et al. (2015), and Alesina et al. (Forthcoming).

The remainder of this paper is structured as follows. Section 1 introduces our data. Section 2 gives a brief overview of the history of migration to the United States. Section 3 identifies the causal effect of ancestry composition on FDI at the extensive margin, discusses various challenges to our identifying assumption, and conducts a range of robustness checks and falsification exercises. Section 4 probes the effect of ancestry composition on FDI at the intensive margin and other outcomes, and illustrates the quantitative implications of our findings using two counterfactual experiments. Section 5 examines the mechanism underlying the effect of ancestry on FDI by testing for spill over effects and by probing the heterogeneity of the effect across countries, counties, and sectors. Section 6 concludes.

1 Data

We collect data on migrations and ancestry, on foreign direct investment and trade, and on origin and destination characteristics. Below is a description of our data, along with their source.

Migrations and ancestry. All migration and ancestry data are constructed from the individual files of the Integrated Public Use Microdata Series (IPUMS) samples of the 1880, 1900, 1910, 1920, 1930, 1970, 1980, 1990, and 2000 waves of the US census, and the 2006-2010 five-year sample of the American Community Survey. We use population data to construct transition matrices between historic countries and US counties (or county groups from 1970 onwards) to countries and counties in their 1990 borders.

Throughout the paper, we use $t-1$ to denote the census wave just before t, o for the foreign country of origin, and d for the US destination county. We construct the number of migrants from origin o to destination d at time t, $I^t_{o,d}$, by counting the number of respondents who live in d, were born in o, and emigrated to the United States between t and $t-1$. The exception to this rule is the 1880 census, which did not record the year of immigration. The variable $I^{1880}_{o,d}$ instead measures the number of residents who were born in o or whose parents were born in o.[6] Since 1980, respondents have also been asked about their ancestry in both the US Census and the American Community Survey, and they can provide multiple answers. $Ancestry_{o,d}$ corresponds to the number of individuals residing in d at time t who report o as first ancestry. Appendix A.1 gives further details on the construction of our migration and ancestry data.

[6] If the own birthplace is in the United States, imprecisely specific (e.g., a continent) or missing, we instead use the parents' birthplace, assigning the person half to his/her father's birthplace and half to his/her mother's birth place.

Foreign direct investment. Our data on FDI is from the US file of the 2014 edition of the Bureau van Dijk ORBIS data set. For each US firm, the database lists the location of its (operational) headquarters, the addresses of its foreign parent entities, and the addresses of its international subsidiaries and branches. We use a 5% ownership threshold to define a parent-subsidiary link. Altogether we have information on 36,108 US firms that have at least one foreign parent or subsidiary. Collectively, these firms have 102,618 foreign parents and 176,332 foreign subsidiaries in 142 countries (in their 1990 borders).[7] We then aggregate this information to the county level. Our main outcome variable, *FDI Dummy*, is 1 if at least one firm within a given destination county has at least one parent or subsidiary in the origin country. For each destination, we also count the total number of FDI linkages (the total number of foreign parents and subsidiaries of all firms within the county), and the total number of unique parents and subsidiaries in both the origin and the destination. We also count the total number of employees working at firms with a foreign parent in a given destination (*# of Employees at Subsidiaries in Destination*) and the total number of employees working at subsidiaries of US firms in a given origin country (*# of Employees at Subsidiaries in Origin*).[8] The ORBIS database also gives the 2007 NAICS code of the sector of the US firm, allowing us to disaggregate these data by 2-digit sector.[9] See Appendix A.2 for details.

Other data. We use data on aggregate trade flows between US states and foreign countries from the 2012 Commodity Flow Survey. In addition, we construct bilateral distances and latitude differences between US counties and foreign countries, as well as various characteristics for countries, counties, and sectors. See Appendix A.3 for details.

Summary statistics. Panel A of Table 1 gives summary statistics on our sample of 3,141 $\times$ 195 origin-destination pairs. Column 1 shows means and standard deviations for all observations. Columns 3-4 show the same statistics for the subsamples of origin-destination pairs containing only observations with non-zero ancestry, and ancestry in the bottom and top quintile, respectively. The table shows that a lot of the variation both in ancestry and FDI is at the extensive margin. Only 1.8% of origin-destination pairs have an FDI link. Conditional on the US county

[7] Although Bureau van Dijk cross checks the data on international subsidiaries and branches using both US and foreign data sources, we cannot exclude the possibility that coverage may be better for some countries than for others. However, all of our specifications control for country fixed effects such that any such variation in coverage at the country level would not affect our results.

[8] When information on the number of employees is missing (which is the case for 6% and 25% of subsidiaries in the destination and origin, respectively), we assume the subsidiary employs one person.

[9] Appendix Table 1 provides a list of sectors and sector groups.

having any population with origins in the foreign country, 3.1% have an FDI link. The larger this population, the larger the probability of finding an FDI link, with 12.8% of the origin-destination pairs in the top quintile having an FDI link. Similarly, about half of the origin-destination pairs have ancestry of zero, reflecting the fact that most destinations in the United States do not have populations with ancestry from all 195 countries. The mean number of individuals with ancestry from a given origin is 318, but is highly skewed, with a mean in the top quintile of 2.8 million individuals. Compared to this stock of ancestry, the flow of immigrants between 1990 and 2000 is relatively small, with 25 on average across the sample. The summary statistics also show that the number of first-generation immigrants measured in the 2010 American Communities Survey (Total Immigrants 2010) is grossly understated (29 on average). This fact is known in the literature and appears to affect only the measurement if immigration flows but not the stock of ancestry (Jensen et al., 2015). For this reason, we exclude the 2000-2010 wave of migrations from our standard specification and instead rely on the pre-2000 census numbers.

Panels B and C show summary statistics following the same format for destination counties and origin countries used in our estimation of heterogenous effects. Appendix Table 3 gives summary statistics on the intensive margin of FDI.

2 Historical background

The 1880 US census counted 50 million residents, 10 million of which were first- or second-generation immigrants from 195 countries. The censuses taken since 1880 counted an additional 60 million immigrants. Our sample period thus covers the vast majority of migration to the United States.[10]

During the first part of this period, up until World War I, migration to the United States was largely unregulated. European migrants in particular faced few or no restrictions to entry and came in large numbers. Figure 1 shows the extent and the changing composition of migration over time. Although the peak of British migration was passed before the beginning of our sample, the numbers for 1880 clearly show the effect of the potato famines and the subsequently large inflow of Irish migrants. The second big wave of migration in our sample is that of Germans in the aftermath of the failed revolutions of 1848 and the consolidation of the German empire under Prussian control in 1871. Similarly disrupted by political changes and an economic crisis in the South, Italian migrants began flocking to the United States in large numbers around 1910,

[10] Daniels (2002) and Thernstrom (1980) provide an excellent introduction to the history of migration to the United States. Also see Goldin (1994) for the political economy of United States immigration policy.

followed by a peak in migrations from Eastern Europe, and in particular from Russia in the years after the October Revolution. The inflow of migrants overall dropped dramatically during World War I, falling below 4 million during the period between 1910 and 1920.

Although economic and political factors in the origin countries dominated the timing of these earlier European migrations, US immigration policies became relatively more important during the 1920s. The first important step toward regulating the inflow of migrants was the Chinese exclusion act of 1882 that ended the migration of laborers, first from China, and then in following incarnations from almost all of Asia. These restrictions were followed by literacy and various other requirements that came into effect after 1917, culminating in the establishment of a quota system in 1921. The quota system limited the overall number of immigrants, reduced the flow of migrants from Southern and Eastern Europe, and effectively shut out Africans, Asians, and Arabs. Combined with the effects of the Great Depression, these new regulations led to negative net migration in the early 1930s and then a stabilization at relatively low levels of immigration. The quota system was abolished in 1965 in favor of a system based on skills and family relationships, leading both to a large increase in the total number of migrants and a shift in composition toward migrants from Asia and the Americas, in particular from Mexico.

Figure 2 maps the spatial settlement pattern of newly arrived immigrants in the United States over time. For each census from 1880 to 2010, we compute the total number of new migrants to destination d, I_d^t,[11] projected on destination and year fixed effects to account for general immigration time trends and persistent destination-specific effects. We show only the residuals from this projection, color coded by decile. Migrants initially settled on the East Coast of the United States (in the mid-19th century), and then the frontier for migrants moved to the Midwest (in the late-19th century), to the West (in the 1930s), and to the South (in the 1980s). Starting in the 1970s, we can also see graphically the increased settlement of migrants in urban centers, with a series of dark dots appearing around large urban areas.

Below we use the interaction of this time-series variation in the relative attractiveness of different destinations within the United States with the staggered arrival of migrants from different origins as the basis of our identification strategy.

[11] Note we treat our first 1880 census differently: because we have no previous census with which to compare it, we define the number of migrants for 1880 as all residents in 1880 who are either foreign born or whose parents are foreign born.

3 Ancestry and Foreign Direct Investment

To evaluate the effect of the presence of descendants of migrants from a given origin on the probability that at least one firm within a given destination has an FDI link (incoming or outgoing) with a firm based in the origin country, we estimate the structural equation,

$$\mathbf{1}\left[FDI_{o,d} > 0\right] = \delta_o + \delta_d + \beta A_{o,d}^{2010} + X'_{o,d}\gamma + \varepsilon_{o,d}, \tag{1}$$

where $\mathbf{1}\left[FDI_{o,d} > 0\right]$ is a dummy variable equal to 1 if any firm headquartered in destination d is either the parent or the subsidiary of any firm headquartered in origin o. $A_{o,d}$ is a measure of common ancestry, usually calculated as the log of 1 plus the number of residents in d that report having ancestors in origin o, measured in thousands. (We choose this functional form in anticipation of non-parametric results discussed below, but also show robustness to a wide range of alternative specifications.) $X'_{o,d}$ is a vector of control variables that always includes the geographic distance between o and d, and the difference in latitude between o and d. δ_o and δ_d represent a full set of origin and destination fixed effects, augmented in most of our specifications by fixed effects for the interaction between destination and continent of origin, and between origin and destination census region.[12] The coefficient of interest is β, which measures the effect of ancestry on the probability that an FDI relationship exists between firms in o and d. The error term $\varepsilon_{o,d}$ captures all omitted influences, including any deviations from linearity.[13] Throughout the main text, we report heteroskedasticity-robust standard errors clustered at the origin-country level. In the appendix, we report standard errors calculated using alternative methods for all the main results of the paper, and show our results are robust.

Equation (1) takes the form of a gravity equation, widely used in the empirical literature, describing the pattern of international trade and FDI. We maintain the same form for consistency with this literature. Moreover, the gravity form is appealing on theoretical grounds because it can be derived in a variety of models.[14] The destination and origin fixed effects absorb all differences in productivity, market size, and market access between origins and destinations that systematically affect prices. We may thus interpret the coefficient β as the effect of ancestry controlling for the large set of conventional economic forces shaping international exchanges.

Equation (1) will consistently estimate the parameter of interest if $Cov\left(A_{o,d}^{2010}, \varepsilon_{o,d}\right) = 0$. This

[12] A census region is a grouping of adjacent US states listed in Appendix Table 2.

[13] We use a simple linear probability model, which allows for a straight-forward interpretation of the coefficient. In robustness checks, we also report results from a probit estimator.

[14] See Arkolakis et al. (2012) for a derivation of the gravity structure of international trade in a variety of theoretical settings. See Carr et al. (2001), Razin et al. (2003), Head and Ries (2008), and Ramondo (2014) for an application of the gravity structure to foreign direct investment.

condition is unlikely to hold in our data. First, past migration flows might be the result of economic transactions such as FDI flows or trade, not their driver. An example of this is the strong concentration of Japanese in Scott County, Kentucky, which emerged after Toyota set up a large manufacturing facility in Georgetown in the 1980s. They were primarily sent to Scott County by their employer. Second, economic transactions and migration flows might be both driven by omitted factors such as similarity in geographic or economic conditions that simultaneously affect $A_{o,d}$ and $\mathbf{1}\left[FDI_{o,d} > 0\right]$. For example, a common narrative is that Scandinavians preferred to settle in Minnesota over other destinations in the United States because of the similarity in climatic conditions to their origin countries. But the same factors might also drive the productive structure of Scandinavia and Minnesota, and affect the probability of economic transactions. Any such challenge would induce a bias in the OLS estimate of β in an indeterminate direction. These challenges are not unique to our data, but are likely concerns for inference from any data where ethnic linkages and economic transactions are simultaneous observed.

To address these concerns, we devise an IV strategy built on a simple dynamic model of migration. The stock of residents of destination d of ancestry from origin o at time t, $A^t_{o,d}$, depends on the past stock of residents with ancestry from o and the newly arrived migrants from o who settle in d. We assume the combination of three forces determines the number of new migrants from o who settle in d: a country-specific *push factor* drives migrants out of country o into the United States; a *pull factor* attracts migrants entering the United States to county d, irrespective of their origin; and a *social network factor* corresponds to the tendency of newly arrived migrants to settle in communities where people with the same ancestry already live.

Formally, the stock of residents in d with ancestry from o at time t follows:

$$A^t_{o,d} = a_t + a_{o,t} + a_{d,t} + b_t A^{t-1}_{o,d} + c_t I^t_o \left(\frac{I^t_d}{I^t} + d_t A^{t-1}_{o,d} \right). \tag{2}$$

The constant terms a_t, $a_{o,t}$, and $a_{d,t}$ control for demographic forces, which may vary over time, over space, and between different ethnic groups. The term $b_t A^{t-1}_{o,d}$ corresponds to the fact that ancestry is a stock variable that evolves cumulatively as new waves of migrants arrive. The constant b_t controls for demographics and for how ties to one's ancestry are passed from one generation to the next. The term I^t_o, the total number of migrants from country o entering the United States, measures the strength of the push factor, the fact that migrants are driven out of country o. The term I^t_d/I^t is the fraction of all migrants entering the United States who settle in county d from all origins. It measures the strength of the economic pull factor, the fact that county d is particularly appealing to migrants at time t. Finally, the term $A^{t-1}_{o.d}$ measures the

strength of the social network factor, the propensity of migrants to settle near their countrymen. The coefficient d_t controls the relative importance of the pull and social network factors.

Equation (2) is recursive, both because ancestry is passed down from generation to generation (the first $A_{o.d}^{t-1}$ term) and because newly arrived migrants' decision of where to settle depends on where past migrants have settled (the second $A_{o.d}^{t-1}$ term). For simplicity, we assume the initial condition $A_{o,d}^{-1880} = 0, \forall\,(o,d)$. Because we are treating our first census (1880) differently than the other censuses, recording not only foreign borns, but also children of foreign borns, this approximation seems to be appropriate for the United States, where the population rapidly increased in the 19th century. Solving equation (2) recursively, we get,

$$A_{o,d}^{2010} = \sum_{t=1880}^{2010} \left(a_t + a_{o,t} + a_{d,t} + c_t I_o^t \frac{I_d^t}{I^t} \right) \prod_{s=t+1}^{2010} (b_s + c_s d_s I_o^s). \tag{3}$$

The reduced-form equation (3) highlights how ancestry can be understood as the result of a sequence of migration waves and their subsequent cumulative effect. In each period t, the interaction of the contemporaneous push factor (I_o^t) and economic pull factor (I_d^t/I^t) determines the flow of migration from o to d. Demographic factors (the b_s's) and the social network factor (the $c_s d_s$'s) then amplify these initial waves of migrants. This simple specification is flexible, allowing for cases in which no migrants from a given origin country exist at some initial period of time. In the absence of a social network factor, $d_t = 0$, ancestry is simply given by the sum of the interaction of the push and pull factors over all time periods.

Crucially, this specification suggests plausibly exogenous variation in $I_o^t \frac{I_d^t}{I^t}$ would allow the construction of an instrument for $A_{o,d}^{2010}$. By interacting a push factor, I_o^t, which is not specific to destination d, but common to all destinations in the United States, and an economic pull factor, I_d^t/I^t, which is not specific to country o but to migrants from all countries, we rule out most plausible sources of endogeneity. However, our exclusion restriction could still be violated if at some point in time, migrants from o to d represent a large fraction of all migrants from o ($I_{o,d}^t$ a large fraction of I_o^t), or a large fraction of all migrants to d ($I_{o,d}^t$ a large fraction of I_d^t), or if migrants from other origins with unobserved similarities to o represent a large fraction of all migrants. To address these concerns, we exclude from the push factor migrants from o going to all destinations in d's census region (a group of four to seven adjacent states[15]), and from the economic pull factor, migrants from all origins in the same continent as o. We replace I_o^t by $I_{o,-r(d)}^t$, the migrants from o who settle in destinations *not* in the same census region as d; and I_d^t/I^t by $I_{-c(o),d}^t/I_{-c(o)}^t$, the fraction of migrants *not* coming from origins in the same continent

[15] Appendix Table 2 displays the allocation of states to census regions.

as o who settle in county d. $-r(d)$ stands for all destinations outside of d's census region, and $-c(o)$ stands for all origins outside of o's continent.

Using the functional form derived in (3), and replacing the $I^t_o \frac{I^t_d}{I^t}$ terms by $I^t_{o,-r(d)} \frac{I^t_{-c(o),d}}{I^t_{-c(o)}}$, our first-stage specification is thus

$$A^{2010}_{o,d} = \delta^f_o + \delta^f_d + \sum_{t=1880}^{2000} \alpha^f_t I^t_{o,-r(d)} \frac{I^t_{-c(o),d}}{I^t_{-c(o)}} + \sum_{n=1}^{5} \delta_n PC_n + X'_{o,d}\gamma^f + \eta_{o,d}, \tag{4}$$

where $\sum_{n=1}^{5} \delta_n PC_n$ stands for the first five principal components summarizing the information contained in the the series $I^s_{o,-r(d)} \cdots I^t_{o,-r(d)} \frac{I^t_{-c(o),d}}{I^t_{-c(o)}}, \forall t < s \leq 2010$. We prefer summarizing the higher-order interactions in (3) as principle components because it avoids adding an excessive number of highly co-linear regressors. Our results are robust to including these terms or not.

Our key identifying assumption is

$$Cov\left(I^t_{o,-r(d)} \frac{I^t_{-c(o),d}}{I^t_{-c(o)}}, \varepsilon_{o,d} | controls \right) = 0. \tag{5}$$

Simply put, it requires that any confounding factors that may make a given destination more attractive for both migration and FDI from a given origin country do not affect the interaction of the settlement of the average migrant originating from another continent with the total number of migrants arriving from the same origin but settling in a different census region.

To further relax this assumption, most of our specifications also control for interactions of fixed effects that are symmetric to the construction of our instruments: the interaction between destination and continent-of-origin fixed effects ($\delta^f_d \times \delta^f_{c(o)}$) and the interaction between origin and destination-census-region fixed effects ($\delta^f_o \times \delta^f_{r(d)}$). For these specifications, the identifying assumption is then merely that any confounding factors do not systematically affect both sides of the interaction across origin-destination pairs. We further corroborate this assumption below using a series of falsification exercises and placebo treatments.

3.1 The First-Stage Relationship

Table 2 shows our basic first-stage regressions, estimates of equation (4). Column 1 is the most parsimonious specification regressing our measure of ancestry on origin and destination fixed effects and the nine simple interaction terms $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_t$. To facilitate the interpretation of the results, we sequentially orthogonalize each of the terms with respect to the interaction terms from the previous censuses. For example, the coefficient marked $I^{1900}_{o,-r(d)}(I^{1900}_{-c(o),d}/I^{1900}_{-c(o)})$

shows the effect of the residual obtained from a regression of $I^{1900}_{o,-r(d)}(I^{1900}_{-c(o),d}/I^{1900}_{-c(o)})$ on the same interaction in 1880, the coefficient marked 1910 shows the effect of the residual from a regression of the 1910 interaction on the interactions from the previous two censuses, and so on. Although this procedure has no effect on the fit and predictive power of the first stage as a whole, we find it useful because it allows us to interpret each coefficient as the marginal effect of the innovation in the migration pattern of the period reported with respect to the previous periods.

All nine coefficients shown in column 1 are positive, and seven are statistically significant at the 5% level. Figure 5 depicts the coefficients graphically. The first main insight from this figure is that even our earliest (pre-1880) snapshot of the cross-sectional variation in economic attractiveness to new migrants has left its imprint on the present-day ancestry composition of US counties: the destinations that were relatively more attractive to the typical migrant arriving pre 1880 continue to the present day to house significantly larger numbers of residents of the ethnic groups that arrived in large numbers prior to 1880. The overall pattern of coefficients suggests a hump-shape, where very recent waves of migrants have a smaller impact on current ancestry than migrations a few decades back, but the effect of past migrations eventually fades after about one century. The pattern is consistent with the effects of two opposing forces: the effect of migrants arriving earlier is amplified over time because migrants tend to pass their ancestry on to more than one descendant and these areas attract further migrants from the same origin to settle in the same destination. However, migrations reaching further back into history are more likely to be forgotten or disseminate through US internal migration. An exception to the general pattern is the coefficient for 1920-30, which is smaller and insignificant. A likely explanation is the Great Depression, which induced large reverse migrations from the United States of recently arrived migrants, demonstrating our model is less well suited for periods with negative net migration.

Taken together, the nine simple interactions incrementally increase the R^2 of the regression by 8 percentage points and explain about 16% of the variation in ancestry not explained by origin and destination fixed effects. Column 2 adds controls for distance and latitude difference. Both columns 1 and 2 estimate equation (2) under the restriction that $d_s = 0$. Column 3 relaxes this restriction and adds the first five principal components of the higher-order interaction terms, raising the R^2 by another three percentage points. Columns 4 and 5 add destination per continent-of-origin fixed effects and origin per destination-census-region fixed effects.

