Genetic Diversity and the Origins of Cultural Fragmentation

Quamrul Ashraf and Oded Galor*

Online Appendix

A Variable Definitions and Sources

A.1 Dependent Variables

Number of ethnic groups [EG]. The total number of distinct ethnic groups in a country's population, as compiled by Fearon (2003). The cross-country variable employed by the empirical analysis is the natural logarithm of one plus the number of ethnic groups. See Fearon (2003) for additional details on primary data sources and methodological assumptions.

Ethnic fractionalization [EF-F/EF-A]. The probability that 2 randomly-selected individuals in a country's population belong to different ethnic groups. Formally, the ethnic fractionalization index for each country is calculated as:

$$FRAC = 1 - \sum_{i=1}^{n} p_i^2,$$

where p_i is the proportional representation of ethnic group i in the national population; and n is the total number of ethnic groups comprising the country's population. Data on ethnic groups (and their proportional representations in the national population) by country are compiled independently by Alesina et al. (2003) and Fearon (2003), thus yielding two separate (but correlated) cross-country measures of ethnic fractionalization. See Alesina et al. (2003) and Fearon (2003) for additional details on primary data sources and methodological assumptions.

Ethnolinguistic fractionalization, level-1 aggregation [ELF-D]. An index of fractionalization, constructed by Desmet, Ortuño-Ortín and Wacziarg (2012), across the *ancestral* categories of the modern linguistic groups in a country's population. The ancestral linguistic divisions in a country's population correspond to the branches that are closest to the "root" of the country-specific phylogenetic linguistic tree. To compute the fractionalization index, the proportional representations (in the national population) of the modern linguistic groups are first aggregated up into different bins, each corresponding to one of these proto-language branches of the linguistic tree. For each country, the index is then calculated across these bins (or ancestral linguistic groups) by applying the same equation as the one underlying the calculation of the ethnic fractionalization index. See Desmet,

*Ashraf: Williams College, Department of Economics, 24 Hopkins Hall Dr., Williamstown, MA 01267 (email: Quamrul.H.Ashraf@williams.edu). Galor: Brown University, Department of Economics, 64 Waterman St., Providence, RI 02912 (email: Oded_Galor@brown.edu).

Ortuño-Ortín and Wacziarg (2012) for additional details on primary data sources and methodological assumptions.

Ethnolinguistic polarization, level-1 aggregation [POL-D]. An index of polarization, constructed by Desmet, Ortuño-Ortín and Wacziarg (2012), across the *ancestral* categories of the modern linguistic groups in a country's population. The ancestral linguistic divisions in a country's population correspond to the branches that are closest to the "root" of the country-specific phylogenetic linguistic tree. To compute the polarization index, the proportional representations (in the national population) of the modern linguistic groups are first aggregated up into different bins, each corresponding to one of these proto-language branches of the linguistic tree. For each country, the index is then calculated across these bins (or ancestral linguistic groups) by applying the following definition of polarization due to Reynal-Querol (2002) and Montalvo and Reynal-Querol (2005):

$$POL = 4\sum_{i=1}^{n} p_i^2 \left[1 - p_i\right],$$

where p_i is the proportional representation of ancestral linguistic group i; and n is the total number of ancestral linguistic groups. See Desmet, Ortuño-Ortín and Wacziarg (2012) for additional details on primary data sources and methodological assumptions.

Ethnolinguistic polarization, Esteban-Ray index [POL-ER]. An index of polarization, constructed by Esteban, Mayoral and Ray (2012), across the ethnic groups in a country's population, where ethnic groups by country are classified according to Fearon (2003) and the definition of polarization that is applied is the one due to Duclos, Esteban and Ray (2004) and Esteban and Ray (2011) that incorporates intergroup distances. Formally, the polarization index for each country is calculated as:

$$POL = \sum_{i=1}^{n}\sum_{j=1}^{n} p_i^2 p_j \left[1 - s_{ij}^{0.05}\right],$$

where p_i is the proportional representation of ethnic group i in the national population; n is the total number of ethnic groups comprising the country's population; and s_{ij} is the "degree of similarity" between the languages spoken by ethnic groups i and j, given by the ratio of the number of common branches (shared by the two languages) to the maximum possible number of branching steps (i.e., 15) in the phylogenetic linguistic tree for all languages worldwide. See Esteban, Mayoral and Ray (2012) for additional details on primary data sources and methodological assumptions.

Ethnolinguistic polarization, Reynal-Querol index [POL-RQ]. An index of polarization, constructed by Esteban, Mayoral and Ray (2012), across the ethnic groups in a country's population, where ethnic groups by country are classified according to Fearon (2003) and the definition of polarization that is applied is the one due to Reynal-Querol (2002) and Montalvo and Reynal-Querol (2005). See Esteban, Mayoral and Ray (2012) for additional details on primary data sources and methodological assumptions.