Column 5 is our standard specification. The F-statistic against the null that the excluded instruments are irrelevant in this specification is 91.5.[16] Column 6 includes third-order polynomials

[16]The Hansen J-test statistic is 18.975 with a p-value of 0.124. We thus fail to reject the null that our instruments are uncorrelated with the error term and correctly excluded from the second-stage regression.

in the distance and latitude difference between origin o and destination d. Columns 7 through 9 successively show variations of our instrumentation strategy: column 7 includes migration data from the 2005-2010 ACS survey, column 8 drops migration prior to 1880, and column 9 estimates our standard specification in levels rather than logs. Throughout all of these variations, we can comfortably reject the null that our instruments are jointly irrelevant in the first stage.

We illustrate our first-stage identification with a stylized graphical illustration, using two specific examples: that of migrations from Ireland, with a migration peak in our first 1880 census (corresponding to the massive migrations from Ireland to the United States in the 1850s), and that of migrations from Germany, with a migration peak in the 1880-1900 period. The top panel of Figure 3 shows the relative attractiveness of US destinations for pre-1880 migrants, where we exclude migrations from Europe – analogously to our regression specification. At the time when Irish migrants arrived en masse to the United States most non-European migrants settled in the East. We expect most Irish migrants from this initial wave to have settled in the East. The bottom panel of Figure 3 shows the distribution of US residents with Irish ancestry in 2010, with disproportionately many in the East. The top panel of Figure 4 shows the relative attractiveness of US destinations for non-European migrants during the 1880-1900 period, that is, the peak of German migrations to the United States. At that time, the preferred destination for migrants had shifted to the Midwest and the West. We expect many German migrants from this first wave to have settled in the Midwest and the West. The bottom panel of Figure 4 shows the distribution of German descendants in 2010, with large populations in the Midwest. Note that for both countries, those initial waves do not predict large populations of Irish and German descendants in Florida and Southern California, two areas with both large Irish and German ancestry, and strong FDI ties to Ireland and Germany. Those settlements may be attributable to variation coming from later migration waves, to some common unobserved factors (precisely what we want to avoid with our IV approach), or to forces outside our simple model.

3.2 Instrumental Variables Results

In our IV estimation, we explicitly test the hypothesis that an increase in the number of descendants from a given origin increases the probability that at least one local firm engages in FDI with that country. In column 1 of Table 3, we estimate equation (1) while instrumenting (the log of) ancestry in 2010 with the simple interaction terms $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_t$ and controlling for origin and destination fixed effects, distance, and the difference in latitude between destination and origin. The coefficient estimate on ancestry is 0.243 (s.e.=0.024) and statistically

significant at the 1% level. The coefficient on distance is not statistically distinguishable from zero, perhaps reflecting the fact that US counties do not differ much in their distance to most foreign countries, and that these smaller differences are irrelevant once we control for the effect of the distance between the United States as a whole and the country in question (it is absorbed in the country fixed effect). By contrast, the difference in latitude is positive and significant, showing that, all else being equal, firms tend to engage in FDI with origin countries that are climatically different from their own location.

In column 2, we add the five principal components of the higher-order interactions to our set of instruments, resulting in a slight fall in the coefficient of interest to 0.190 (s.e.=0.024). The estimate implies that doubling the number of residents with ancestry from a given origin relative to the sample mean (from 318 to 636) increases by 4.1 percentage points the probability that at least one firm engages in FDI with that origin.

Another useful way to gauge the relative importance of ancestry is its partial R^2 relative to the controls included in the specification. Taken together, the standard gravity terms, that is, the origin and destination fixed effects as well as distance and latitude difference, explain 20.3% of the variation in FDI Dummy. Adding ancestry to these variables in a simple OLS specification (shown in panel B) accounts for an additional 9 percentage points, though this effect is not necessarily causal. Adding our nine simple interactions to the standard gravity terms, thus running the most parsimonious reduced form, raises the R^2 by 1.5 percentage points, and adding them in combination with the five principal components raises the R^2 by 1.8 percentage points. These numbers are a lower bound on the importance of common ancestry for FDI, because it only accounts for part of the causal effect.

Column 3 shows our standard specification. It includes destination per continent-of-origin fixed effects and origin per destination-census-region fixed effects. For a given origin country, this demanding specification uses only variation across different destinations within the same census region while controlling for the fact that each destination may have an idiosyncratically high or low propensity to interact with the continent containing the origin country. Symmetrically, for each destination, it uses only variation across different origins within the same continent while controlling for the fact that each origin may have an idiosyncratically high or low propensity to interact with the census region containing the destination county. Reassuringly, the coefficient estimate is almost identical to that in column 2 (0.197, s.e.=0.030). Comparing this estimate with the same column in panel B shows that it is about 25% larger than the corresponding OLS coefficient. The endogenous assignment of migrants to destinations within the United States thus appears to induce a downward bias in the OLS coefficient.

The remaining columns probe the robustness of this result. The coefficient estimate remains remarkably stable and highly statistically significant across specifications. Column 4 adds a third-degree polynomial in distance and latitude difference to capture a potentially non-linear effect of distance; column 5 adds an interaction term for the contemporaneous 2010 migrations; and column 6 adds a more stringent set of origin-state fixed effects, exploiting only variation within US states. All of these variations leave our coefficient of interest virtually unchanged.

Figure 6 presents our results graphically. It plots the coefficients from a reduced-form regression of the FDI Dummy on the simple simple interaction terms $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_t$ (again orthogonalized as in Table 2) and destination and origin fixed effects. All nine coefficients are greater than zero, and seven of them are statistically significant at the 5% level. Destinations that received an (exogenous) increase in the number of migrants from a given origin in any of the nine consecutive waves of immigration thus tend to have a significantly higher probability of engaging in bilateral FDI with these origin countries today.

3.3 Validity of the Exclusion Restriction

To corroborate the validity of our exclusion restriction, we run a series of robustness checks.

Placebo Test. Our first robustness check uses a placebo treatment to assess whether our instrument really picks up push factors that are specific only to one country, or correlated with omitted variables that affect FDI with other countries in a systematic fashion.

The results are presented in Appendix Table 4. In panel A, we assign the interaction between push and pull factors for a given origin to a quasi-randomly selected other country: its nearest neighbor in alphabetic order. To further check whether the same push factors might affect two countries in different continents, panel B assigns the interaction between push and pull factors for a given country to its nearest neighbor in alphabetic order in a *different* continent. Across all specifications, our placebo treatment is always statistically insignificant, and the point estimates are near zero. We conclude from this placebo test that our instrument is not picking up any artificial correlation (positive or negative) between the push factors for different countries.

Standard Errors. Appendix Table 5 shows our standard specification from column 3 of Table 3 using alternative standard errors clustered by origin, destination, state, state-county, and state-continent, as well as various block-bootstrapped standard errors. Across all these specifications, clustering by origin, as we do throughout the paper, is the most conservative choice.

An alternative way to detect any tendency to over-reject the null is to reassign the "treatment" to a different set of outcome observations, in the spirit of Fisher's randomization inference procedure. We repeatedly assign the interaction between push and pull factors for country o to randomly selected other countries. Appendix Figure 1 shows the distribution (histogram) of t-statistics on the estimated coefficient on ancestry across 200 random assignments. The t-stats across those random assignments are centered around zero, with no noticeable tendency for positive or negative estimates. Reassuringly, the rate of false positives is 2.5%.

The Communist Natural Experiment. Our second robustness check combines our identification strategy with a more explicit natural experiment, making use of specific historical episodes of economic isolation between the United States and former communist countries, during which FDI was impossible, and not expected to be possible in the near future. These countries are the Soviet Union from 1918 to 1990, China from 1945 to 1980, Vietnam from 1975 to 1996, and Eastern Europe (the non-Soviet members of the Warsaw pact) from 1945 to 1989.

Table 4 shows estimates of (1) for each of these countries or sets of countries, using as instruments only migration waves that occurred during the period of exclusion. We can confidently assume the prospect of FDI, which was outlawed for political reasons, did not drive migrations during those periods. For all countries, we find a large causal impact of ancestry on FDI. The magnitude of the estimated impact of ancestry is broadly similar to the one we estimated for all countries in Table 3, and the estimated coefficients are statistically significant at the 1% and 5% level for the Soviet Union and Eastern Europe, respectively. The coefficients are not statistically significant for China and Vietnam, most likely because most migration from these countries occurred before or after our period of exclusion.

The fact that we find similar results in these more restrictive natural experiments as in our baseline specification in Table 3 bolsters our confidence that our exclusion restriction is valid, and that neither reverse causality nor omitted variables drive our baseline results

Ancestry and immigration. According to our reduced-form model of migration, the number of migrants arriving at a given destination is a function of the economic attractiveness of the destination at the time (measured by the interaction of our pull and push factors) and the stock of descendants of migrants from the same origin (the social network factor). To provide direct evidence for this model, we estimate the specification

$$I^t_{o,d} = \delta_o + \delta_d + \alpha I^t_{o,-r(d)} \frac{I^t_{-c(o),d}}{I^t_{-c(o)}} + \beta A^{t-1}_{o,d} + X'_{o,d}\gamma + \varepsilon_{o,d} \tag{6}$$

for $t = 2000, 1990$ (the census years for which we have information on lagged ancestry), where we again instrument for $A_{o,d}^{t-1}$ using (4).

Table 5 shows that both the coefficient on the interaction and the coefficient on lagged ancestry are indeed positive and significant predictors of current migration, although the results for migration between 1980 and 1990 are significant only at the 10% level. The coefficient on the interaction is in the ballpark of 1 in both specifications. This finding is what we would expect if newly arrived migrants distributed equally at each point in time, independent of their origin.

Overview of additional robustness checks. Finally, we run a battery of additional robustness checks. The results are presented in the appendix for concision.

Appendix Table 6 shows our results separately for the five largest origins (by number of descendants in the United States), and the five largest destinations (in total number of foreign ancestry).[17] The estimated impact of ancestry on FDI is similar across these specifications.

Appendix Table 9 shows plausible variations of our leave-out instrument, removing or not different sets of migrants from the $I_{o,-r(d)}^{t}(I_{-c(o),d}^{t}/I_{-c(o)}^{t})$ interaction terms. In panel A, we do not remove any migrants from o to d when computing our push and pull factors, using directly the $I_{o}^{t}(I_{d}^{t}/I^{t})$ interaction terms. In panel B, we only remove migrants from o in the pull factor and migrants to d in the push factor, that is, the interactions $I_{o,-d}^{t}(I_{-o,d}^{t}/I_{-o}^{t})$. In panel C, we additionally remove migrants from the same continent from the pull factor, $I_{o,-d}^{t}(I_{-c(o),d}^{t}/I_{-c(o)}^{t})$. In panel D, we go further than in our standard specification, additionally removing migrants from all adjacent states in calculating the pull factor. Across these variations, the coefficient of interest in our standard specification (column 3) is stable between 0.174 (s.e.=0.028) in panel B and 0.200 (s.e.=0.024) in panel D, where as expected, less aggressive leave-out instruments produce estimates that are marginally closer to the OLS coefficient (0.155, s.e.=0.018).

Finally, in Appendix Tables 10 and 11, we experiment with a non-parametric specification as well as various alternative functional forms for our measure of ancestry. We discuss these results in more detail in section 5 because they shed light on the mechanism through which ancestry affects FDI. We only note here that our main result, the strong and significant causal impact of ancestry on FDI, is robust to various functional-form specifications.

[17] Appendix Table 7 shows the result from separate regressions for all countries. Appendix Table 8 shows the results from separate regressions for all sectors.

4 The Effect of Ancestry on Other Outcomes

4.1 Inward and outward FDI

We first distinguish between inward and outward FDI. To do so, we estimate our standard specification from column 3 of Table 3 separately for inward FDI, where the outcome variable is a dummy equal to 1 if at least one firm in US county d is a foreign affiliate of a parent in foreign country o, and for outward FDI, where the outcome variable is a dummy equal to 1 if at least one firm in US county d is the parent of a foreign subsidiary in country o. The results are in Figure 7, showing the two coefficient estimates on ancestry and their 95% confidence intervals. Both coefficients are positive and statistically significant. We find a stronger impact of ancestry on outward FDI, an elasticity of 0.2, than on inward FDI, an elasticity of 0.15, although both coefficients are not statistically distinguishable from each other.

4.2 The intensive margin of FDI

So far, we have studied the impact of ancestry on the extensive margin of FDI, the probability that at least one firm engages in FDI. We now turn to the impact of ancestry on the intensive margin of FDI: conditional on being positive, how large are FDI flows for a given size of the local population with a given foreign ancestry?

In Table 6, we estimate by IV various specifications of the form,

$$\ln FDI_{o,d} = \delta_o + \delta_d + \beta A_{o,d}^{2010} + X'_{o,d}\gamma + \varepsilon_{o,d}. \tag{7}$$

where $FDI_{o,d}$ corresponds to various measures of the volume of FDI between o and d and where we instrument $A_{o,d}^{2010}$ with the same first-stage equation (4) as earlier.[18]

Because of the log specification, cases of zero FDI will automatically be dropped from our sample. This creates a selection problem, as counties with non-zero FDI are likely to be systematically different from those with zero FDI. To correct for this potential selection bias, we implement a simple Heckman (1979) correction. Formally, we first estimate an IV probit regression for the extensive margin of FDI,

$$\rho_{o,d} = \Pr\left(FDI_{o,d} > 0 | observables\right) = \Phi\left(\delta_o + \delta_d + \beta A_{o,d}^{2010} + X'_{o,d}\gamma\right), \tag{8}$$

[18] Appendix Figure 2 presents our results graphically, showing a scatterplot with $\log\left(ancestry\right)$ on the horizontal axis, and $\log\left(FDI\ volume\right)$ on the vertical axis, for all county-country pairs in our sample. The figure shows a positive relationship between the size of the local community in county d with ancestry from country o and the number of FDI links between county d and country o.

where $A_{o,d}^{2010}$ is again instrumented as in equation (4). We extract an estimate for $\hat{z}_{o,d} = \Phi^{-1}(\hat{\rho}_{o,d})$, the predicted latent variable that determines non-zero FDI. We then include a simple inverse Mills ratio term $\left(\hat{\mu}_{o,d} = \frac{\varphi(\hat{z}_{o,d})}{\Phi(\hat{z}_{o,d})}\right)$ within our set $X_{o,d}$ of controls in the intensive margin equation (7), where φ denotes the density function. This correction for selection, the extensive margin of FDI, is similar to the procedure in Helpman et al. (2008) for international trade.[19]

We use various measures for the volume of FDI. In panel A of Table 6, we count the total number of FDI relationships, that is, the sum of the number of firms in d which are either parent or subsidiary of a firm in o and the number of firms in o which are parent or subsidiary of a firm in d. In panel B, we only count the number of firms in d which are a subsidiary of a firm in o, a measure of inward FDI. In panel C, we only count the number of firms in o which are parent of a firm in d, an alternative measure of inward FDI. In panel D, we measure the total local employment in county d at subsidiaries of firms in o, giving us a measure of the impact of inward FDI on local employment. Panels E, F and G use similar measures for outward FDI: the number of foreign subsidiaries of local firms in E, the number of local parents of foreign subsidiaries in F, and the number of foreign workers employed by subsidiaries of local firms in panel G.

Across most specifications, and for most measures of the intensive margin of FDI, we find a positive impact of ancestry on the volume of foreign investment. The effect of ancestry on the intensive margin of FDI, the coefficient β in equation (7), is around 0.3 across most specifications. This coefficient is large. A doubling of the number of residents in county d who report ancestry from country o (from the mean, 318, to 636) increases local employment at subsidiaries of foreign firms by 35%. Alternatively, increasing the number of residents in d with ancestry from o by one s.d. from the mean increases local employment at foreign subsidiaries by 7% of a s.d.

With all measures of the volume of FDI, the estimated impact of ancestry is larger in our IV specification (column 2) compared to the OLS specification (column 1). This finding is similar to our result for the extensive margin of FDI. For all measures of the volume of FDI, the impact of ancestry is larger when we include our complete set of interacted fixed effects (column 2) than when we use a simple gravity specification without interacted fixed effects (column 3). Finally, correcting for selection using a Heckman type procedure always leads to a lower estimated impact

[19]Note Helpman et al. (2008) correct for both the selected presence of zeros, as well as for the unobserved selection of which firm engages in foreign activities, export in their case. For several reasons, we only correct for the first term (the presence of zeros), not for the second (the selection of firms). There are several reasons for that. First, we are not interested in how ancestry affects the volume of FDI of one individual firm, as Helpman et al. (2008) are interested in how various covariates affect the export of one individual firm, but rather in how ancestry affects the *total* volume of FDI between a US county and a foreign country. Second, we directly use firm-level data, so that we do not require an explicit correction for firm selection. Finally, at the very fine level of geographic disaggregation we use, US counties as opposed to entire countries in Helpman et al. (2008) – the simple structural model they use to motivate their correction for firm selection is unlikely to be appropriate.

of ancestry on the volume of FDI (column 4).[20] In the case of our measures of inward FDI (panels E, F and G), the impact of ancestry becomes insignificant, with slightly negative point estimates. Except for those three cases, correcting for selection does not change our results substantially.

Figure 10 illustrates these results graphically by estimating equation (7) using data only for Germany (top panel) and Britain (bottom panel). It shows a conditional scatterplot of the number of German (British) subsidiaries of firms in each US county over the their predicted German (British) ancestry. Both panels show a positive and significant slope close to the corresponding full-sample estimate in column 3 of Table 6 and no obvious outliers. Figure 11 shows corresponding plots using only data for Los Angeles county and Cook county.

The conclusion from Table 6 is that foreign ancestry not only affects the extensive margin of FDI, but also has a sizable impact on the intensive margin of FDI. More descendants of foreign migrants increases the likelihood that local firms engage in FDI, the number of firms that do so, and the local employment by foreign-owned firms.

4.3 International trade

To conclude this section, we compare our methodology to the existing literature on the effects of common ancestry, which has primarily focused on international trade.

We do not have access to micro data on international trade by US firms that would allow us to measure trade flows at the level of individual US counties. Instead, we use readily available trade data from the Commodity Flow Survey at the level of US states. Our instrument for the composition of ancestry is the same as in equation (4) for FDI, except that all variables are defined at the state level, and not at the county level.

To compare the magnitude of the effect of ancestry on international trade, we focus our attention on the intensive margin of trade, rather than the extensive margin of trade, as most of the literature has done. Table 7 presents the results of the estimation of various specifications of equation (7), where the dependent variable is now total exports from US state d to foreign country o, or total imports by US state d from foreign country o. For the intensive margin of FDI, we again allow for a Heckman-type correction for the selection bias due to zero trade. Panel A shows the results for the intensive margin of FDI aggregated to the state level for comparison. Panels B and C show the corresponding results for exports and imports. The main finding

[20] Note the number origin and destination fixed effects is too large for a probit estimation of the extensive margin of FDI to be computationally feasible. Moreover, Greene et al. (2002) suggest using Monte Carlo simulations that probit regressions tend to give biased estimates in the presence of a large number of fixed effects. For both reasons, in column 4, we only implement a Heckman-type correction for selection in a specification without origin and destination fixed effects.

emerging from the table is that when we properly instrument for ancestry, and include a full set of origin and US state fixed effects, we continue to find a robustly positive and significant effect of ancestry on FDI, but we do not find a robust positive impact of ancestry on the patterns of international trade of US states. This result can be seen in columns 3 and 4 of Table 7, where the effect of ancestry on trade is either insignificant or negative.

This finding is in stark contrast to earlier findings in the literature, started by the seminal contribution of Gould (1994) and Rauch and Trindade (2002), and the recent IV results of Cohen et al. (2015) for trade with Japan, and Parsons and Vezina (2014) for trade with Vietnam, that all find the presence of migrants facilitates both exports and imports.[21] We do not find any such positive impact of ethnic ties (ancestry) on international trade. A closer look at the data suggests two important features are essential in reaching this negative conclusion: When either a formal identification is missing (OLS in column 1), or no control for destination (US state) fixed effect is included (column 2), we erroneously find a positive and significant estimated impact of ancestry on trade. But when both are present (columns 3 and 4), we find none.

4.4 Quantitative implications

Having estimated the impact of ancestry on a range of different outcomes, we illustrate the quantitative implications of our findings using two thought experiments. First, we estimate how investment relations between US counties and China might have evolved if Chinese migrants had not been effectively barred from entering the United States between 1882 and 1965. Second, we report how FDI relationships between Los Angeles and the world might have evolved if Los Angeles had had an influx of population in the 1800s similar to that resulting from the San Francisco Gold Rush.

The effect of Chinese exclusion. The US government passed the Chinese Exclusion Act into law in 1882 in response to increased immigration from China. It essentially closed the United

[21] Rauch and Trindade (2002) find the larger the ethnic Chinese communities in two countries, the more these countries trade. Combes et al. (2005) find the larger migration between regions within France are, the more trade between these regions. Using a data set similar to ours, Gould (1994) finds migration between foreign countries and US states are correlated with international trade at the state level, a finding confirmed by Dunlevy (2006). Felbermayr and Toubal (2012) use data on the composition of the stock of migrants and bilateral trade flows among OECD countries and find a strong correlation between migrations and trade. Two papers use an IV approach: Cohen et al. (2015) use Japanese internment camps during WWII as an instrument for the current distribution of Japanese migrants in the United States, and find a strong impact of the presence of Japanese migrants on trade and FDI to Japan; Parsons and Vezina (2014) use the quasi-random allocation of Vietnamese boat people in the United States as an instrument for the current distribution of Vietnamese migrants in the United States, and find a strong impact on trade with Vietnam.

States to legal immigration of laborers from China. It was in force until 1943, when it was replaced by the Magnuson Act, which allocated a quota of 105 immigrants per year from China, and was in effect until 1965, when the removal of the quota system allowed for large-scale Chinese immigration for the first time. We refer to the entire period from 1882 through to 1965 as the "Chinese Exclusion." How different would the ancestry composition and FDI of US counties be today had it not been for Chinese Exclusion?

We first derive a prediction for the time path of aggregate Chinese migration to the United States in this counterfactual scenario. We aggregate our immigration data at the time-census region-origin level to run a regression of the form

$$I^t_{o,r} = \delta^t_r + \delta_o + \xi \cdot D^t_{China} + \nu^t_{o,r},$$

where D^t_{China} is a dummy equal to 1 for Chinese migration between 1880 and 1970 , and $\delta_{t,r}$ and δ_o are time $\times$ census region and origin fixed effects, respectively. The coefficient of interest, ξ, estimates the average impact of the Chinese Exclusion Act across the years of its existence and census regions on immigration from China. Defining the counterfactual time path of immigration as $\tilde{I}^t_{o,r} \equiv I^t_{o,r} - \xi \cdot D^t_{China}$, we then predict the change in ancestry using the estimates from our standard first-stage regression as

$$dA_{o,d} \equiv \sum_t \hat{\alpha}^f_t \cdot \left(\tilde{I}^t_{o,-r(d)} - I^t_{o,-r(d)} \right) \frac{I^t_{-c(o),d}}{I^t_{-c(o)}},$$

where $\hat{\alpha}^f_t$ are the estimated first-stage coefficients from our standard specification in column 5 of Table 2. The hypothetical incidence of FDI relations with China at the county level is $dY_{o,d} \equiv \hat{\beta} \cdot dA_{o,d}$, where $\hat{\beta}$ is the estimated second-stage coefficient from column 3 in Table 3.

Our estimates suggests that in the absence of the Chinese Exclusion Act, the number of individuals with Chinese ancestry in 2010 would be 1.3 million higher. This increase would have been highly unequally distributed: some states would have virtually no additional inhabitants with Chinese ancestry, whereas the average county in California would have 15 thousand additional individuals with Chinese ancestry. This change translates into large but heterogenous changes in the incidence of FDI relationships with China. Figure 8 depicts the expected change the probability of positive FDI with China. Free immigration from China would have resulted in substantially stronger FDI ties with the Northeast, the Midwest and the Southwest.

Los Angeles Gold Rush. To gauge the size of the estimated intensive margin effects, we derive predictions on the intensity of FDI relationships between Los Angeles county and the world under the hypothetical scenario that Los Angeles had experienced a Gold Rush similar to San Francisco. In particular, we derive predictions on the intensity of FDI relationships with the world if the number of immigrants pre-1880 had been fivefold the actual number of immigrants to Los Angeles. Table 8 presents the results of this thought experiment for the 10 foreign countries with the biggest predicted change in their ancestry group in Los Angeles in 2010. Column 1 presents the actual number of individuals (in thousands) of each ancestry in Los Angeles County in 2010. Column 2 presents the total number of FDI links recorded in our data between Los Angeles County and the respective origin countries. Columns 3 through 5 present the predictions of our though experiment. The calculations in column 3 are based on the IV specification corresponding to column 2 of Table 6, with the only difference that we do not include the principle components as instruments. A 'Gold Rush' in Los Angeles would have resulted in sizeable effects on the intensity of foreign direct investment relations with those countries that were the source of immigration pre-1880: The intensity of foreign direct investment between Los Angeles County on the one side and German and Ireland on the other side would have roughly doubled. Column 4 replicates the same exercise using the reduced form regression corresponding to column 3 as a basis for the counterfactual predictions. The results are moderately smaller and in the same ballpark as those derived from the IV regression. Finally, column 5 presents the predicted absolute change in the size of the ancestry groups, based on a reduced form regression analogous to column 4 with *Ancestry 2010* (in levels) as outcome variable. It suggests that the population of Irish and German descent living in Los Angeles County today would counting roughly 100.000 individuals more, respectively.

5 Understanding the Effect of Ancestry

A clear advantage of the fact that our instrumentation strategy can be flexibly applied to the entire set of origin countries is that it delivers the statistical power and a sufficiently large number of instruments to probe in some detail the nature of this effect.

5.1 First-generation immigrants

Having shown the historical stock of ancestry predicts subsequent migrations, we ask whether the effect of ancestry on FDI requires a sustained inflow of migrants from the same origin. To address this question, Table 9 compares the (causally identified) effect of ancestry to that of

foreign born, that is, first-generation immigrants. Column 1 replicates our standard specification for comparison. Column 2 replaces our measure of ancestry in equation (1) with the log of 1 plus the number of foreign born from a given origin alive in 2010 (measured in thousands, using the same functional form as for our measure of ancestry), instrumenting as in equation (4). As expected, we obtain a positive and statistically significant coefficient on foreign born (the correlation between the two variables is 0.59). However, when we simultaneously include both endogenous variables in the specification, the coefficient on ancestry remains positive and statistically significant at the 1% level, whereas the coefficient on foreign born in 2010 is close to zero and insignificant in the OLS specification in column 3 and turns negative in the IV specification in column 4. Because each foreign born also increases the number of individuals with foreign ancestry and the coefficient on foreign born is smaller than the one on ancestry in absolute terms, we may interpret this result as stating that foreign born have a positive effect on the probability of bilateral FDI, but their effect is smaller than that of their descendants.

Using the number of foreign born in 1970 as a proxy for second-generation immigrants, columns 5-7 shows that, by contrast, the effect of second-generation immigrants is not significantly different from that of the average descendant of migrants with a foreign ancestry. These results suggest first-generation immigrants have a significantly smaller effect on FDI than their descendants and that the effect of ancestry on FDI fully develops only over long periods of time, consistent with the temporal pattern of reduced-form coefficients shown in Figure 6.

5.2 Heterogenous Effects

We next explore how the effect of ancestry on FDI varies across origins, destinations, and sectors.

Heterogeneous effect across origins. We begin by dropping the destination fixed effects from equation (1) and running 112 separate IV regressions of the FDI dummy on ancestry for each origin country that has at least one FDI link with the United States. The top panel in Figure 9 plots the coefficients on ancestry over the reciprocal of the standard error, yielding a funnel plot, where the size of the circles is proportional to the share of each origin country in the total foreign ancestry of the United States. All coefficients to the northeast of the red line are statistically significant at the 5% level. Of the 112 coefficients, 74 are positive and statistically significant at this level level. For easier readability, the plot excludes coefficient estimates larger than 1 and with a reciprocal of the estimated standard error exceeding 150 (the full set of results is in Appendix Table 7). The plot shows the estimates for larger origin countries tend to be more precise and clustered in a tight band. The coefficients for the five largest origin countries range

from 0.174 (s.e.=0.011) for Mexico to 0.265 (s.e.=0.009) for the UK. The plot also suggests some heterogeneity may exist in the size of the effect.

In panel A of Table 10, we explore the heterogeneous effect of ancestry on the extensive margin of FDI across countries. We use the simple IV specification of column 2 in Table 3 and add interactions of ancestry with measures of distance and various country characteristics. Columns 1-3 show the interaction of ancestry with the geographic distance and measures of genetic, linguistic, and religious distance between the United States and the origin country as defined by Spolaore and Wacziarg (2015). The results show a consistently positive and statistically significant effect on the interaction between ancestry and geographic distance. Once we account for this interaction, the interactions with the three other measures of distance are statistically insignificant, suggesting the effect of ancestry on FDI increases with geographic distance but not other measures of cultural distance. The remaining columns also show a consistently positive effect on the interaction between ancestry and judicial quality as defined by Nunn (2007), suggesting ancestry has a relatively larger effect on FDI when the origin country has good institutions.

Panel B shows broadly similar results for the intensive margin of FDI. The only qualitative difference in the results is that we find a positive and significant effect of the interaction with ethnolinguistic fractionalization in the origin country (as provided by Alesina et al. (2003)) once we control for the interactions with geographic distance and judicial quality, suggesting ancestry from a given origin has a relatively larger effect at the intensive margin when the origin country is more ethnically diverse.

Heterogeneous effect across destinations. The bottom panel in Figure 9 shows coefficients and standard errors from 100 separate IV regressions of the FDI dummy on ancestry for the 100 largest counties within the United States, where again the size of the circles is proportional to the size of the total population of the destination. The coefficient is positive and statistically significant for 99 of these 100 regressions. The size of the coefficient varies from 0.093 (s.e.=0.045) for Pierce County in Washington to 0.457 (s.e.=0.060) for Orleans County in Louisiana.

Table 11 probes the heterogeneity of this effect in more detail by again reverting to our simple IV specification from column 2 in Table 3 and interacting our measure of ancestry with the share of the county's population that are of any foreign ancestry (column 1) and a measure of the ethnic diversity within the destination county (column 2). The table shows when either fewer residents with any foreign ancestry (a lower foreign share in column 1) or a more diverse set of ancestries (a higher ethnic diversity in column 2) are present, the effect of ancestry is relatively stronger. However, when both measures of the composition of ancestry are included, only the

interaction with ethnic diversity remains significant: ancestry has a stronger impact on FDI flows to and from more ethnically diverse US counties, but the overall share of residents with foreign ancestry matters little. These results suggest US counties with more diverse populations may act as hubs for FDI, where the presence of descendants from a wide variety of origins enhances the effect of ancestry on FDI in each bilateral link.

Heterogeneous effect across sectors. We reconstruct our data set separately for all 20 2-digit NAICS sectors, considering in each case only FDI links sent and received by US firms in that sector. Appendix Table 8 shows the coefficients estimated in separate IV regressions for each sector, where we find a statistically significant positive effect of ancestry on FDI in 18 of the 20 sectors. However, because the number of US firms in some of these sectors is less than 500, interpreting these results may be hard. Panel A of Table 12 instead aggregates sectors into five groups with a more comparable number of firms. It shows the effect is largest in manufacturing (0.175, s.e.=0.029) and smallest in sectors dealing in natural resources (0.036, s.e.=0.010). The overall pattern of results appears consistent with the view that the effect of ancestry on FDI may be larger in sectors that involve more differentiated inputs (Nunn (2007)), but we do not have detailed data on enough sectors to test this hypothesis formally.

Panel B of Table 12 presents the IV coefficient of ancestry on FDI separately for firms producing final consumption goods and firms producing intermediate inputs.[22] The impact of ancestry on FDI is somewhat larger for intermediates than for final goods. This finding suggests unobserved local tastes that may affect both migrations and the location of multinationals do not play a major role.

Panel C of Table 12 presents the IV coefficients of ancestry on FDI separately for the subset of large versus small firms. We find the impact of ancestry on FDI is positive for both categories of firms, but it is substantially larger, about twice as large, for large than small firms. This suggests larger firms are better able to take advantage of the local ethnic diversity.

5.3 The effect of diversity on FDI

Spillovers. In Table 13, we test for the presence of spillovers within states and between migrants from proximate origins. In column 1 of panel A, we use our simple specification from column 2 in Table 3, but add the total number of descendants of ancestry o at the state level.

[22]To separate firms into final-goods producers and intermediate-goods producers, we use the upstreamness index from Antràs et al. (2012). A sector is labelled as final goods (intermediate input) if its upstreamness index is below (above) 2.

We are able to identify the effect of this spillover at the state level by aggregating our instruments from equation (4) to the state level and including them as a separate set of instruments in the specification, such that both endogenous variables are identified. The coefficient on our measure of ancestry at the state level is -0.029 (s.e.=0.012), suggesting a negative and significant spillover from a larger presence of descendants from the same origin in the state on the effect of ancestry on FDI at the county level. In column 2, we instead include (and instrument for) the number residents in the nearest adjacent county with ancestry from the same origin country, where we find a negative, albeit insignificant, effect. Column 3 includes the number of descendants from origins within the same continent, where we find a positive but again insignificant effect. However, when we include in column 4 a measure for the number of descendants from the closest neighboring country, we find a negative and highly significant effect.

Overall, the evidence thus points to the presence of negative spillovers. The effect of ancestry on the extensive margin of FDI falls with the population of migrants from the same origin in the state as a whole, or with origin from neighboring countries. For example, a large Polish contingent in a given county has a lower effect on the probability of FDI with Poland if the state overall contains a large Polish contingent. Similarly, if the destination county also hosts a large number of descendants from a nearby origin, such as the Czech Republic, the Polish contingent has a smaller marginal effect on FDI. Thus, some substitutability seems to exist between migrants that come from geographically proximate countries.

In panel B, we repeat the same estimation for the intensive margin of FDI, but appear to lack the statistical power to identify significant spillovers.

Concavity of the effect. Up to this point, we have used a concave function linking the probability of FDI to the number of descendants from a given origin country. This concave functional implies a lower marginal effect of the 1001st descendant of Irish origin than for the first. However, an immediate implication of such concavity is that a more ethnically diverse population, combining many smaller communities from different origins, should generate more FDI than one large community of foreign descent. That is, the functional form linking ancestry to FDI is itself informative about the effect of ethnic diversity on FDI. To investigate the functional form linking FDI to ancestry, we perform several non-parametric and parametric tests.

We start with a flexible non-parametric estimation of the impact of ancestry on the probability of finding an FDI link, where we divide the absolute numbers of individuals of a given ancestry, $Ancestry_{o,d}^{2010}$, into quantiles, including the same covariates as in our simple specification from

column 2 in Table 3. We experiment with different number of quantiles: $Q = 3, 4, 5$.[23] We present the results in panel A of Appendix Table 10. Across all specifications, with terciles, quartiles, and quintiles, we systematically find the impact of ancestry is concave.

In panel B of Appendix Table 10, we offer a formal test to justify the functional form we use throughout the paper, $A_{o,d}^{2010} = \ln\left(1 + \frac{1}{1000} Ancestry_{o,d}^{2010}\right)$. To that end, we perform a non-linear least squares estimation of

$$\mathbf{1}\left[FDI_{o,d} > 0\right] = \delta_o + \delta_d + \beta \ln\left(1 + \kappa Ancestry_{o,d}^{2010}\right) + X'_{o,d}\gamma + \varepsilon_{o,d},$$

again including the same covariates as in our simple specification from column 2 in Table 3. We find a point estimate of $\beta = 0.157$ and $\kappa = 0.001$. This finding forms the basis for our choice of functional form applied throughout the paper. The functional form using natural logarithms is convenient because it offers a compact way to model the non-linear impact of ancestry. For small ancestry ($Ancestry_{o,d} \ll 1000$), the function $\ln\left(1 + Ancestry_{o,t}/1000\right)$ is approximately linear in $Ancestry_{o,d}$. For large ancestry ($Ancestry_{o,d} \gg 1000$), the function $\ln\left(1 + Ancestry_{o,t}/1000\right)$ behaves approximatively like $\ln Ancestry_{o,d}$. So for a large number of residents with foreign ancestry, the coefficient β in equation (1) is simply the elasticity of the extensive margin of FDI with respect to ancestry; for a small number of residents with foreign ancestry, the coefficient β measures the proportional impact of ancestry on the extensive margin of FDI.

In Appendix Table 11, we further explore the robustness of our results to alternative functional forms. In column 1, we simply measure ancestry in levels, and find a positive and significant effect. In column 2, we use $\ln\left(Ancestry_{o,d}^{2010}\right)$, and use the value -1 instead of $-\infty$ for $Ancestry_{o,d}^{2010} = 0$, and find a result similar to our baseline specification. In column 3, we use $\left(Ancestry_{o,d}^{2010}\right)^{1/3}$ as an alternative concave function, and find again a robust positive and significant effect of ancestry. In columns 4, 5, and 6, we replicate our results for different dates, 1980, 1990 and 2000, instead of 2010, and change the dates for our IV interaction terms accordingly. The estimated impact of ancestry on FDI varies little when we move the cutoff date to measure ancestry.

We conclude from this exploration of the functional form of the impact of ancestry on FDI that (i) the effect of ancestry is robust across different functional forms, and (ii) the data suggests a strongly non-linear impact of ancestry on FDI: changes in the number of residents with foreign ancestry matter more for FDI when few foreign descendants are present than when many are.

[23] The cutoffs for the number of residents in county d with ancestry from country o are $\{0; 145; 655; +\infty\}$ for terciles, $\{0; 108; 282; 1144; +\infty\}$ for quartiles, and $\{0; 92; 187; 455; 1660; +\infty\}$ for quintiles.

6 Conclusion

The economic effects of migration loom large in public debates about illegal immigration to the U.S. and the ongoing flow of migrants to Europe from places such as Syria, Afghanistan, and the Balkans. Much of the academic debate on the subject has focused on the relatively short-term consequences, identifying effects of immigration on local labor markets and consumer prices (Card, 1990; Cortes, 2008). In this paper, we add to this debate by showing causally identified evidence of an effect of migration, and the ethnic diversity resulting from it, on the propensity of firms based in the areas receiving migrants to interact economically with the migrants' origin countries. This effect of ancestry on FDI operates over long periods of time, spanning generations rather than decades, and explains an economically large share of the variation in patterns of FDI across US counties and states.

Our identification strategy uses 130 years of census data to isolate variation in today's ancestry composition of US counties that derives solely from the interaction of time-series variation in the relative attractiveness of different destinations within the United States with the staggered timing of factors that drove out-migration from the migrants' countries of origin. This approach allows us to generate four main insights.

First, we are able to causally identify and quantify the effect of ancestry on FDI in a setting with a high degree of external validity while guarding against a wide range of possible confounding factors, including unobserved origin and destination effects. We find that a doubling of a US county's residents with ancestry from a given foreign country relative to the mean increases by 4.2 percentage points the probability that at least one local firm engages in FDI with that country.

Second, the presence of descendants of first-generation immigrants rather than first-generation immigrants themselves seems to largely generate the majority of the effect of ancestry on FDI. The effect of ancestry on FDI is thus long lasting and appears to unfold over generations rather than years, where even the earliest migrations for which we have data going back to the 19th century significantly affect the pattern of FDI today.

Third, the effect of ancestry on FDI increases with the geographic distance to the origin country and the quality of its institutions. Once these factors are controlled for, other measures of genetic, linguistic, and religious distance do not significantly affect the size of the effect of ancestry on FDI. The effect also does not appear to vary systematically between firms producing final and intermediate goods and small and large firms.

Finally, we find a range of results that show a positive effect of ethnic diversity on FDI. The most obvious of these findings is the strong indication of concavity in the number of descendants

of migrants from a given origin, such that a more ethnically diverse population, combining many smaller communities from different origins, should generate more FDI than one large community of foreign descent. Further, we find negative spillovers both within states and between migrants from geographically proximate countries, such that a larger community of the same ethnic descent in surrounding counties or a larger community of descent from a neighboring country decreases the effect of ancestry on FDI. In addition, the effect of ancestry on FDI significantly increases with the diversity of the community of residents with foreign ancestry. All three findings taken together suggest ethnic diversity may be a quantitatively important driver of FDI.

Taken together, our results suggest that receiving migration from a foreign country has a positive long-term effect on the ability of local firms to interact economically with the migrants' country of origin. This effect increases with the institutional quality of the origin country, suggesting that, for example, receiving migrants from a war-torn country may have larger positive effects once the country stabilizes. The collage of our results also appears more consistent with a model in which common ancestry mitigates informational frictions, but does not operate though contract enforcement or common tastes. In the presence of informational frictions, common ancestry acts as a conduit for transmitting information. This information channel is even more important for remote countries (the positive interaction with distance). Information transmissions tends to follow the shortest path, so that a small increment in the number of residents with common ancestry matters more when few residents have information about a country than when many do (the concave impact of ancestry and the negative spillover effects). On the other hand, common ancestry does not facilitate FDI precisely when in weak judicial environments (the positive interaction with judicial quality), so it does not seem to be a substitute for contract enforcement. Finally, the fact that the effect of ancestry on FDI does not seem to vary across firms producing final versus intermediate goods appears to exclude mechanisms that rely on common tastes between descendants of migrants and their countries of origin. However, all of our evidence on the mechanism through which ancestry facilitates FDI is indirect. We leave a thorough study of this mechanism for future research.

References

Ager, P. and M. Brückner (2013): "Cultural divsersity and economic growth: Evidence from the US during the age of mass migration," *European Economic Review*, 64, 76–97.

Aleksynska, M. and G. Peri (2014): "Isolating the Network Effect of Immigrants on Trade," *The World Economy*, 37, 434–45.

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

Alesina, A., J. Harnoss, and H. Rapoport (2015): "Birthplace Diversity and Economic Prosperity," Working paper.

Alesina, A., S. Michalopoulos, and E. Papaioannou (Forthcoming): "Ethnic Inequality," *Journal of Political Economy*.

Antràs, P., D. Chor, T. Fally, and R. Hillberry (2012): "Measuring the Upstreamness of Production and Trade Flows," *American Economic Review Papers and Proceedings*, 102, 412–416.

Arkolakis, C., A. Costinot, and A. Rodríguez-Clare (2012): "New Trade Models, Same Old Gains?" *American Economic Review*, 102, 94–130.

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

Bartik, T. J. (1991): *Who benefits from state and local economic development policies?*, no. wbsle in Books from Upjohn Press, W.E. Upjohn Institute for Employment Research.

Borjas, G. J. (1994): "The Economics of Immigration," *Journal of Economic Literature*, XXXII, 1667–1717.

——— (2003): "The Labor Demand Curve is Downward Sloping: Reexamining the Impact of Immigration on the Labor Market," *The Quarterly Journal of Economics*, 118, 1335–1374.

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

Card, D. (1990): "The Impact of the Mariel Boatlift on the Miami Labor Market," *Industrial and Labor Relations Review*, 43, 245–257.

Card, D. and J. Di Nardo (2000): "Do Immigrant Inflows Lead to Native Outflows?" *American Economic Review*, 90, 360–367.

Carr, D. L., J. R. Markusen, and K. E. Maskus (2001): "Estimating the Knowledge-Capital Model of the Multinational Enterprise," *American Economic Review*, 91, 693–708.

Chaney, T. (2014a): "The network structure of international trade," *The American Economic Review*, 104, 3600–3634.

——— (2014b): "Networks in International Trade," in *Oxford Handbook of the Economics of Networks*, ed. by Y. Bramoulle, A. Galleoti, and B. Rogers, Oxford University Press.

Cohen, L., U. Gurun, and C. Malloy (2015): "Resident Networks and Firm Value," *The Journal of Finance*, forthcoming.

Combes, P., M. Lafourcade, and T. Mayer (2005): "The trade-creating effects of business and social networks: Evidence from France." *Journal of International Economics*, 66 (1), 1–29.