A.2 Independent and Instrumental Variables

Migratory distance from East Africa. The great circle distance from Addis Ababa, Ethiopia to a country's modern capital city along a land-restricted path forced through one or more of five aforementioned intercontinental waypoints, including Cairo, Egypt; Istanbul, Turkey; Phnom Penh, Cambodia; Anadyr, Russia; and Prince Rupert, Canada. Distances are calculated using the Haversine formula and are measured in units of a thousand kilometers. The methodology underlying the construction of this measure is adopted from Ramachandran et al. (2005). The geographical coordinates of the waypoints are obtained from Ramachandran et al. (2005) and those of the modern capital cities are obtained from the Central Intelligence Agency's (CIA) *World Factbook* online. See Ashraf and Galor (2013) for additional details.

Genetic diversity. The expected heterozygosity (genetic diversity) of a country's contemporary national population, as developed by Ashraf and Galor (2013). This measure is based on migratory distances from East Africa to the year 1500 locations of the ancestral populations of the country's component ethnic groups in 2000 and on the pairwise migratory distances among these ancestral populations. The source countries of the ancestral populations are identified from the *World Migration Matrix, 1500–2000* (Putterman and Weil, 2010), and the modern capital cities of these countries are used to compute the aforementioned migratory distances. The measure of genetic diversity is then computed by applying (i) the coefficients obtained from regressing expected heterozygosity on migratory distance from East Africa at the ethnic group level, using a worldwide sample of 53 ethnic groups comprising the *Human Genome Diversity Cell Line Panel*, compiled by the Human Genome Diversity Project (HGDP) and the Centre d'Étude du Polymorphisme Humain (CEPH); (ii) the coefficients obtained from regressing pairwise genetic distance on pairwise migratory distance in a sample of 1,378 HGDP-CEPH ethnic group pairs, and (iii) the ancestry weights representing the fractions of the year 2000 national population (i.e., of the country for which the measure is being computed) that can trace their ancestral origins to different source countries in the year 1500. The ethnic group (and group-pair) level data on expected heterozygosities, geographical coordinates, and pairwise genetic distances are obtained from Ramachandran et al. (2005), and the country level data on ancestry weights are obtained from the *World Migration Matrix, 1500–2000* website. See Ashraf and Galor (2013) for a detailed discussion of the methodology underlying the construction of this measure.

Years since Neolithic. The number of thousand years elapsed (as of the year 2000) since the majority of the population residing within a country's modern national borders began practicing sedentary agriculture as the primary mode of subsistence. This measure, reported by Putterman (2008), is compiled using a wide variety of both region- and country-specific archaeological studies as well as more general encyclopedic works on the transition from hunting and gathering to agriculture during the Neolithic Revolution. See the *Agricultural Transition Data Set* website for additional details on primary data sources and methodological assumptions.

Settlement duration. The maximum duration, in tens of thousands of years, of uninterrupted settlement by anatomically modern humans across locations within a country's modern national

borders, as reported by Ahlerup and Olsson (2012). See Ahlerup and Olsson (2012) for additional details on primary data sources and methodological assumptions.

Duration as colony. The duration, in centuries, of experience by a country (or any subregion thereof) as a colony of one or more colonial powers, including the United Kingdom, France, Spain, Portugal, the Netherlands, Belgium, Italy, Germany, and the United States. In cases where different regions within a country's modern national borders were simultaneously colonized by different colonial powers, the durations of experience as a colony is aggregated across these regions. This variable is constructed by the authors of the current paper, based on colonization and decolonization dates obtained from a wide range of online resources, including (but not limited to) the CIA's *World Factbook*, the *Encyclopaedia Brittanica*, and *Country Studies* of the Library of Congress. Additional details on primary data sources and methodological assumptions are available from the authors upon request.

Absolute latitude. The absolute value of the latitude of a country's geodesic centroid, as reported on the *Gothos* website, based on online metadata from (i) the National Geospatial-Intelligence Agency's (NGA) *GEOnet Names Server (GNS)* and (ii) the United States Geological Survey's (USGS) *Geographic Names Information System (GNIS)*.

Mean land quality. A geospatial index of the suitability of land for agriculture, based on ecological indicators of climate suitability for cultivation, such as growing degree days and the ratio of actual to potential evapotranspiration, as well as on ecological indicators of soil suitability for cultivation, such as soil carbon density and soil pH. This index was initially developed at a half-degree resolution by Ramankutty et al. (2002), and it has been aggregated up to the country level by Michalopoulos (2012), by averaging values across the grid cells that are located within a country's national borders. The variable employed by the current analysis is thus the aggregate measure reported by Michalopoulos (2012). See Michalopoulos (2012) for additional details.