Cortes, P. (2008): "The Effect of Low-Skilled Immigration on U.S. Prices: Evidence from CPI Data," *Journal of Political Economy*, 116, pp. 381–422.

Daniels, R. (2002): *Coming to America*, HarperCollins Publishers.

Docquier, F., G. Peri, and I. Ruyssen (2014): "The Cross-country Determinants of Potential and Actual Migration," *International Migration Review*, 118, S37–S99.

Dunlevy, J. A. (2006): "The influence of corruption and language on the protrade effect of immigrants: Evidence from the American states," *Review of Economics and Statistics*, 88, 182–186.

Felbermayr, G. J. and F. Toubal (2012): "Revisiting the trade-migration nexus: Evidence from new OECD data," *World Development*, 40, 928–937.

Friedberg, R. (2001): "The impact of mass migration on the Israeli labor market." *Quarterly Journal of Economics*, 116 (4), 1373–1408.

Fuchs-Schuendeln, N. and T. A. Hassan (2015): "Natural Experiments in Macroeconomics," Working Paper 21228, National Bureau of Economic Research.

Fulford, S. L., I. Petkov, and F. Schiantarelli (2015): "Does It Matter Where You Came From? Ancestry Composition and Economic Performance of U.S. Counties, 1850-2010," Institute for the Study of Labor (IZA) Discussion Paper No. 9060.

Garmendia, A., C. Llano, A. Minondo, and F. Requena (2012): "Networks and the disappearance of the intranational home bias," *Economics Letters*, 116, 178–182.

Goldin, C. (1994): "The Political Economy of Immigration Restriction in the United States, 1890 to 1921," in *The Regulated Economy: A Historical Approach to Political Economy*, ed. by C. Goldin and G. D. Libecap, University of Chicago Press, 223–258.

Gould, D. M. (1994): "Immigrant links to the home country: Empirical implications for U.S. bilateral trade flows." *The Review of Economics and Statistics*, 76, 302–316.

Granovetter, M. (1973): "The strength of weak ties," *American Journal of Sociology*, 78, 1360–1380.

Greene, W., C. Han, and P. Schmidt (2002): "The bias of the fixed effects estimator in nonlinear models," Unpublished Manuscript, Stern School of Business, NYU.

Guiso, L., P. Sapienza, and L. Zingales (2009): "Cultural biases in economic exchange." *The Quarterly Journal of Economics*, 124, 1095–1131.

Head, K. and J. Ries (1998): "Immigration and Trade Creation: Econometric Evidence from Canada," *Canadian Journal of Economics*, 31, 47–62.

——— (2008): "FDI as an Outcome of the Market for Corporate Control: Theory and Evidence," *Journal of International Economics*, 74, 2–20.

Heckman, J. J. (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 153–161.

Helpman, E., M. Melitz, and Y. Rubinstein (2008): "Estimating Trade Flows: Trading Partners and Trading Volumes," *Quarterly Journal of Economics*, 123, 441–487.

Jensen, E. B., R. Bhaskar, and M. Scopilliti (2015): "Demographic Analysis 2010: Estimates of Coverage of the Foreign-Born Population in the American Community Survey," Tech. rep., U.S. Census.

Katz, L. F. and K. M. Murphy (1992): "Changes in Relative Wages, 1963-1987: Supply and Demand Factors," *Quarterly Journal of Economics*, 107, 35–78.

Kaufmann, D., A. Kraay, and M. Mastruzzi (2003): "Governance Matters III: Governance Indicators for 1996–2002," Working Paper No. 3106, World Bank.

Nunn, N. (2007): "Relationship-Specificity, Incomplete Contracts, and the Pattern of Trade," *Quarterly Journal of Economics*, 122, 569–600.

Ottaviano, G. I. and G. Peri (2006): "The economic value of cultural diversity: evidence from US cities," *Journal of Economic Geography*, 6, 9–44.

Parsons, C. and P.-L. Vezina (2014): "Migrant Networks and Trade: The Vietnamese Boat People as a Natural Experiment," Mimeo University of Oxford.

Peri, G. (2012): "The Effect of Immigration on Productivity: Evidence from U.S. States," *The Review of Economics and Statistics*, 94, 348–358.

Portes, R. and H. Rey (2005): "The Determinants of Cross-Border Equity Flows," *Journal of International Economics*, 65, 269–296.

Putterman, L. and D. N. Weil (2010): "Post-1500 Population Flows and the Long-Run Determinants of Economic Growth and Inequality," *The Quarterly Journal of Economics*, 125, 1627–1682.

Ramondo, N. (2014): "A Quantitative Approach to Multinational Production," *Journal of International Economics*, 93, 108–122.

Rauch, J. and V. Trindade (2002): "Ethnic Chinese Networks In International Trade," *The Review of Economics and Statistics*, 84, 116–130.

Razin, A., Y. Rubinstein, and E. Sadka (2003): "Which countries export FDI, and how much?" Tech. rep., National Bureau of Economic Research.

Spolaore, E. and R. Wacziarg (2015): "War and Relatedness," *The Review of Economics and Statistics*, forthcoming.

THERNSTROM, S. (1980): *Havard encyclopedia of American ethnic groups*, Harvard University Press, Cambridge MA.

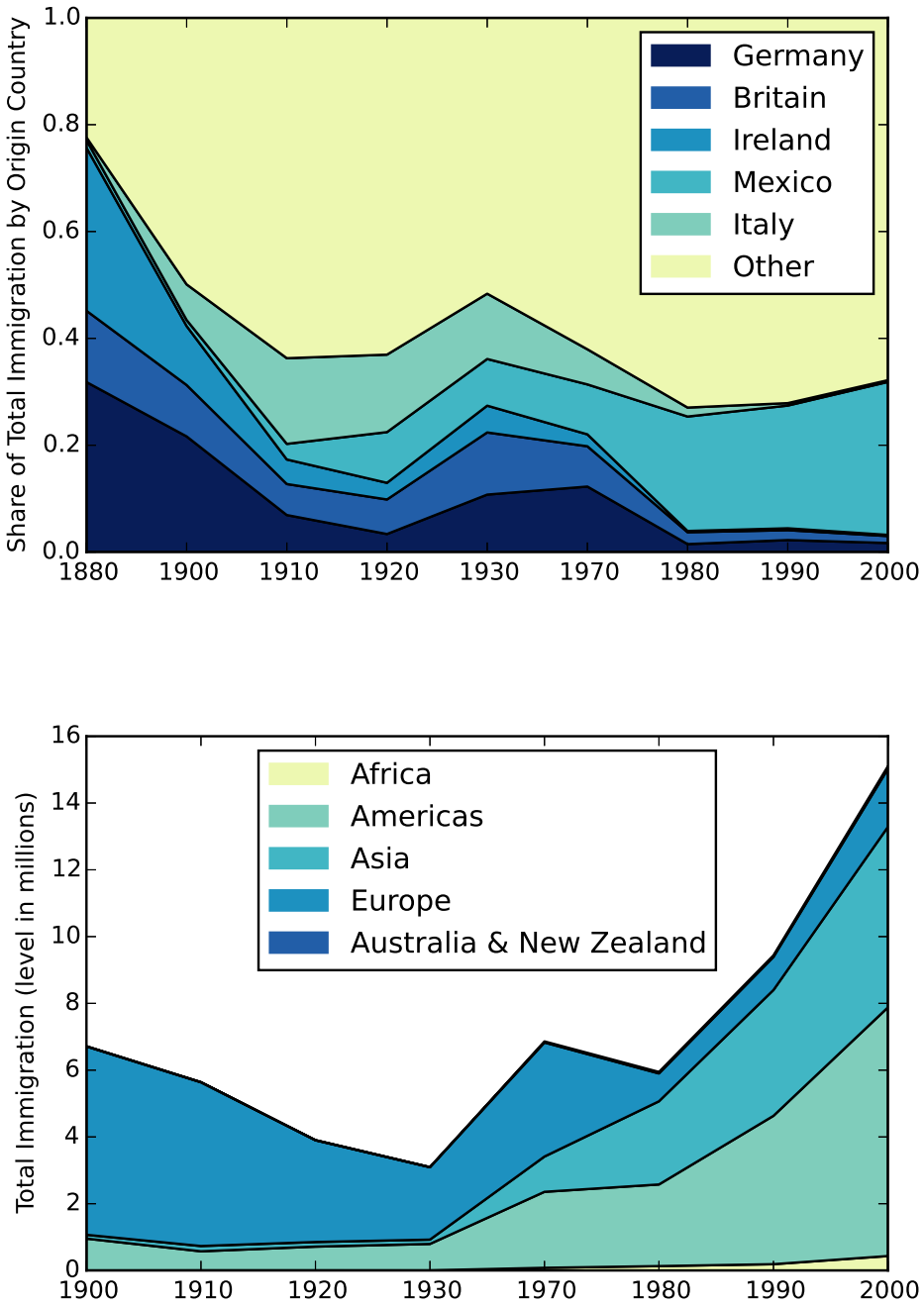


FIGURE 1: ORIGINS OF IMMIGRANTS TO THE UNITED STATES, PRE-1880 TO 2000

Notes: The upper sub-figure depicts the share of total immigration to the United States for the largest five origin countries of US residents that claim foreign ancestry: Germany, Britain, Ireland, Mexico, and Italy. The lower panel shows the the number of migrants (in millions) by continent of origin.

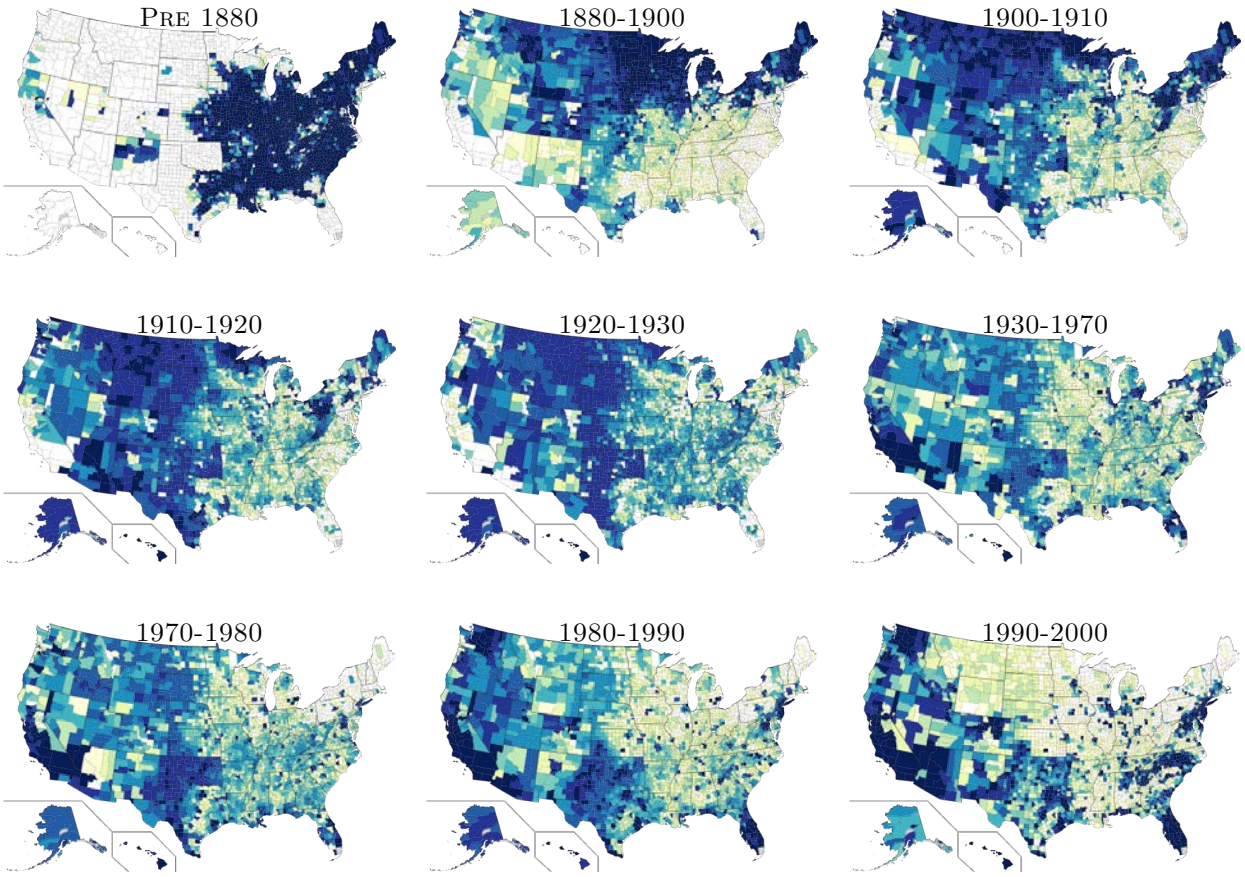


Figure 2: Destinations of Immigrants to the United States, pre-1880 to 2000

Notes: This figure maps the residual immigration flow into the United States by census decades. We regress the log number of immigrants into US county d at time t, I_d^t, on destination county d and year t fixed effects, and calculate the residuals. The maps' color coding depicts the residuals' decile in the distribution of residuals across counties and census years. Darker colors indicate a higher decile.

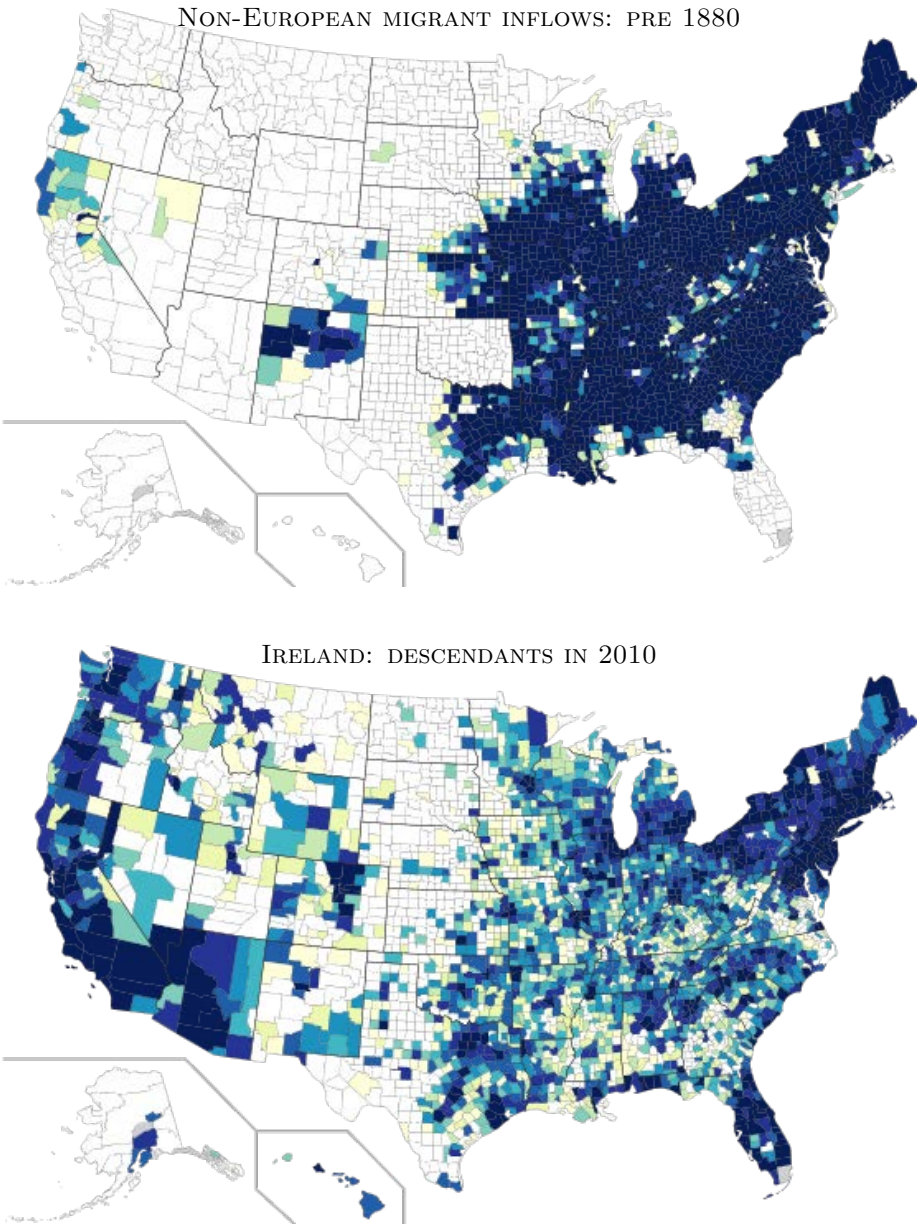


Figure 3: Migrants and Ancestors: The Case of Ireland pre-1880

Notes: The upper panel presents the pre-1880, non-European residual immigration flow into the United States. The residuals are generated in the same way as in Figure 2. The lower part of the map presents the residual Irish ancestry in 2010. We regress our measure of ancestry in 2010 for all origin-destination pairs on destination county d and origin country o fixed effects and calculate the residuals of Irish ancestry in 2010. The maps' color coding depicts the residuals' decile in the distribution of Irish ancestry residuals across counties. Darker colors indicate a higher decile.

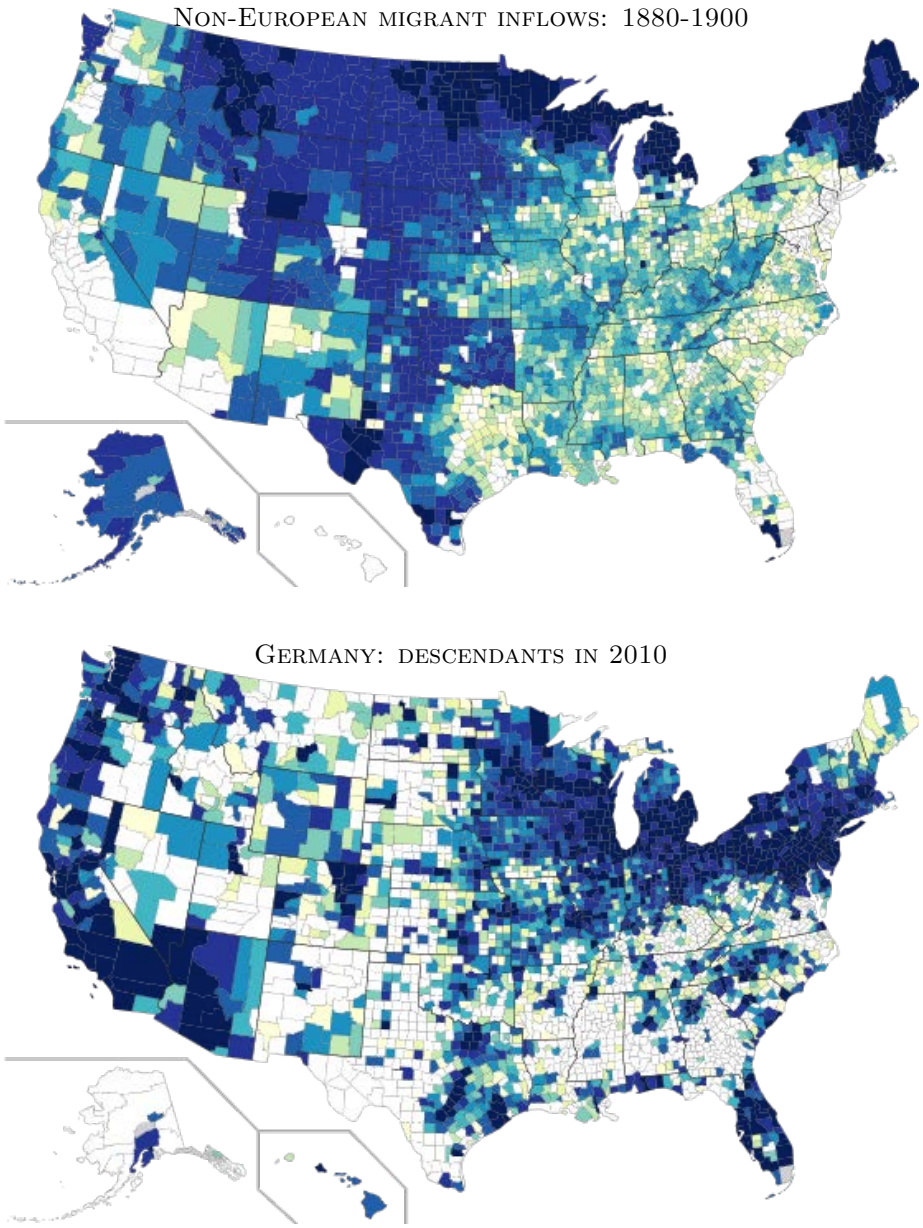


Figure 4: Migrants and Ancestors: The Case of Germany 1880-1900

Notes: The upper map presents the 1880-1900, non-European residual immigration flow into the United States. The residuals are generated in the same way as in Figure 2. The lower part of the map presents the residual German ancestry in 2010. We regress our measure of ancestry in 2010 for all origin-destination pairs on destination county d and origin country o fixed effects and calculate the residuals of German ancestry in 2010. The maps' color coding depicts the residuals' decile in the distribution of German ancestry residuals across counties. Darker colors indicate a higher decile.

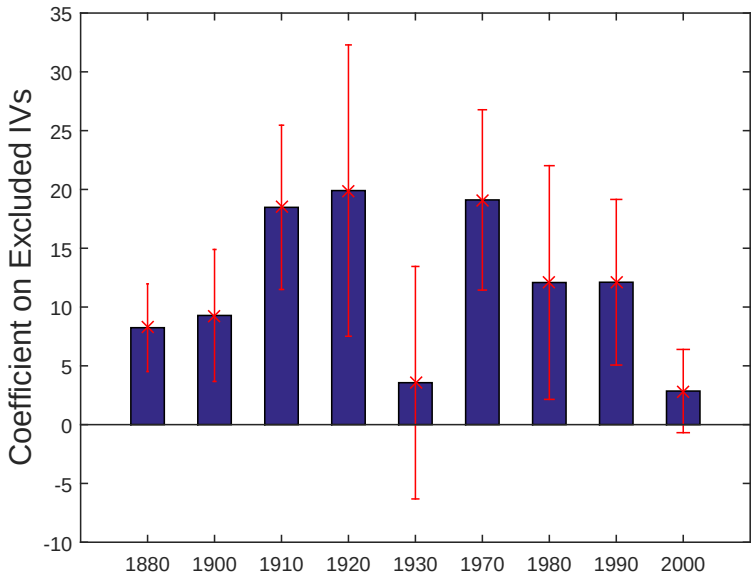


FIGURE 5: FIRST-CTAGE COEFFICIENTS

Notes: Coefficient estimates (bars) and 95% confidence intervals (red lines) on the excluded instruments $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_{t=1880..2000}$ from Table 2, column 2. The dependent variable is Log Ancestry 2010. Robust standard errors are clustered at the origin-country level.

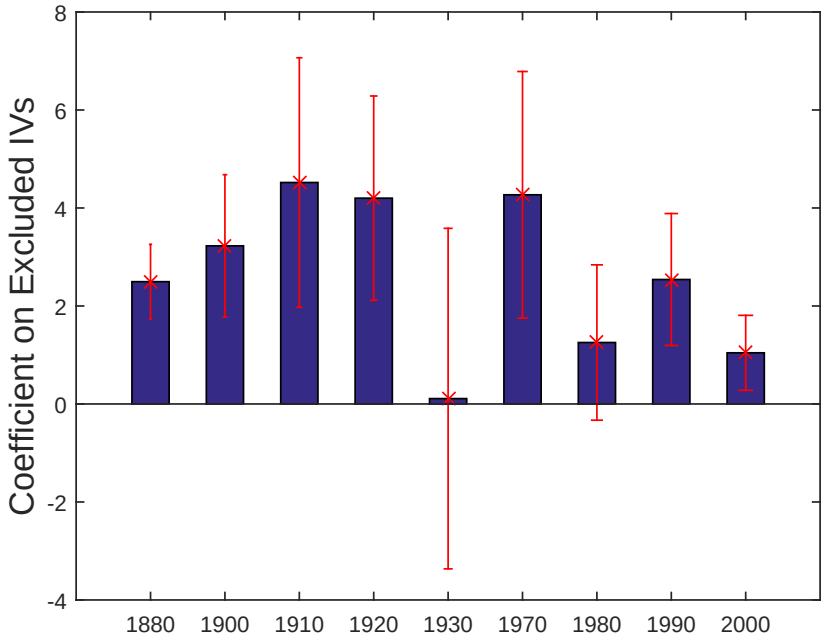


FIGURE 6: SECOND-STAGE COEFFICIENTS

Notes: Coefficient estimates (bars) and 95% confidence intervals (red lines) on the excluded instruments $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_{t=1880..2000}$ from a reduced form regression similar to Table 2, column 2. The dependent variable is the 2014 FDI dummy. All coefficients are multiplied by 100. Robust standard errors are clustered at the country level. The R^2 of this regression is 0.218.

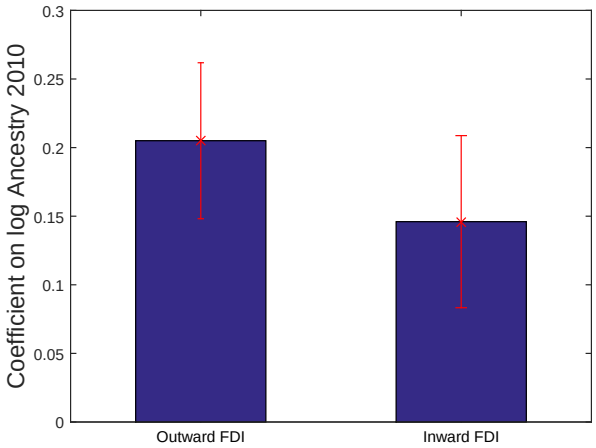


FIGURE 7: ANCESTRY, INWARD AND OUTWARD FDI (EXTENSIVE MARGIN)

Notes: The figure depicts coefficient estimates (and 95% confidence intervals) on the Log ancestry 2010 from extensive margin regressions. The specification is the same as that in Table 3, column 3. The dependent variable for the left bar is a dummy variable on outward FDI, which is equal to 1 if at least one firm in a destination county is the parent of a foreign subsidiary in origin country o. The dependent variable for the right bar is a dummy variable on inward FDI, which is equal to 1 if at least one firm in a destination county is a foreign affiliate of a parent firm in the origin country. The bars in the figure present the respective coefficient estimates; the red lines give 95% confidence intervals. Standard errors are clustered at the country level to account for potential heteroscedasticity.

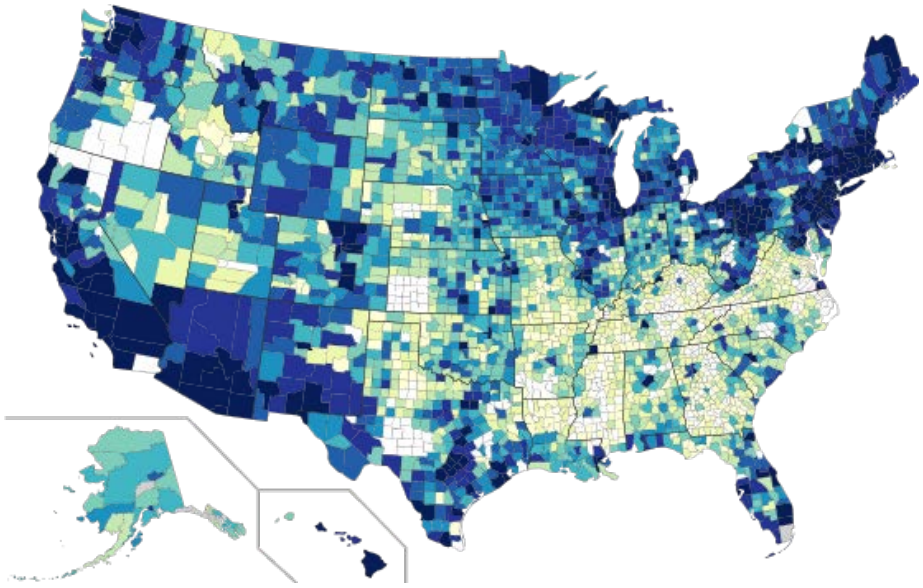


FIGURE 8: COUNTERFACTUAL EXPERIMENT: REMOVING THE CHINESE EXCLUSION ACT

Notes: The map depicts for each US. county the expected increase in the probability of having positive FDI relations with China in a counterfactual world where the "Chinese Exclusion" Act of 1882 and the Magnuson Act of 1943 had never been passed, that is, if Chinese immigration to the United States had been free from 1882 to 1965. Darker color indicates bigger increase. The cutoff values in the categories (from light to dark) are 0.00002, 0.00006, 0.00011, 0.00018, 0.000285, 0.000433, 0.000745, 0.00138 and 0.00322. The details for the construction of this counterfactual are presented in section 4.4.

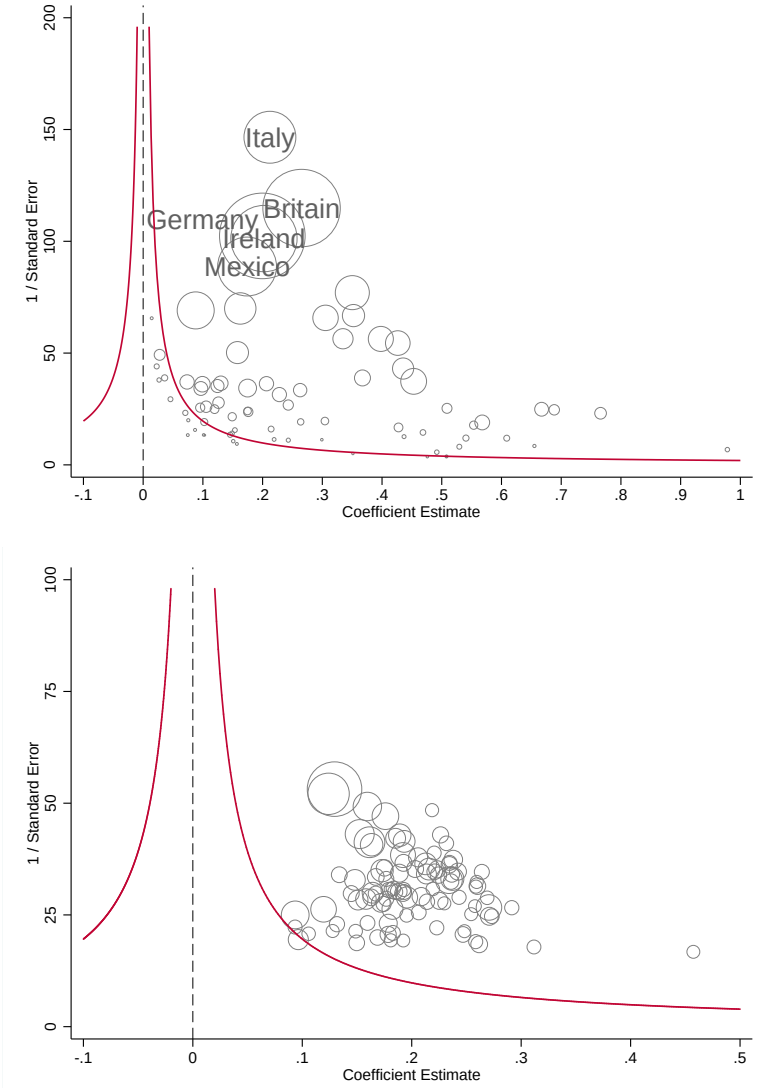


FIGURE 9: HETEROGENEOUS ESTIMATES ACROSS COUNTRIES AND COUNTIES

Notes: The figure depicts funnel plots at both country (upper part) and county (lower part) levels. To generate the country-level plot, we run an IV regression of the FDI dummy on the log 2010 ancestry for each origin country. To generate the county-level plot, we run an IV regression of the FDI dummy on the log 2010 ancestry for each destination county. In both parts, we use $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_{t=1880..2000}$ and principal components as excluded instruments, and control for log distance as well as latitude difference. In both parts, we plot the estimated coefficients (x axis) against the reciprocal of estimated standard errors on ancestry. The size of the circle is proportional to the size of country ancestry (upper part) and to the size of county population (lower part). The imposed curve is $y = 1.96/x$ for positive x region and $y = -1.96/x$ for negative x region, and circles to the right of the curve indicate statistically significant coefficients.

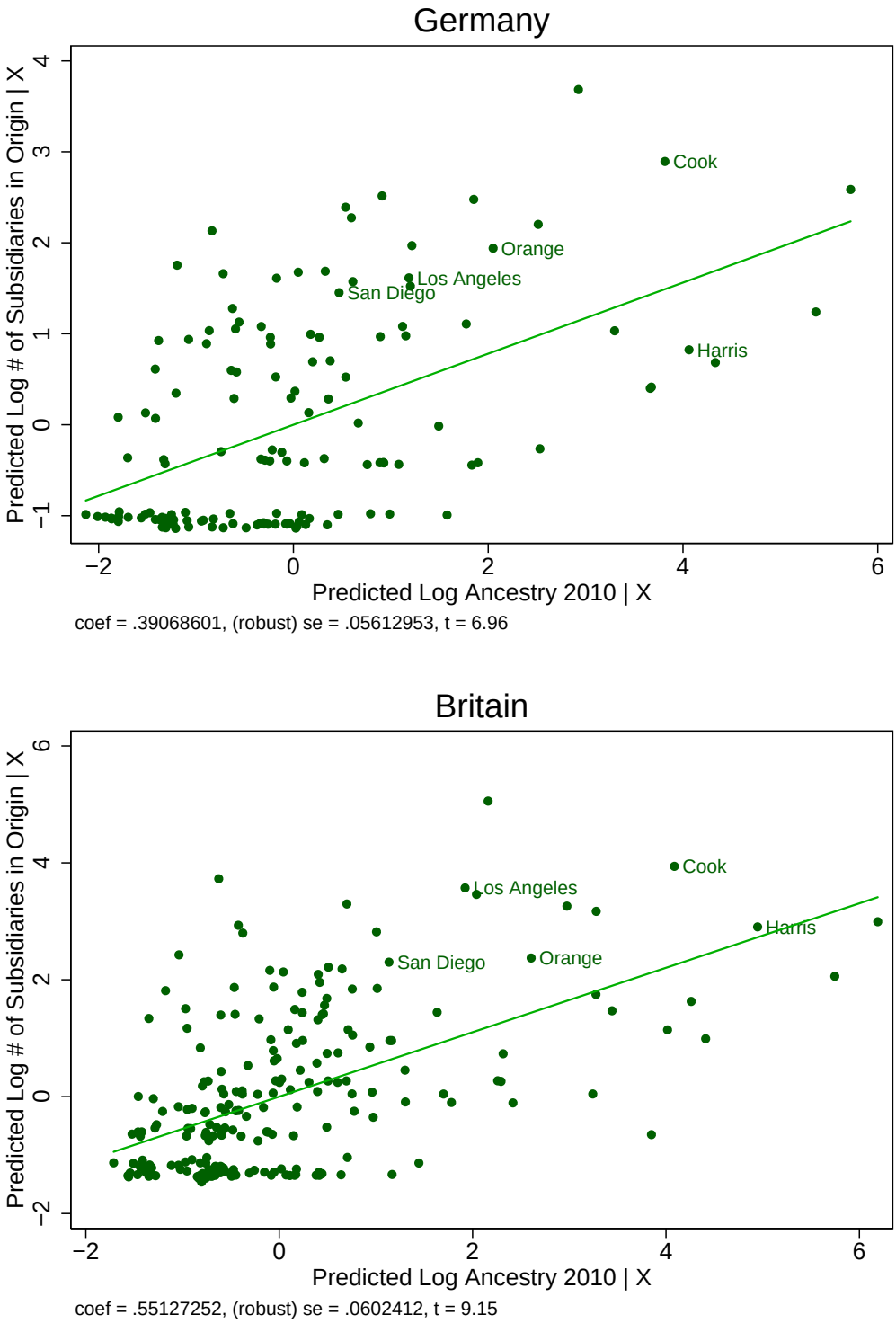
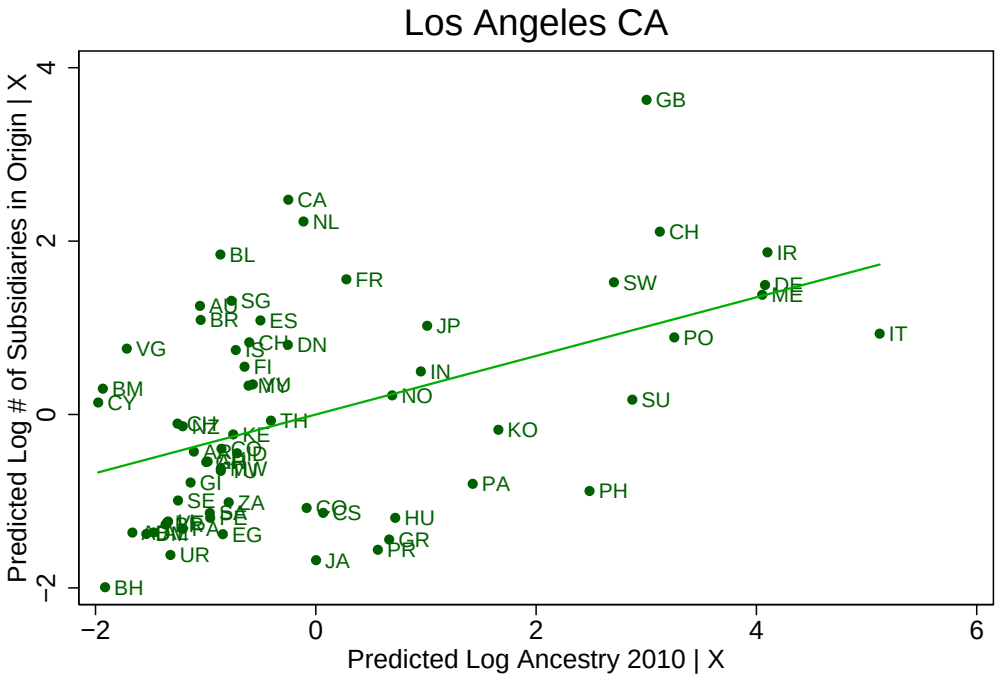


FIGURE 10: ANCESTRY AND FDI: GERMANY AND BRITAIN

Notes: The upper part of the figure is a conditional scatterplot of log 2010 German ancestry and log # of subsidiaries in origin country. The lower part is for British ancestry. The corresponding regression uses the same specification as in Table 6, column 3. The solid line depicts the fitted regression line.



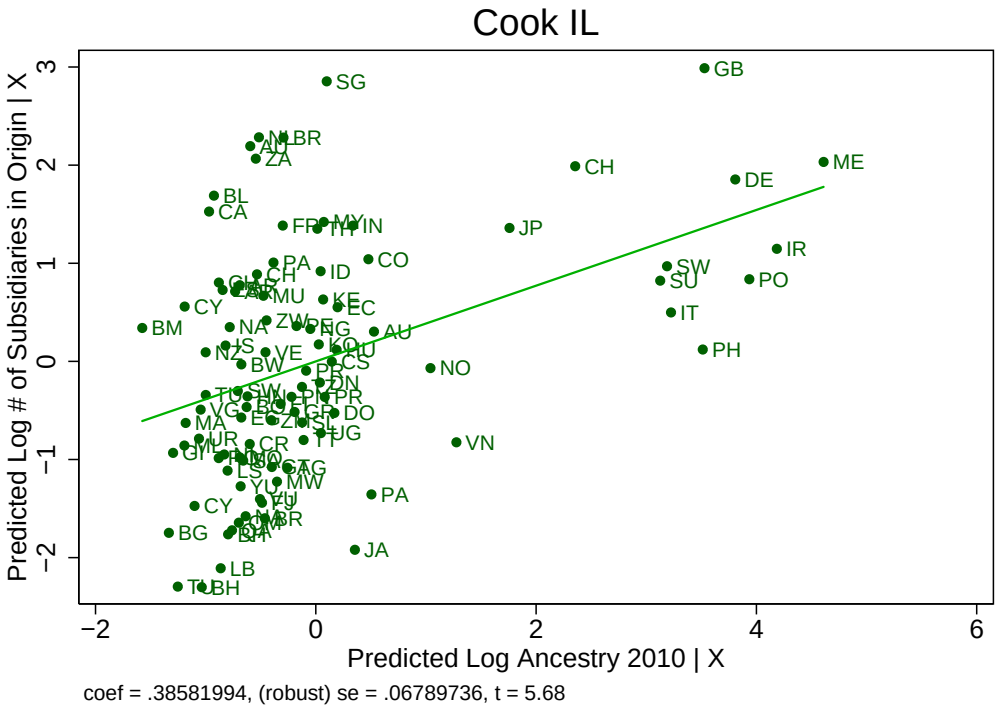


FIGURE 11: ANCESTRY AND FDI: LA AND COOK COUNTIES

Notes: The upper part of the figure is a conditional scatterplot of log 2010 ancestry in Los Angeles, CA and log # of subsidiaries in origin country. The lower part is for Cook County, IL. The corresponding regression uses the same specification as in Table 6, column 3. The solid line depicts the fitted regression line.

Table 1: Summary Statistics

	(1)	(2)	(3)	(4)
	All	*Ancestry > 0*		
		All	*Bottom Quintile*	*Top Quintile*
Panel A: Origin-destination pairs				
FDI Dummy	0.018	0.031	0.003	0.128
	(0.132)	(0.173)	(0.053)	(0.334)
Ancestry 2010 (in thousands)	0.318	0.580	0.000	2.871
	(5.378)	(7.246)	(0.000)	(16.000)
Total Immigrants 2010 (in thousands)	0.029	0.052	0.001	0.226
	(0.352)	(0.474)	(0.013)	(1.018)
Immigrants between 1990-2000 (in thousands)	0.025	0.045	0.001	0.200
	(1.024)	(1.382)	(0.007)	(3.071)
Geographic Distance (km)	9122.393	8397.379	9123.499	7454.962
	(3802.105)	(3763.707)	(4315.761)	(2921.581)
Latitude Difference (degree)	19.440	16.319	19.064	13.512
	(11.312)	(10.902)	(11.397)	(8.632)
# of Total FDI	0.196	0.352	0.028	1.623
	(5.490)	(7.401)	(1.456)	(16.305)
# of Subsidiaries in Origin	0.033	0.060	0.003	0.269
	(1.345)	(1.813)	(0.283)	(3.842)
# of Parents in Destination	0.015	0.027	0.001	0.125
	(0.407)	(0.548)	(0.103)	(1.200)
# of Employees at Subsidiary in Origin (in thousands)	0.039	0.069	0.010	0.319
	(4.941)	(6.661)	(1.298)	(14.749)
# of Subsidiaries in Destination	0.068	0.122	0.011	0.564
	(1.903)	(2.565)	(0.543)	(5.667)
# of Parents in Origin	0.079	0.143	0.012	0.665
	(2.282)	(3.077)	(0.575)	(6.810)
# of Employees at Subsidiary in Destination (in thousands)	1.873	3.398	0.087	16.531
	(86.649)	(116.896)	(5.650)	(260.727)
N	612495	336382	67277	67276
Panel B: Counties				
2010 Share of Population with Foreign Ancestry	0.693	0.694	0.661	0.709
	(0.300)	(0.299)	(0.312)	(0.237)
2010 Diversity of Ancestries	0.785	0.784	0.761	0.831
	(0.073)	(0.073)	(0.068)	(0.076)
2010 Population # (in thousands)	98.389	98.433	6.778	384.363
	(313.143)	(313.238)	(4.873)	(621.566)
2010 Per Capita Income (in thousand dollar)	34.100	34.097	32.708	39.058
	(7.805)	(7.807)	(8.475)	(9.068)
N				
Panel C: Countries				
Genetic Distance	0.103	0.084	0.106	0.064
	(0.053)	(0.041)	(0.050)	(0.036)
N	155	119	18	25
Linguistic Distance	0.950	0.937	0.990	0.922
	(0.110)	(0.121)	(0.010)	(0.114)
N	132	103	8	26
Religious Distance	0.820	0.807	0.923	0.730
	(0.129)	(0.137)	(0.050)	(0.129)
N	131	101	8	25
2010 Per Capita GDP (in thousand dollar)	14.698	18.144	31.903	27.175
	(22.917)	(24.990)	(40.999)	(21.956)
N	165	126	24	27
Judicial Quality	0.503	0.537	0.546	0.681
	(0.208)	(0.214)	(0.224)	(0.197)
N	144	115	15	26
2010 Country Diversity	0.439	0.405	0.433	0.220
	(0.270)	(0.256)	(0.246)	(0.185)
N	163	122	20	27

Notes: The table presents means (and standard deviations). Variables in Panel A refer to our sample of (country,county) pairs used in Tables 2, 3, 4, 9, 6, 5, 13, and Appendix Table 4. Variables in Panel B refer to our sample of counties used in Table 11. Variables in Panel C refer to our sample of counties used in Table 10. Column 1 shows data for all observations. Columns 2 to 4 show all bottom quintile and top quintile of observations with positive ancestry. In Panel A, the FDI dummy is a dummy variable equal to 1 if the destination county has either subsidiaries or shareholders in the origin country. In Panel B, population is missing for for three counties and per-capita income is missing for eight counties. The ancestry-diversity variable is computed as 1 minus the Herfindahl index of ethnolinguistic group shares in each county. The details of variables in Panel C are given in the Data Appendix.

Table 2: First-stage: The Effect of Historical Immigration on Ancestry

	Log Ancestry 2010								*Ancestry 2010*
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$I^{1880}_{o,-r(d)} \frac{I^{1880}_{-c(o),d}}{I^{1880}_{-c(o)}}$	8.38***	8.24***	13.55***	10.36***	8.28***	8.28***	8.77***		3.07***
	(1.91)	(1.90)	(1.54)	(1.30)	(1.16)	(1.16)	(1.11)		(0.37)
$I^{1900}_{o,-r(d)} \frac{I^{1900}_{-c(o),d}}{I^{1900}_{-c(o)}}$	9.57***	9.28***	16.48***	11.19***	8.19***	8.22***	10.14***	15.73***	2.73*
	(2.90)	(2.86)	(3.71)	(3.71)	(2.99)	(2.99)	(2.86)	(2.72)	(1.65)
$I^{1910}_{o,-r(d)} \frac{I^{1910}_{-c(o),d}}{I^{1910}_{-c(o)}}$	18.40***	18.48***	20.48***	15.25***	12.17***	12.18***	11.47***	10.61**	3.88**
	(3.58)	(3.56)	(4.08)	(4.23)	(4.40)	(4.39)	(4.10)	(4.24)	(1.80)
$I^{1920}_{o,-r(d)} \frac{I^{1920}_{-c(o),d}}{I^{1920}_{-c(o)}}$	19.40***	19.90***	26.61***	23.10***	23.80***	23.81***	19.24***	29.64***	9.47***
	(6.33)	(6.32)	(5.65)	(4.93)	(2.97)	(2.98)	(3.70)	(4.21)	(2.27)
$I^{1930}_{o,-r(d)} \frac{I^{1930}_{-c(o),d}}{I^{1930}_{-c(o)}}$	3.67	3.57	4.44	4.57	5.30	5.38	3.95	8.05	9.56***
	(5.03)	(5.04)	(4.91)	(4.76)	(3.89)	(3.92)	(4.35)	(4.98)	(2.49)
$I^{1970}_{o,-r(d)} \frac{I^{1970}_{-c(o),d}}{I^{1970}_{-c(o)}}$	18.93***	19.11***	18.48***	15.78***	14.90***	14.89***	17.16***	19.24***	6.33***
	(3.91)	(3.91)	(3.52)	(3.45)	(3.03)	(3.03)	(2.53)	(3.22)	(0.51)
$I^{1980}_{o,-r(d)} \frac{I^{1980}_{-c(o),d}}{I^{1980}_{-c(o)}}$	12.11**	12.08**	15.50***	15.03***	14.41***	14.40***	15.60***	24.66***	10.28***
	(5.06)	(5.07)	(5.61)	(5.61)	(4.96)	(4.97)	(3.55)	(6.65)	(1.96)
$I^{1990}_{o,-r(d)} \frac{I^{1990}_{-c(o),d}}{I^{1990}_{-c(o)}}$	11.95***	12.11***	11.02***	11.01***	9.31**	9.32**	11.44***	13.27**	11.21***
	(3.58)	(3.59)	(3.70)	(3.63)	(3.75)	(3.76)	(3.57)	(5.78)	(3.46)
$I^{2000}_{o,-r(d)} \frac{I^{2000}_{-c(o),d}}{I^{2000}_{-c(o)}}$	2.88	2.86	3.89**	4.12***	4.25***	4.26***	2.91	7.54***	2.37*
	(1.80)	(1.81)	(1.72)	(1.54)	(1.44)	(1.45)	(2.46)	(2.25)	(1.24)
$I^{2010}_{o,-r(d)} \frac{I^{2010}_{-c(o),d}}{I^{2010}_{-c(o)}}$							140.21***		
							(26.51)		
R^2	0.58	0.59	0.62	0.71	0.75	0.75	0.77	0.76	0.44
F Stat on excluded IVs	12.7	13.1	2769.1	433.4	91.5	91.1	244.0	85.2	1021.0
p-value on F Stat	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Distance	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Latitude Difference	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
3rd order poly in dist and lat	No	No	No	No	No	Yes	No	No	No
Destination × Continent FE	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Origin × Census Region FE	No	No	No	No	Yes	Yes	Yes	Yes	Yes
Principal Components	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The table presents coefficient estimates of our first stage equation (4) at the country-county level. All specifications control for origin and destination fixed effects. Standard errors are given in parentheses and are clustered at the country level. One, two, and three asterisks denote statistical significance at the 10%, 5%, and 1% levels, respectively. Except for column (9), which uses the level of the 2010 ancestry stock as the dependent variable, all other dependent variables are the log of the 2010 ancestry. All coefficients and standard errors are multiplied by 100.

Table 3: Second-stage: The effect of Ancestry on FDI

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: IV			*FDI 2014 (Dummy)*			
Log Ancestry 2010	0.243***	0.190***	0.197***	0.197***	0.190***	0.201***
	(0.024)	(0.024)	(0.030)	(0.030)	(0.033)	(0.031)
Log Distance	0.009	0.004	0.026		0.024	-0.024
	(0.010)	(0.009)	(0.031)		(0.029)	(0.027)
Latitude Difference	0.007**	0.006***	0.006**		0.006**	0.004
	(0.003)	(0.002)	(0.003)		(0.003)	(0.004)
N	612495	612495	612495	612495	612495	612300
Panel B: OLS			*FDI 2014 (Dummy)*			
Log Ancestry 2010	0.176***	0.176***	0.155***	0.155***	0.155***	0.171***
	(0.016)	(0.016)	(0.018)	(0.018)	(0.018)	(0.019)
R^2	0.2963	0.2963	0.3633	0.3633	0.3633	0.3932
N	612495	612495	612495	612495	612495	612495
Principal Components	No	Yes	Yes	Yes	Yes	Yes
$I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})$	No	No	No	No	Yes	No
Destination × Continent FE	No	No	Yes	Yes	Yes	Yes
Origin × Census Region FE	No	No	Yes	Yes	Yes	No
Origin × State FE	No	No	No	No	No	Yes
3rd order poly in dist and lat	No	No	No	Yes	No	No

Notes: The table presents coefficient estimates from IV (Panel A) and OLS (Panel B) regressions of equation (1) at the country-county level. The dependent variable in all panels is a dummy indicating an FDI relationship between origin o and destination d in 2014. The main variable of interest is *Log Ancestry 2010*. For all columns in Panel A, we include $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_{t=1880..2000}$ as excluded instruments. All specifications control for log distance, latitude difference, origin, and destination fixed effects. Standard errors are given in parentheses. Standard errors are clustered at the country level. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. (We also run an IV probit regression using a similar specification as in column 2, and the estimated coefficient on *Log Ancestry 2010* is 0.118***(0.039).)

Table 4: The Effect of Ancestry on FDI: The Case of Communist Countries

	Soviet Union	China	Viet-nam	Eastern Europe
Years excluded	1918-1990	1949-1980	1975-1996	1945-1989
	FDI 2014 (Dummy)			
Log Ancestry 2010	0.251***	0.588	0.128	0.109**
	(0.056)	(0.432)	(0.104)	(0.055)
R^2	0.1016	0.3038	0.1967	0.1742
N	3141	3141	3141	18846
F Stat on excluded IVs	5.076	0.404	13.267	37.315

Notes: The table presents coefficient estimates from IV regressions of equation (1) at the country-county level. Each column uses data from a subset of origin countries: Soviet Union (column 1), China (column 2), Vietnam (column 3), as well as Albania, Bulgaria, Czechoslovakia, Hungary, Poland, and Romania (column 4). The dependent variable in all columns is the dummy of FDI in 2014. The main variable of interest in all columns is the log of 2010 ancestry. All columns use as instruments the same set of variables as column 3 of Table 3, but only the immigration terms between *Closure start* and *Closure end* are excluded instruments; the remaining variables are included as controls. All specifications control for log distance, latitude difference, and origin fixed effects. The coefficient estimates on these specifications are not reported in the interest of space. Standard errors are given in parentheses. Standard errors are clustered at the country level. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Table 5: The Effect of Ancestry on Immigration

	(1)		(2)	
	Immigration 1990-2000		*Immigration 1980-1990*	
Log Ancestry 1990	9.4796*	9.1846**		
	(5.3527)	(4.5328)		
Log Ancestry 1980			5.3511*	5.4062*
			(3.1992)	(2.9951)
$I^{2000}_{o,-r(d)} \frac{I^{2000}_{-c(o),d}}{I^{2000}_{-c(o)}}$		0.7409***		
		(0.1264)		
$I^{1990}_{o,-r(d)} \frac{I^{1990}_{-c(o),d}}{I^{1990}_{-c(o)}}$				2.4135*
				(1.2713)
N	612495	612495	612495	612495
F Stat on excluded IVs	11.771	9.353	14.278	12.977

Notes: The table presents the coefficient estimates from IV regressions of equation (6) at the country-county level. The dependent variable is the immigration flow from 1990 to 2000 in clolumn 1 and the immigration flow from 1980 to 1990 in column 2. For all columns, we include $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_{t=1880..1980}$ as excluded instruments. All specifications control for log distance, latitude difference, origin-region, and destination-continent fixed effects. The coefficient estimates on these are not reported in the interest of space. Standard errors are given in parentheses. Standard errors are clustered at the country level. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Table 6: The Effect of Ancestry on the Intensive Margin of FDI

	(1)	(2)	(3)	(4)
	OLS	IV	IV	IV
Panel A	*Log Total # of FDI relationships*			
Log Ancestry 2010	0.256***	0.501***	0.350***	0.153***
	(0.051)	(0.047)	(0.022)	(0.030)
R^2	0.7706			
N	10851	10851	10851	10851
Panel B	*Log # of subsidiaries in destination*			
Log Ancestry 2010	0.282***	0.515***	0.315***	0.249***
	(0.053)	(0.071)	(0.015)	(0.043)
R^2	0.7649			
N	9082	9082	9082	9082
Panel C	*Log # of parents in origin*			
Log Ancestry 2010	0.272***	0.519***	0.325***	0.247***
	(0.056)	(0.086)	(0.016)	(0.041)
R^2	0.7651			
N	9082	9082	9082	9082
Panel D	*Log # of workers employed at subsidiaries in destination*			
Log Ancestry 2010	0.607***	1.419***	0.734***	0.404*
	(0.155)	(0.182)	(0.107)	(0.226)
R^2	0.6931			
N	9082	9082	9082	9082
Panel E	*Log # of subsidiaries in origin*			
Log Ancestry 2010	0.104**	0.402***	0.244***	-0.051
	(0.042)	(0.045)	(0.024)	(0.044)
R^2	0.7375			
N	4065	4065	4065	4065
Panel F	*Log # of parents in destination*			
Log Ancestry 2010	0.126***	0.484***	0.267***	-0.045
	(0.039)	(0.033)	(0.019)	(0.031)
R^2	0.7614			
N	4065	4065	4065	4065
Panel G	*Log # of workers employed at subsidiaries in origin*			
Log Ancestry 2010	0.219	0.839***	0.450***	-0.179
	(0.161)	(0.081)	(0.058)	(0.175)
R^2	0.6661			
N	4065	4065	4065	4065
Origin × Census Region FE	Yes	Yes	No	No
Destination × Continent FE	Yes	Yes	No	No
Heckman Correction	No	No	No	Yes

Notes: The table presents the OLS (column1) and IV (columns 2-4) estimates of equation (7). The dependent variables are specified for each panel in the table. The main variable of interest is *Log Ancestry 2010*. All IV columns use as instruments the same set of variables as column 3 of Table 3. All specifications control for log distance, latitude difference, origin, and destination fixed effects. The coefficient estimates on these specifications are not reported in the interest of space. Standard errors are given in parentheses. Standard errors are clustered at the country level. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

TABLE 7: THE EFFECT OF ANCESTRY ON THE INTENSIVE MARGIN OF TRADE (STATE LEVEL)

	(1)	(2)	(3)	(4)
	OLS		IV	
Panel A	*Log Total # of FDI relationships*			
Log Ancestry 2010	0.195***	0.943***	0.210**	0.086*
	(0.036)	(0.061)	(0.086)	(0.047)
R^2	0.8490			
N	2384	2384	2384	2199
Panel B	*Aggregate Export*			
Log Ancestry 2010	0.080**	1.054***	-0.206***	-0.163
	(0.034)	(0.089)	(0.053)	(0.111)
R^2	0.8373			
N	7904	7904	7904	4751
Panel C	*Aggregate Import*			
Log Ancestry 2010	0.292***	1.350***	-0.319***	-0.088
	(0.053)	(0.109)	(0.084)	(0.154)
R^2	0.7746			
N	6210	6210	6210	3831
Panel D	*Export To Vietnam*			
Log Ancestry 2010	1.163***	1.204***		
	(0.121)	(0.120)		
R^2	0.6745			
N	51	51		
Panel E	*Export To Japan*			
Log Ancestry 2010	0.869***	1.076***		
	(0.199)	(0.125)		
R^2	0.4262			
N	51	51		
Destination	Yes	No	Yes	Yes
Heckman Correction	No	No	No	Yes

Notes: The table presents the OLS (column 1) and IV (columns 2-4) estimates of equation (7) at the state level for FDI and trade. The main variable of interest is *Log Ancestry 2010*. The dependent variables are # of FDI links, aggregate exports, aggregate imports, exports to Vietnam, and exports to Japan in panels A, B, C, D, and E, respectively. The dependent variables in all other panels have the same as in Table 6. For all columns, we use $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_{t=1880..2000}$ and principal components as excluded instruments. All specifications control for log distance, latitude difference, and origin fixed effects. Standard errors are given in parentheses. Standard errors are clustered at the country level. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Table 8: Counterfactual Experiment: A Gold Rush in Los Angeles in 1880

	(1)	(2)	(3)	(4)	(5)
			Predicted Counterfactual Change		
Country Name	Ancestry 2010	FDI #	*FDI # (in %, IV)*	*FDI # (in %, RF)*	*Ancestry 2010*
Germany	365450	241	+102.83	+95.95	+98086
Ireland	265444	40	+95.32	+89.03	+92849
UK	441787	582	+33.69	+31.81	+40274
Norway	43718	55	+5.12	+4.87	+6930
Sweden	56442	71	+4.42	+4.20	+6001
France	81055	278	+3.80	+3.61	+5178
Canada	29862	531	+2.84	+2.70	+3880
Switzerland	11282	162	+2.71	+2.57	+3702
Czechoslovakia	19661	4	+2.34	+2.23	+3212
Netherlands	41841	121	+1.80	+1.71	+2473

Notes: The table presents the number of individuals of select ancestries living in Los Angeles County (column 1), the number of FDI links between Los Angeles County and the countries of origin (column 2), and the predicted changes in these variables under a counterfactual scenario where the pre-1880 pull factor of Los Angeles is 5 times as large the true size (columns 3 through 5). Column 3 shows the predicted change of *Total # of FDI relationships* (in percent) based on the IV regression of *Log Total # of FDI relationships* on *Log Ancestry 2010*, instrumented for by $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_{t=1880..2000}$. Column 4 shows the predicted change of *Total # of FDI relationships* (in percent) based on the reduced-form regression corresponding to the IV regression in column 4. Column 5 shows the predicted absolute change in ancestry, based on the same regression as column 4 but with *Ancestry 2010* as dependent variable. All three regressions control for log distance and latitude difference and include a origin $\times$ census region, and destination $\times$ continent fixed effects. Only the 10 countries with the highest absolute change in ancestry are shown in the interest of space. The details for the construction of this counterfactual are presented in section 4.4.

Table 9: The effect of Ancestry versus Immigration on FDI

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	FDI 2014 (Dummy)						
	IV	IV	OLS	IV	IV	IV	IV
Log Ancestry 2010	0.195***		0.146***	0.277***		0.184***	
	(0.011)		(0.019)	(0.031)		(0.015)	
Log Foreign-born 2010		0.257***	0.033	-0.135**			
		(0.023)	(0.044)	(0.056)			
Log Foreign-born 1970					0.523***	0.052	
					(0.045)	(0.053)	
Log Ancestry 2000							0.285***
							(0.029)
Log Foreign-born 2000							-0.149***
							(0.058)
N	612495	612495	612495	612495	612495	612495	612495

Notes: The table presents the OLS (column 3) and IV (all other columns) estimates of equation (1), contrasting the effect of ancestry and immigration on FDI. The dependent variable is the dummy of FDI in 2014. All IV columns use as instruments the same set of variables as column 3 of Table 3. All specifications control for log distance, latitude difference, origin-region, and destination-continent fixed effects. The coefficient estimates on these control variables are not reported in the interest of space. Standard errors are given in parentheses. Standard errors are clustered at the country level. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. For column 4, the Kleibergen-Paap statistic on the excluded instruments is 21.939 with p-value 0.056. For column 5, the Kleibergen-Paap statistic on the excluded instruments is 21.851 with p-value 0.058.

Table 10: Heterogeneous Effects across Countries

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A			*FDI 2014 (Dummy)*			
Log Ancestry × Geographic Distance	0.118***	0.117***	0.134**			0.199**
	(0.036)	(0.031)	(0.063)			(0.079)
Log Ancestry × Genetic Distance	-0.812					
	(1.217)					
Log Ancestry × Linguistic Distance		-0.343				
		(0.515)				
Log Ancestry × Religious Distance			-0.710			
			(0.446)			
Log Ancestry × Judicial Quality				0.147*		0.379*
				(0.089)		(0.202)
Log Ancestry × Fractionalization					-0.227**	0.545
					(0.092)	(0.379)
N	486855	414612	411471	452304	508842	446022
Panel B			*Log Total # of FDI relationships*			
Log Ancestry × Geographic Distance	0.445***	0.607***	0.549**			0.848***
	(0.086)	(0.091)	(0.230)			(0.194)
Log Ancestry × Genetic Distance	-6.133					
	(4.391)					
Log Ancestry × Linguistic Distance		-3.562**				
		(1.402)				
Log Ancestry × Religious Distance			-1.542			
			(1.600)			
Log Ancestry × Judicial Quality				1.335***		2.111***
				(0.237)		(0.464)
Log Ancestry × Fractionalization					-1.639***	2.765***
					(0.352)	(0.931)
N	10034	9407	9283	10150	10231	10149

Notes: The table presents coefficient estimates from IV regressions at the country-county level. The dependent variable for Panel A is the dummy of FDI in 2014. The dependent variable for Panel B is the log of the number of FDI links in 2014. We use $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_{t=1880..2000}$ and principal components as IVs. All specifications control for log distance, latitude difference, origin, and destination fixed effects. Standard errors are given in parentheses. Standard errors are clustered at the country level. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Table 11: Heterogeneous Effects across Counties

	(1)	(2)	(3)
Panel A	*FDI 2014 (Dummy)*		
Log Ancestry × Foreign Share	-0.210**		0.067
	(0.085)		(0.143)
Log Ancestry × Ethnic Diversity		1.373***	1.544***
		(0.241)	(0.429)
N	611910	612495	611910
Panel B	*Log Total # of FDI relationships*		
Log Ancestry × Foreign Share	-0.957**		-0.618
	(0.481)		(0.435)
Log Ancestry × Ethnic Diversity		4.895***	4.637***
		(1.115)	(0.874)
N	10851	10851	10851

Notes: The table presents coefficient estimates from IV regressions at the country-county level. The dependent variable for Panel A is the dummy of FDI in 2014. The dependent variable for Panel B is the log of the number of FDI links in 2014. We use $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_{t=1880..2000}$ and principal components as IVs. All specifications control for log distance, latitude difference, origin and destination fixed effects.Standard errors are given in parentheses. Standard errors are clustered at the country level. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

TABLE 12: HETEROGENEOUS EFFECTS ACROSS SECTORS AND FIRMS

	FDI 2014 (Dummy)		
	Log Ancestry 2010	N	FDI Dummy
Panel A: Individual Sectors			
Manufacturing	0.175***	5549	.009
	(0.029)		
Trade	0.172***	3212	.005
	(0.029)		
Information, Finance, Management, and other Services	0.150***	3628	.006
	(0.027)		
Construction, Real Estate, Accomodation, Recreation	0.134***	1637	.003
	(0.024)		
Health, Education, Utilities, and other Public Services	0.046***	689	.001
	(0.022)		
Natural Resources	0.036***	669	.001
	(0.010)		
Panel B: Final vs. Intermediate Goods			
Intermediate Goods	0.180***	5842	.01
	(0.028)		
Final Goods	0.170***	4201	.007
	(0.030)		
p-value of χ^2 test, H_0: equality of coefficients	0.000		
Panel C: Small vs. Large Firm Size			
Above Median	0.119***	1840	.003
	(0.021)		
Below Median	0.058***	723	.001
	(0.027)		
p-value of χ^2 test, H_0: equality of coefficients	0.000		

Notes: The table presents coefficient estimates on *Log Ancestry 2010* from IV regressions at the country-county level for each of the five sector groups (panel A), for firms producing final goods versus intermediate inputs (panel B), and for small- versus large-size firms (panel C). The composition of sector groups in panel A is given in Appendix Table 1. Final goods and intermediate inputs are defined as 4-digit NAICS sectors with upstreamness index below and above 2, respectively, where we use the upstreamness index from Antràs et al. (2012). The cutoff value between small and big firms is the median employee number, which is 1380 for subsidiaries and 1057 for parents. The dependent variable is the dummy of FDI in 2014. We use $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_{t=1880..2000}$ and principal components as intrumental variables. "N" refers the number of country-county pairs that have an FDI link in the corresponding sector. "FDI Dummy" refers to the mean of the FDI dummy in the corresponding sector group. All specifications control for log distance, latitude difference, origin-region, and destination-continent fixed effects. Standard errors are given in parentheses. Standard errors are clustered at the country level. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Table 13: Spillovers Effects

	(1)	(2)	(3)	(4)
Panel A: IV	*FDI 2014 (Dummy)*			
Log Ancestry 2010	0.179***	0.180***	0.179***	0.168***
	(0.034)	(0.028)	(0.028)	(0.028)
Log Ancestry 2010, State Level	-0.029**			
	(0.012)			
Log Ancestry 2010 of Nearest County within State		-0.022		
		(0.023)		
Log Ancestry 2010, Continent Level			0.020	
			(0.015)	
Log Ancestry 2010 of Nearest Origin				-0.070***
				(0.022)
N	612495	612495	612495	612495
Panel B: IV	*Log Total # of FDI relationships*			
Log Ancestry 2010	0.215**	0.091	0.166***	0.125***
	(0.099)	(0.074)	(0.048)	(0.040)
Log Ancestry 2010, State Level	-0.179			
	(0.125)			
Log Ancestry 2010 of Nearest County within State		0.042		
		(0.054)		
Log Ancestry 2010, Continent Level			-0.100*	
			(0.059)	
Log Ancestry 2010 of Nearest Origin				-0.046
				(0.064)
N	10851	10851	10851	10851

Notes: The table presents coefficient estimates from the extensive-margin equation (1) (Panel A) and the intensive-margin equation (7) (Panel B) at the country-county level. The dependent variable for Panel A is the dummy of FDI in 2014. The dependent variable for Panel B is the log of total FDI in 2014. For all columns, we use $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_{t=1880..2000}$ and principal components as IVs. All specifications control for log distance, latitude difference, origin-region, and destination fixed effects. Standard errors are given in parentheses. Robust standard errors are calculated to account for potential heteroskedasticity. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Online Appendix

"Migrants, Ancestors, and Investment"

Konrad B. Burchardi

Thomas Chaney

Tarek A. Hassan

A Data Appendix

A.1 Details on the construction of migration and ethnicity data

To construct the migration and ancestry data up until the year 2000, we downloaded the 1880, 1900, 1910, 1920, 1930, 1970, 1980, and 2000 waves of the Integrated Public Use Microdata Series (IPUMS) from https://usa.ipums.org/usa-action/samples. For each wave, we selected the largest available sample; for example, if a 1% and 10% sample was available for 1880 data, we used the 10% sample. To construct the 2010 data, we used the 2006-2010 American Community Survey (ACS) sample provided on the IPUMS website.

For each sample, we obtained the following variables: year, datanum, serial, hhwt, region, statefip, county, cntygp97, cntygp98, puma, gq, pernum, perwt, bpl, mbpl, fbpl, nativity, ancestr1, yrimmig, mtongue, mmtongue, fmtongue, and language.

We constructed the number of migrants from origin country o to destination county d in t, $I^t_{o,d}$, as well as the measure of ancestry $A^t_{o,d}$ from 1980 onward. We first aggregated the individual-level census data to counts of respondents at the level of historic US counties (or country groups from 1970 onwards) and foreign countries, and then transformed the data into 1990 country-county level using various transition matrices. Details are given in the following sections.

Transition matrices

The aim was to create transition matrices that would help us transform non-1990 countries to 1990 countries and non-1990 counties to 1990 counties/county groups, respectively.

- Birthplace-to-country: The aim was to construct transition matrices to map all the birthplace answers into 1990 countries. In each wave of the US Census, respondents were asked to report their country of birth. All possible answers (across time) are listed here: https://usa.ipums.org/usa-action/variables/BPL#codes_section. The censuses from 1850-2012 contain roughly 550 possible different answers to the question of birthplace. In each census data set, they are saved in the variable "bpld." What follows is our procedure for building those matrices:

  1. We started with a transition matrix of zeros, with all possible answers to the 1990 birthplace question as rows and all 1990 countries as columns. A cell in row r and column c of the transition matrix answers the question, "What is the probability that an individual who claims his/her birthplace as r refers to the area that in 1990 is country c?" So all cells contain values in [0,1], and rows sum up to 1.

2. For each row r in the transition matrix, if r with certainty refers to the area that in 1990 is country c, we simply changed the entry in cell (r,c) from 0 to 1 and moved to next row; if r does refer to an area that in 1990 is in multiple countries, then we searched for the 1990 population of each possible country, and assigned probabilities in proportion to the population data. We could usually find the population information on the Worldbank database.

- Ancestry-to-country:The aim was to construct transition matrices to map all the answers to the ancestry question into 1990 countries. The 1980, 1990, 2000, and 2010 census data provide information on the ancestry (ancestr1, 3-digit version). All possible answers (across time) are listed here: https://usa.ipums.org/usa-action/variables/ANCESTR1/#codes_section. The procedure was the same as in the birthplace-to-country part.

- Group-to-county & puma-to-county: The aim was to construct transition matrices to map all the county groups/pumas into counties. For the years 1970, 1980, and 1990, the US census data are at the historic US county group level. A "county group" is an agglomeration of US counties. For the years 2000 and 2010, the census data are at the puma level. A "puma" is also an agglomeration of US counties and is state dependent. To construct transition matrices from county agglomeration level to county level, we downloaded the corresponding matching files from the IPUMS website. We used data on the population of each county (within each county group/puma) to assign a probability that an observation from county group/puma g in year t is from county c in year t. This approach gives a transition matrix from year t county groups to year t counties.

- County-to-county: The aim was to construct transition matrices to map all the non-1990 counties into 1990 ones. The list and boundaries of US counties changed over time. To merge non-1990 US county-level data with 1990 county-level data, we needed a transition matrix. One transition matrix exists for each census year (1880, 1900, 1910, 1920, 1930, 1970, 1980, 2000, 2010). Such a transition matrix has as rows all US counties, indexed c, in year t, and as columns all US counties, indexed m, in year 1990. Each cell of the transition matrix takes a value that answers the question, "Which fraction of the area of the county c in year t is in 1990 part of county m?" A step-by-step tutorial for building those matrices follows:

  1. Download the year-specific map files.For 1880 us counties, obtain maps from Atlas: http://publications.newberry.org/ahcbp/downloads/united_states.html. Download the 503MB GIS file and find out the 1880 part. For 1900, 1910, 1920, and 1930 counties, obtain maps from IPUMS: https://usa.ipums.org/usa/volii/ICPSR.shtml. For 1970, 1980, and 1990 counties, obtain maps from NHGIS: https://data2.nhgis.org/main.
  2. Project non-1990 maps onto 1990 ones. We used the intersect command in ArcGIS to map year-specific counties onto 1990 counties based on areas. This approach gives a transition matrix from non-1990 counties to 1990 counties.

Post-1880 flow of immigrants

For each census wave after 1880, we counted the number of individuals in each historic US county d with historic foregin country o as birthplace (as identified by birthplace variable "bpld" in the raw data) that had immigrated to the United States since the last census that contains the immigration variable (not necessarily 10 years earlier). Then we transformed these data

- from the non-1990 foreign-country ("bpld") level to the 1990 foreign-country level using bpld-to-country transition matrices.

- from the US-county group/puma level to the US-county level using group/puma-to-county transition matrices.
- from the non-1990 US-county level to the 1990 US-county level using county-to-county transition matrices.
- from the post-1990 US-county level to the 1990 US county level based on the information from https://www.census.gov/geo/reference/county-changes.html.

Pre-1880 stock of immigrants

For the year 1880, we calculated for each historic US county d the number of individuals who have a historic foreign country o (no matter when they immigrated). We added to those calculations the number of individuals in county d who were born in the United States, but whose parents were born in historic foreign country o. (If the parents were born in different countries, we counted the person as half a person from the mother's place of birth, and half a person from the father's place of birth). Then we transformed these data

- from the pre-1880 foreign-country ("bpld") level to the 1990 foreign-country level using the pre-1880 country-to-country transition matrix.
- from the pre-1880 US-county level to the 1990 US-county level using the pre-1880 county-to-county transition matrix.

Stock of ancestry (1980, 1990, 2000, and 2010)

For the years 1980, 1990, 2000, and 2010, we calculated for each US county group the number of individuals who state as primary ancestry ("ancestr1" variable) some nationality/area. We transformed the data

- from the ancestry-answer ("ancestr1") level to the 1990 foreign-country level using ancestry-to-country transition matrices.
- from the US-county group/puma level to the US county-level using group/puma-to-county transition matrices.
- from the non-1990 US-county level to the 1990 US-county level using county-to-county transition matrices.
- from the post-1990 US-county to the 1990 US-county level based on the information from https://www.census.gov/geo/reference/county-changes.html.

A.2 Details on the construction of FDI data

We purchased the data from the Bureau van Dijk ORBIS. For each US firm, the raw data set lists the location of its (operational) headquarters, the addresses of its foreign parent entities, and the addresses of its international subsidiaries and branches. It also provides the number of employees for both US and foreign firms. The steps for building the data follow:

Clean up the postcode information

A firm's postcode is a unique identifier for the county location of the US firm, so we wanted to make sure one county corresponds to one postcode. If more than one country has the same postcode, we dropped the one with the smaller population (according to Google 2012 population data). We made the remaining duplicate postcodes for the same county unique by using random sampling. In the last step, we hand-coded missing postcodes that we took from main data set. Only one such case existed: 75427 for Dallas.

Build the parent data

The parent data contain 38 variables: "Mark" "Company name" "BvD ID number" "Ticker symbol" "Country ISO Code" "City" "Postcode" "Type of entity" "NAICS 2007 Core code (4 digits)" "NAICS, text description" "NACE Rev. 2 Core code (4 digits)" "Core NACE Rev. 2, text description" "Operating revenue (Turnover) th USD 2013" "Number of employees 2013" "Total assets th USD 2013" "Current market capitalisation th USD" "Listed/Delisted/Unlisted" "No of recorded shareholders" "No of recorded subsidiaries" "No of recorded branch locations" "Shareholder - Name" "Shareholder - Ticker symbol" "Book value per share USD" "Shareholder - BvD ID number" "Shareholder - City" "Shareholder - Postal code" "Shareholder - Telephone number" "Shareholder - Type" "Shareholder - NACE Rev. 2, Core code" "Shareholder - NACE Rev. 2, text description" "Shareholder - NAICS 2007, Core code" "Shareholder - NAICS 2007, text description" "Shareholder - Country ISO code" "Shareholder - Direct %" "Shareholder - Total %" "Shareholder - Operating revenue (Turnover) mil USD" "Shareholder - Number of employees" "Shareholder - Total assets mil USD." Here "shareholder" is equivalent to "parent" in our context, so we can use them interchangeably. The data-building steps are as follows:

1. Data clean-up:
    - Blank rows were filled in within the same company with more than one parents.
    - Numerate string variables: "Listed," "Unlisted," and "Delisted" received a numerical counterparty listed with 1, 0, -1, respectively.
    - Numerical variables in string were adjusted to match Stata format, e.g., n.a., - into ., canceling string, inside numbers.
    - Assigned number to Shareholder Direct and Shareholder Total:
        - Mysterious values such as "MO" and "WO" were recovered from a paper using the same data set for a simlar topic: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2407845. It says, "When the stake of a shareholder is described by the following initials, we replace it with the appropriate number as follows: MO, majority owned, is replaced by '75%'; JO, jointly owned, is replaced by '50%'; NG, negligence, is replaced by '0%'; CQP1,—-, is replaced by '50.01%'; BR, branch, is replaced by '5.01%'; and if the holding of a shareholder is wholly owned (WO), then we delete the firm from the sample, as this firm should not be considered a publicly traded company."
        - Ambiguous numbers were made clear: for values with a ">" before, e.g., $> 25.00 > 30.00 > 50.00 > 75.00$, were replaced by the original number plus 10; for values with a "<" before, e.g., $< 34.00 < 50.00 < 75.00$, were replaced by the original number minus 10; for values with a "±" before, e.g. ±25.00 ±75.00, are replaced by the original number.
    - After adjustment, numerical variables in string were destringed.

2. Postcode matching: We matched both companies and their US parents (foreign parents were ignored in this step), with the postcode data just built. Besides the original string variable postcode, we generated new variables postcode5digit and postcodeextension and labeled them "Postal code (5 digit)" and "Postal code (extension)." Similarly, shareholders had shareholderpostcodeUS5digit and shareholderpostcodeUSextension (note the spelling postal code in shareholder variables was unified to postcode).

3. Country-code matching: We matched both companies and their parents. Each firm had four country variables: numerical country code, country name, and 2- and 3- digit ISO country code. Then we adjusted those modern country codes to 1990 codes based on the information on post-1990 country changes.

Build the subsidiary data

The subsidiary data contain 53 variables: "Mark" "Company name" "BvD ID number" "Ticker symbol" "Country ISO Code" "City" "Postcode" "Type of entity" "NACE Rev. 2 Core code (4 digits)" "Core NACE Rev. 2, text description" "NAICS 2007 Core code (4 digits)" "NAICS, text description" "Operating revenue (Turnover) th USD 2007" "Total assets th USD 2007" "Number of employees 2007" "Current market capitalisation th USD" "Listed/Delisted/Unlisted" "No of recorded branch locations" "No of recorded subsidiaries" "No of recorded shareholders" "Book value per share USD" "Subsidiary - Name" "Subsidiary - BvD ID number" "Subsidiary - Ticker symbol" "Subsidiary - Country ISO code" "Subsidiary - City" "Subsidiary - Postal code" "Subsidiary - Telephone number" "Subsidiary - Type" "Subsidiary - NACE Rev. 2, Core code" "Subsidiary - NACE Rev. 2, text description" "Subsidiary - NAICS 2007, Core code" "Subsidiary - NAICS 2007, text description" "Subsidiary - Operating revenue (Turnover) mil USD" "Subsidiary - Total assets mil USD" "Subsidiary - Number of employees" "Subsidiary - Level" "Subsidiary - Direct %" "Subsidiary - Total%" "Subsidiary - Status" "Branch - Name" "Branch - BvD ID number" "Branch - Country ISO code" "Branch - City" "Branch - Postcode" "Branch - Telephone number" "Branch - NAICS 2007, Core code" "Branch - NACE Rev. 2, Core code" "Branch - Operating revenue (Turnover) mil USD" "Branch - Total assets mil USD" "Branch - Number employees" "Branch - NACE Rev. 2, text description" "Branch - NAICS 2007, text description." Most cleaning processes are similar to that of the parent data. The extra cleaning for subsidiary data follows:

1. Subsidiary status: Similarly to shareholder status, strings are coded in subsidiary status. From an online documentation of the data set, subsidiary status is UO+ if all links found in the path have a percentage over 97.99% or are "UO links" indicated by a source; *UO if all links found in the path have a percentage over 50%, but one or more links are between 50.01% and 97.99%; *UO- if all links have a percentage over 25%, but one or more links are between 25.01% and 50%. To prevent confusion in later analysis, we did not code these strings to the average percentage of their range, but to integers: UO-=1, UO=2, UO+=3 and missing values "-" to be 0.

2. We merged subsidiaries with branches using variables to denote the difference; that is, we renamed 13 branch variables to subsidiary* to match the 13 of the 19 subsidiary variables and reshape the data to move branch rows under subsidiaries rows. Thus, each observation is a subsidiary or branch of the firm. Denoting variables are issubsidiary (=1 if subsidiary and =0 if branch), companyBranchcount (count of branches under each firm), companySubsidiarycount (count of subsidiaries under each firm), and companyRowcount (count of both branches and subsidiaries under each firm).

A.3 Details on the construction of our other data

International trade.— The data on trade between US states and foreign countries, both at the aggregate level and at the sectoral level, are from the Commodity Flow Survey for the year 2012. The data are collected by the US Census Bureau. A representative sample of establishments are surveyed every five years, and information on their shipments collected. The value of all shipments crossing the US international border are recorded as international trade, along with their foreign origin/destination country. We only used thes readily available data aggregated at the US state and foreign country level. Although they do not cover all of the US foreign trade (the data com from a representative survey, not from the universe of foreign transactions), they are the only publicly available source of international data disaggregated at a geographic level below that of the entire United States. For each origin country and destination state, $Import_{o,d}$ are aggregate imports (in dollars) from country o to US state d in 2012, and $Export_{o,d}$ are aggregate exports (in dollars) from US state d to country o in 2012, where we keep the convention of using o for foreign countries and d for US administrative units, states or counties.

Bilateral distances and latitude differences.— To compute the distance between US counties or states and foreign countries, we used the coordinates for all postal codes within a county or state, and the coordinates of the main city for foreign countries.[24] We define the latitude and longitude of a US county as the unweighted average of the latitudes and longitudes of all postal codes within the county. We define the latitude and longitude of a US state as the unweighted average of the latitude and longitude of all counties within the state. The distance between foreign country o and a US county or state d, $Distance_{o,d}$, is computed as the great circle distance between the two, measured in kms. The latitude difference between a foreign country o and a US county or state d, $Latitude\,Difference_{o,d}$, is the absolute difference between the latitudes of the two, measured in degrees.

Country characteristics.— To shed light on the mechanism through which the presence of foreign ancestry affects the patterns for foreign investment, we constructed several measures of foreign country and US county characteristics. $Genetic\,Distance_o$ is a measure of the genetic distance between foreign country o and the United States, normalized to take values between 0 and 1. $Linguistic\,Distance_o$ is a measure of the linguistic distance between foreign country o and the United States; it measures the probability that a randomly selected person in the United States speaks the same language as a randomly selected person from country o. $Religious\,Distance_o$ measures the religious distance between foreign country o and the United States, with a similar construction as the linguistic distance.[25] A higher index for $GeneticDistance_o$, $LinguisticDistance_o$, or $ReligiousDistance_o$ corresponds to a greater distance between the United States and country o. $Judicial\,Quality_o$ is a measure of the judicial quality in country o.[26] A higher index for $JudicialQuality_o$ corresponds to a higher-quality judicial system. $Fractionalization_o$ is a measure of the ethnolinguistic fractionalization of country o.[27]

US county characteristics.— We define three US-county level measures. $Ethnic\,Diversity_d$ is a measure of the diversity of ethnic groups in county d.[28] $Foreign\,Share_d$ measures the share of residents in county d who claim foreign ancestry. $Population_d$ is the population size of US county d. $IncomePerCapita_d$ is a measure of the per-capita income (in thousand dollars) within US county d. Both population and income data are taken from the Bureau of Economic Analysis.

[24] The geo-coordinates are downloaded from `www.geonames.org`and `www.cepii.fr`, respectively.

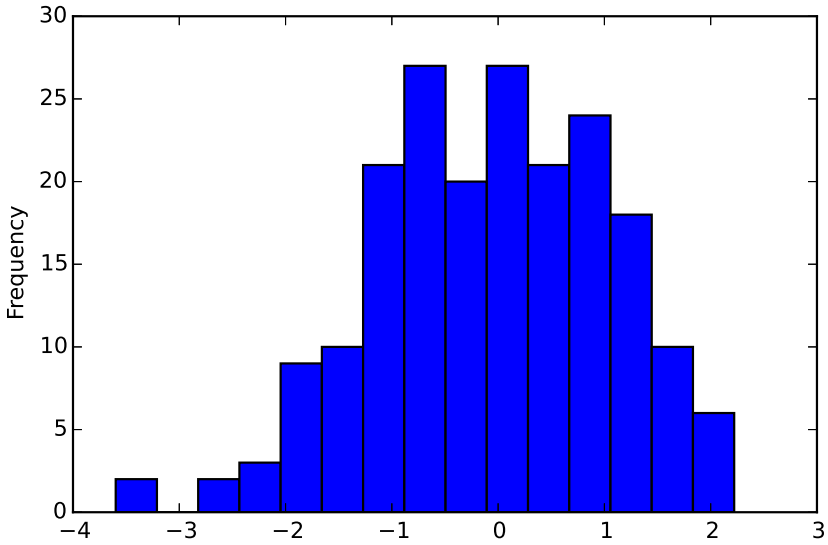
[25] Both genetic and religious distance measures come from Spolaore and Wacziarg (2015).

[26] The measure of judicial quality comes from Kaufmann et al. (2003) and is used in Nunn (2007). It is based on a weighted average of variables measuring perceptions of the effectiveness of the judiciary and the enforcement of contracts.

[27] The measure of fractionalization comes from Alesina et al. (2003). It is equal to 1 minus the Herfindahl index of ethnolinguistic group shares.

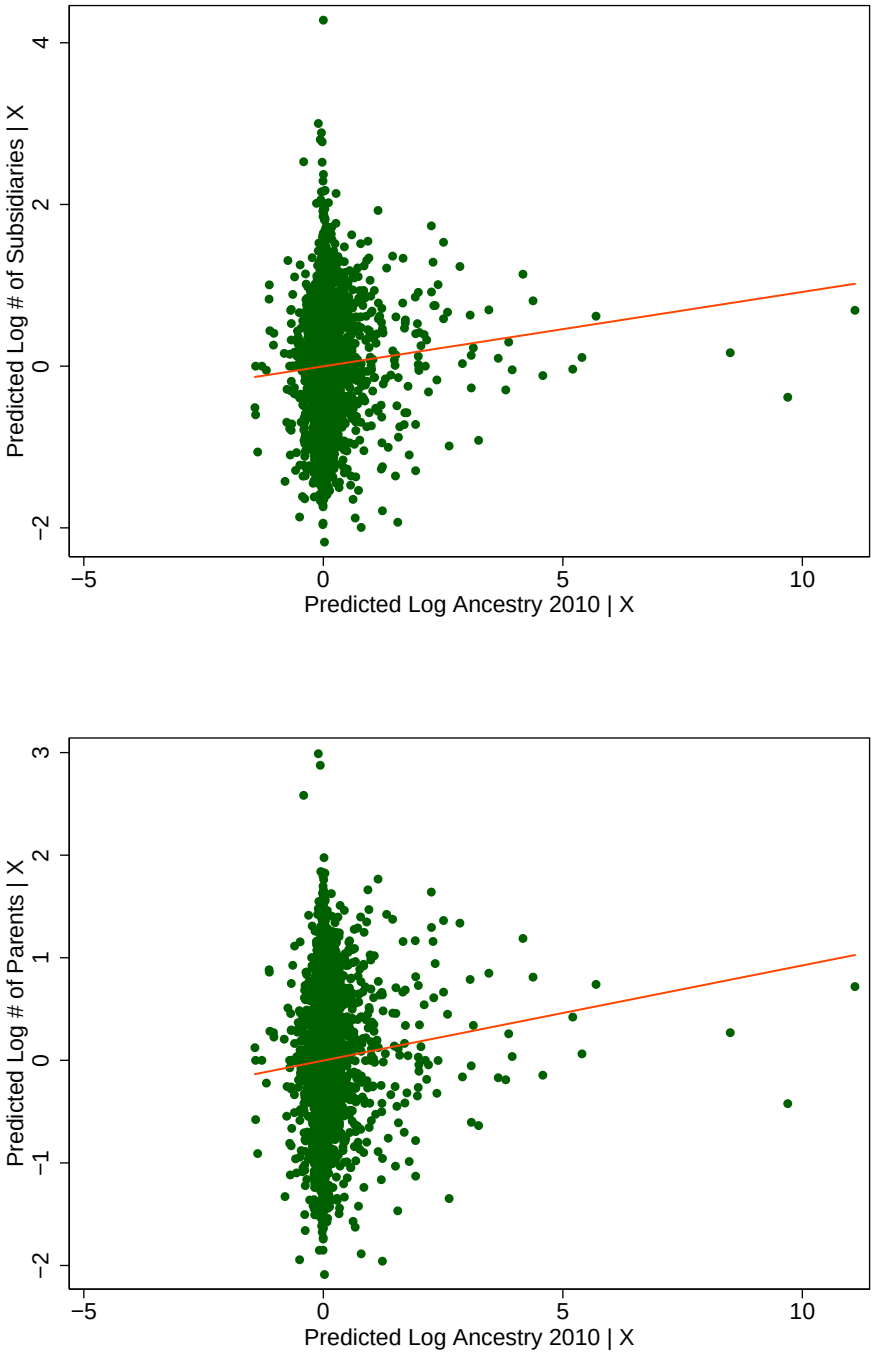
[28] It is equal to 1 minus the Herfindahl index of ancestry, measured as the sum of squared fractions of all possible ancestry among people who report foreign ancestry within US county d.

Sectoral characteristics.— We separated sectors into final consumption goods and intermediate inputs. To do so, we use the measure of upstreamness from Antràs et al. (2012). We classified 4-digit NAICS sectors as "final goods" if their upstreamness index is below 2, and as "intermediates" if their upstreamness index is above 2.



Appendix Figure 1: Placebo Experiment: Histogram of t-stats

Notes: The figure presents a placebo experiment as an extension to Table 4 Panel B. Each time, we assign each country to some random country on a different continent, run the same specification as in Table 4 Panel B Column 3, and report the t-statistic on the estimated coefficient of *Log Ancestry 2010*. We repeat the procedure 200 times and generate the histogram. Using ± 1.96 as cut-off values, the false positive rate is 2.5% and the false negative rate is 4.5%.



Appendix Figure 2: Ancestry and the Intensive Margin of FDI

Notes: The figure is a conditional scatterplot of log 2010 ancestry against log (FDI volume). The upper part depicts the log of the total number of domestic subsidiaries of foreign firms plus foreign subsidiaries of domestic firms. The lower part depicts the log of the total number of domestic parents of foreign firms plus foreign parents of domestic firms. The corresponding regression uses the same specification as in Table 6, column 3. The solid line depicts the fitted regression line. The slope of the upper part line is 0.092 with standard error 0.017. The slope of the lower part line is 0.093 with standard error 0.016.

Appendix Table 1: Composition of Sector Groups

Group	Sectors	# of US Firms
Manufacturing	Manufacturing	10009
Trade	Wholesale Trade Retail Trade	7191
Information, Finance, Management, and Other Services	Information Finance and Insurance Professional, Scientific, and Technical Services Management of Companies and Enterprises Administrative and Support and Waste Management and Remediation Services Other Services (Except Public Administration)	10052
Construction, Real Estate, Accomodation, Recreation	Construction Transportation and Warehousing Real Estate and Rental and Leasing Arts, Entertainment, and Recreation Accommodation and Food Services	3039
Health, Education, Utilities, and Other Public Services	Utilities Educational Services Health Care and Social Assistance	1257
Natural Resources	Agriculture, Forestry, Fishing and Hunting Mining, Quarrying, and Oil and Gas Extraction	871

Appendix Table 2: Assignment of States to Census Regions

Census Region	State Names
New England	Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Vermont
Middle Atlantic	New Jersey, New York, Pennsylvania
East North Central	Illinois, Indiana, Michigan, Ohio, Wisconsin
West North Central	Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota, South Dakota
South Atlantic	Delaware, District Of Columbia, Florida, Georgia, Maryland, North Carolina, South Carolina, Virginia, West Virginia
East South Central	Alabama, Kentucky, Mississippi, Tennessee
West South Central	Arkansas, Louisiana, Oklahoma, Texas
Mountain	Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, Wyoming
Pacific	Alaska, California, Hawaii, Oregon, Washington

Appendix Table 3: Summary Statistics on the Intensive Margin of FDI

Origin-destination pairs	(1)	(2)	(3)
Ancestry 2010 (in thousands)	9.796	10.729	15.645
	(36.739)	(39.193)	(55.681)
# of Total FDI	11.057		
	(39.766)		
# of Subsidiaries in Destination		4.572	
		(14.951)	
# of Parents in Origin		5.354	
		(17.975)	
# of Employees at Subsidiary in Destination (in thousands)		11.872	
		(44.522)	
# of Subsidiaries in Origin			5.018
			(15.739)
# of Parents in Destination			2.318
			(4.431)
# of Employees at Subsidiary in Origin (in thousands)			5.812
			(60.380)
N	10851	9082	4065

Appendix Table 4: Placebo Regressions

	(1)	(2)	(3)	(4)	(5)	(6)
	FDI 2014 (Dummy)					
Panel A	*Assign to alphabet neighbor*					
Log Ancestry 2010	-0.012	-0.006	0.014	0.014	0.000	0.017
	(0.021)	(0.015)	(0.031)	(0.031)	(0.032)	(0.035)
N	612495	612495	612495	612495	612495	612300
Panel B	*Assign to alphabet neighbor on a different continent*					
Log Ancestry 2010	-0.031	-0.022	0.012	0.012	-0.005	0.016
	(0.022)	(0.014)	(0.036)	(0.036)	(0.035)	(0.041)
N	612495	612495	612495	612495	612495	612300
Principal Components	No	Yes	Yes	Yes	Yes	Yes
$I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})$	No	No	No	No	Yes	No
Destination × Continent FE	No	No	Yes	Yes	Yes	Yes
Origin × Census Region FE	No	No	Yes	Yes	Yes	No
Origin × State FE	No	No	No	No	No	Yes
3rd order poly in dist and lat	No	No	No	Yes	No	No

Notes: The table presents coefficient estimates from two placebo regressions at the country-county level. In Panel A, we assign each country to its alphabet neighbor. In Panel B, we assign each country to its alphabet neighbor on a different continent. The dependent variable in all panels is the dummy of FDI in 2014. For all columns, we include $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_{t=1880..2000}$ as excluded instruments. All specifications control for log distance, latitude difference, origin, and destination fixed effects. Standard errors are given in parentheses. Standard errors are clustered at the country level. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Appendix Table 5: Alternative Standard Error Specifications

	(1)	(2)	(3)
		FDI Dummy (2014)	
Panel A: Bootstrap	Raw	Country Panel	County Panel
Log 2010 Ancestry	0.1970***	0.1970***	0.1970***
	(0.0104)	(0.0350)	(0.0188)
N	612495	612495	612495
Panel B	Robust	County Cluster	State Cluster
Log 2010 Ancestry	0.1970***	0.1970***	0.1970***
	(0.0101)	(0.0200)	(0.0221)
N	612495	612495	612495
Panel C	Country Cluster	State-Continent Cluster	State-Country Cluster
Log 2010 Ancestry	0.1970***	0.1970***	0.1970***
	(0.0300)	(0.0151)	(0.0133)
N	612495	612495	612495

Notes: The table presents results from experiments of alternative standard errors. Panel A runs a bootstrap experiment with 50 repetitions in each column. Panel A, column 1 runs bootstrap on the pooled data. Panel A, column 2 runs bootstrap but keeps the country as panel (so it randomly draws 200 country panels with replacement). Panel A column 3 runs bootstrap but keeps the county as panel (so it randomly draws 3,000 county panels with replacement). Panels B and C present results from IV regressions. The specification is the same as that in Table 3, column 3. Standard errors are given in parentheses. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Appendix Table 6: The Effect of Ancestry on FDI: 5 Largest Countries and Counties

	FDI 2014 (Dummy)
Panel A: Top 5 Ancestries	*Log Ancestry 2010*
Germany	0.200***
	(0.010)
Britain	0.265***
	(0.009)
Ireland	0.202***
	(0.010)
Mexico	0.174***
	(0.011)
Italy	0.212***
	(0.007)
Panel B: Largest 5 Counties	*Log Ancestry 2010*
Los Angeles,California	0.129***
	(0.019)
Cook,Illinois	0.124***
	(0.019)
Harris,Texas	0.161***
	(0.024)
San Diego,California	0.152***
	(0.023)
Orange,California	0.159***
	(0.020)

Notes: The table presents coefficient estimates from IV regressions at the country-county level. The dependent variable in both panels is the dummy of FDI in 2014. Panel A presents the coefficient on *Log Ancestry 2010* for each of the top five ancestries. Panel B presents the coefficient on *Log Ancestry 2010* for each of the largest five counties. We use $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_{t=1880..2000}$ and principal components as IVs. All specifications control for log distance and latitude difference. Standard errors are clustered at the country level. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Appendix Table 7: The Effect of Ancestry on FDI: Country Specific Effects

	FDI Dummy		
	Point Estimate	Standard Deviation	N
United Arab Emirates	14.971	2.972	60
Kuwait	7.777	2.298	22
Finland	3.919	0.482	180
New Zealand	2.898	0.500	107
Oman	2.815	1.857	6
British Virgin Islands	2.810	0.758	100
Australia	2.236	0.384	369
Malaysia	2.058	0.435	90
South Africa	1.804	0.261	80
Tunisia	1.664	0.481	9
Iceland	1.310	0.268	25
Saudi Arabia	1.257	0.178	29
Puerto Rico	1.118	0.265	26
Bahamas	1.047	0.339	44
Israel	0.978	0.146	137
Switzerland	0.766	0.043	371
BelgiumLuxembourg	0.688	0.041	354
Denmark	0.667	0.040	278
Uruguay	0.655	0.118	21
Thailand	0.609	0.084	68
Japan	0.568	0.053	575
Brazil	0.553	0.056	140
Chile	0.541	0.083	73
Panama	0.529	0.123	44
Austria	0.509	0.040	148
Liberia	0.508	0.264	6
Barbados	0.508	0.267	38
Costa Rica	0.492	0.176	30
Malta	0.476	0.273	11
Turkey	0.469	0.069	48
Norway	0.453	0.027	239
Indonesia	0.437	0.079	29
Senegal	0.436	0.362	2
Canada	0.435	0.023	809
Argentina	0.428	0.060	64
Netherlands	0.426	0.018	398
Sweden	0.398	0.018	323
Republic of Korea	0.367	0.026	155
Spain	0.352	0.015	300
Kenya	0.351	0.197	5
France	0.350	0.013	528
India	0.335	0.018	233
China	0.305	0.015	248
Venezuela (Bolivarian Republic of)	0.304	0.051	32
Belize	0.299	0.089	14
UK	0.265	0.009	664

Egypt	0.264	0.052	23
Colombia	0.263	0.030	45
Samoa	0.243	0.091	5
Peru	0.243	0.037	30
Hungary	0.228	0.032	52
Morocco	0.219	0.088	11
Nigeria	0.214	0.062	18
Italy	0.212	0.007	489
Portugal	0.207	0.028	85
Ireland	0.202	0.010	247
Germany	0.200	0.010	608
Romania	0.176	0.042	23
Czechoslovakia	0.175	0.029	54
Pakistan	0.175	0.041	23
Mexico	0.174	0.011	259
USSR	0.163	0.014	97
Philippines	0.158	0.020	50
Sri Lanka	0.157	0.107	6
Bulgaria	0.154	0.064	11
Ghana	0.151	0.094	6
Lebanon	0.149	0.046	20
Bolivia (Plurinational State of)	0.149	0.071	8
Trinidad and Tobago	0.146	0.074	15
Greece	0.130	0.027	42
Jamaica	0.126	0.036	15
Socialist Yugoslav	0.124	0.028	29
Honduras	0.120	0.040	14
Cameroon	0.118	0.094	2
Guatemala	0.105	0.038	14
Fiji	0.103	0.075	5
Nicaragua	0.102	0.052	7
Algeria	0.101	0.075	3
Viet Nam	0.099	0.028	18
Dominican Republic	0.097	0.029	16
Ecuador	0.096	0.039	15
Poland	0.088	0.014	63
Paraguay	0.087	0.064	4
Democratic People's Republic of Korea	0.081	0.088	1
Jordan	0.076	0.050	7
Sudan	0.075	0.075	1
El Salvador	0.074	0.027	13
Albania	0.071	0.043	3
Bangladesh	0.046	0.034	2
Cambodia	0.036	0.026	3
Haiti	0.028	0.020	2
Ethiopia	0.027	0.026	1
Syrian Arab Republic	0.023	0.023	1
Myanmar	0.014	0.015	1
Afghanistan	0.004	0.004	1

Guyana	0.003	0.003	1
Iraq	0.002	0.002	1
Cuba	-0.000	0.000	1
Libya	-0.020	0.021	1
Mauritania	NA	NA	0
Equatorial Guinea	NA	NA	0
Iran (Islamic Republic of)	NA	NA	0
Somalia	NA	NA	0
Mongolia	NA	NA	0
Greenland	NA	NA	0
Sierra Leone	NA	NA	0
State of Palestine	NA	NA	0
Cape Verde	NA	NA	0
Tonga	NA	NA	0
Lao People's Democratic Republic	NA	NA	0
Nepal	NA	NA	0
Yemen	NA	NA	0

Notes: The table is an extension of Table 6 Panel A, where we only show the results for top five ancestries. The fourth column presents the number of counties that have an FDI link with the corresponding country.

Appendix Table 8: The Effect of Ancestry on FDI: Sector-Specific Effects

	FDI Dummy		
	Point Estimate	Standard Deviation	N
	20 Sectors Based on 2007 NAICS code		
Manufacturing	17.520	2.850	5549
Wholesale Trade	15.600	3.053	2513
Professional, Scientific, and Technical Services	13.210	2.712	1925
Information	9.234	2.096	906
Retail Trade	9.012	2.249	846
Transportation and Warehousing	8.941	1.795	620
Administrative and Support and Waste Management and Remediation Services	8.869	2.052	855
Real Estate and Rental and Leasing	8.316	2.260	662
Finance and Insurance	7.399	2.187	1143
Other Services (except Public Administration)	5.664	1.594	301
Management of Companies and Enterprises	5.389	1.526	524
Construction	4.442	1.752	510
Accommodation and Food Services	3.400	1.011	239
Arts, Entertainment, and Recreation	3.034	0.677	131
Mining, Quarrying, and Oil and Gas Extraction	2.928	0.986	528
Health Care and Social Assistance	2.828	1.326	291
Utilities	2.461	1.407	338
Educational Services	0.899	0.551	111
Agriculture, Forestry, Fishing and Hunting	0.701	0.288	149
Public Administration	0.118	0.095	10

Notes: The table presents coefficient estimates on *Log Ancestry 2010* from IV regressions for each of the 20 sectors at the country-county level. The dependent variable is the dummy of FDI in 2014. We use $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_{t=1880..2000}$ and principal components as IVs. All specifications control for log distance, latitude difference, origin-region, and destination-continent fixed effects. Standard errors are given in parentheses. Standard errors are clustered at the country level. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Appendix Table 9: The Effect of Ancestry on FDI: Robustness

	Extensive Margin					
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A	$I_o * I_d/I$					
Log Ancestry 2010	0.209***	0.208***	0.179***	0.179***	0.181***	0.228***
	(0.020)	(0.020)	(0.027)	(0.027)	(0.031)	(0.019)
N	612495	612495	612495	612495	612495	612300
Panel B	$I_{o,-d} * I_{-o,d}/I_{-o}$					
Log Ancestry 2010	0.217***	0.207***	0.174***	0.174***	0.179***	0.224***
	(0.021)	(0.020)	(0.028)	(0.028)	(0.031)	(0.019)
N	612495	612495	612495	612495	612495	612300
Panel C	$I_{o,-d} * I_{-c(o),d}/I_{-c(o)}$					
Log Ancestry 2010	0.231***	0.225***	0.189***	0.188***	0.186***	0.240***
	(0.023)	(0.022)	(0.027)	(0.027)	(0.031)	(0.021)
N	612495	612495	612495	612495	612495	612300
Panel D	$I_{o,-adj(d)} * I_{-c(o),d}/I_{-c(o)}$					
Log Ancestry 2010	0.240***	0.209***	0.200***	0.200***	0.193***	0.248***
	(0.024)	(0.021)	(0.024)	(0.024)	(0.030)	(0.021)
N	640764	640764	640764	640764	640764	640560
Principal Components	No	Yes	Yes	Yes	Yes	Yes
$I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})$	No	No	No	No	Yes	No
Destination × Continent FE	No	No	Yes	Yes	Yes	Yes
Origin × Census Region FE	No	No	Yes	Yes	Yes	No
Origin × State FE	No	No	No	No	No	Yes
3rd order poly in dist and lat	No	No	No	Yes	No	No

Notes: The table presents coefficient estimates from IV (Panel A) and OLS (Panel B) regressions at the country-county level. The dependent variable in all panels is the dummy of FDI in 2014. The main variable of interest is *Log Ancestry 2010*. The excluded instruments are indicated by the title in each panel. In Panel D, "adj" refers to the adjacent states for the state of county d. All specifications control for log distance, latitude difference, origin, and destination fixed effects. Standard errors are given in parentheses. Standard errors are clustered at the country level. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Appendix Table 10: Non-parametric Estimation

	(1)	(2)	(3)
Panel A: OLS	*FDI 2014 (Dummy)*		
Ancestry 2010 Quantile 1	0.003	0.003	0.003
	(0.008)	(0.008)	(0.007)
Ancestry 2010 Quantile 2	0.019	0.012	0.012
	(0.018)	(0.013)	(0.011)
Ancestry 2010 Quantile 3	0.128***	0.046*	0.025
	(0.034)	(0.024)	(0.018)
Ancestry 2010 Quantile 4		0.174***	0.068**
		(0.039)	(0.027)
Ancestry 2010 Quantile 5			0.209***
			(0.042)
N	612495	612495	612495

Panel B	*Nonlinear Least Squares*	
	β	γ
Estimates	0.1574***	0.0011***
	(0.0011)	(0.0000)

Notes: The table presents coefficient estimates from OLS (Panel A) and nonlinear least squares (Panel B) at the country-county level. The dependent variable in both panels is the dummy of FDI in 2014. In column 1, the cutoff values are 144.89 and 654.91; in column 2, the cutoff values are 107.73, 281.58, and 1144.13; and in column 3, the cutoff values are 91.90, 186.68, 454.95, and 1659.56. All the numbers are in unit. Standard errors are given in parentheses. Standard errors are clustered at the country level. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. All coefficients and standard errors are multiplied by 100. For Panel B, we obtain the optimal β and γ by solving a nonlinear least squares problem as mentioned in the text.

Appendix Table 11: Alternative Functional Forms

	FDI 2014 (Dummy)					
	(1)	(2)	(3)	(4)	(5)	(6)
Ancestry 2010	0.003***					
	(0.001)					
Log Ancestry 2010 (-1 for $-\infty$)		0.197**				
		(0.098)				
$(\text{Ancestry 2010})^{1/3}$			0.203***			
			(0.026)			
Log Ancestry 1980				0.221***		
				(0.037)		
Log Ancestry 1990					0.215***	
					(0.034)	
Log Ancestry 2000						0.202***
						(0.030)
N	612495	612495	612495	612495	612495	612495

Notes: The table presents coefficient estimates from IV regressions at the country-county level. The dependent variable is the dummy of FDI in 2014. The main variable of interest is the various ancestry variables as indicated by the first column of the table. In the second row, we use Log(Ancestry/1000) instead of Log(1+Ancestry/1000), and replace Log(0) with -1. The specification is the same as that in Table 3, column 3. Standard errors are given in parentheses. Standard errors are clustered at the country level. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.