Variation in land quality. The standard deviation of the agricultural suitability index (as discussed above) across the grid cells (at a half-degree resolution) that are located within a country's national borders, as reported by Michalopoulos (2012). See Michalopoulos (2012) for additional details.

Mean elevation. The average elevation of a country, in thousands of kilometers above sea level, calculated using geospatial data at a 1-degree resolution from the *Geographically based Economic data (G-ECON)* project (Nordhaus, 2006), which is, in turn, based on similar data at a 10-minute resolution from New et al. (2002). The measure is aggregated up to the country level by averaging across the grid cells that are located within a country's national borders. See the *G-ECON* project website for additional details.

Variation in elevation. The standard deviation of elevation (as discussed above) across the grid cells (at a 1-degree resolution) that are located within a country's national borders. See the *G-ECON* project website for additional details.

Dispersion in elevation. The difference between the maximum and minimum values of elevation (as discussed above) across the grid cells (at a 1-degree resolution) that are located within a country's national borders. See the *G-ECON* project website for additional details.

Percentage of arable land. The fraction of a country's total land area that is arable, as reported for the year 2000 by the World Bank's *World Development Indicators* online.

Distance to waterways. The distance, in thousands of kilometers, from a geospatial grid cell to the nearest ice-free coastline or sea-navigable river, averaged across the grid cells that are located within a country's national borders. This variable, developed by Gallup, Sachs and Mellinger (1999), is available from the online *Research Datasets* repository maintained by Harvard University's Center for International Development.

Total land area. The total land area of a country, in millions of square kilometers, as reported for the year 2000 by the World Bank's *World Development Indicators* online.

Temperature. The average monthly temperature of a country, in units of ten degrees Celsius per month, over the 1961–1990 time period, calculated using geospatial data on average monthly temperature for this period at a 1-degree resolution from the *G-ECON* project (Nordhaus, 2006), which is, in turn, based on similar data at a 10-minute resolution from New et al. (2002). The measure is aggregated up to the country level by averaging across the grid cells that are located within a country's national borders. See the *G-ECON* project website for additional details.

Precipitation. The average monthly precipitation of a country, in units of ten millimeters per month, over the 1961–1990 time period, calculated using geospatial data on average monthly precipitation for this period at a 1-degree resolution from the *G-ECON* project (Nordhaus, 2006), which is, in turn, based on similar data at a 10-minute resolution from New et al. (2002). The measure is aggregated up to the country level by averaging across the grid cells that are located within a country's national borders. See the *G-ECON* project website for additional details.

Percentage of land in tropical and subtropical climate zones. The fraction of a country's total land area that is located in regions classified as tropical or subtropical by the Köppen-Geiger climate classification system. This variable, developed by Gallup, Sachs and Mellinger (1999), is available from the online *Research Datasets* repository maintained by Harvard University's Center for International Development.

Disease richness. The total number of different types of infectious diseases in a country, as reported by Fincher and Thornhill (2008), based on the *Global Infectious Disease and Epidemiology Network (GIDEON)* online database. See Fincher and Thornhill (2008) for additional details.

Island nation dummy. An indicator for whether or not a country shares a land border with any other country, as reported by the CIA's *World Factbook* online.

Landlocked dummy. An indicator for whether or not a country is landlocked, as reported by the CIA's *World Factbook* online.

B Supplementary Results

TABLE B.1: First-Stage Regressions

	(1) OLS First stage of Columns 2–6	(2) OLS First stage of Columns 7–8
Migratory distance from East Africa	-0.598***	-0.609***
	(0.084)	(0.094)
Years since Neolithic	0.066	0.109
	(0.096)	(0.101)
Settlement duration	0.122**	0.070
	(0.053)	(0.061)
Duration as colony	-0.041	-0.036
	(0.087)	(0.088)
Absolute latitude	0.032	0.009
	(0.034)	(0.035)
Mean land quality	0.031	0.882
	(1.122)	(1.240)
Variation in land quality	-0.387	-1.036
	(1.887)	(1.876)
Mean elevation	0.670	0.371
	(0.519)	(0.647)
Variation in elevation	-4.092**	-5.016**
	(1.661)	(1.988)
Dispersion in elevation	0.658*	0.921**
	(0.360)	(0.439)
Observations	143	129
Adjusted R^2	0.77	0.78
F-test of excluded instrument	50.24	41.84

Notes: This table reports the results from the first-stage regressions associated with the 2SLS regressions in Columns 2–8 of Table 1 of the paper, where genetic diversity (adjusted for post-1500 migrations) is instrumented using migratory distance from East Africa. All regressions include controls for the percentage of arable land, distance to waterways, total land area, temperature, precipitation, the percentage of land in tropical and subtropical climate zones, disease richness, and island, landlocked, and continental fixed effects. Heteroskedasticity-robust standard errors are reported in parentheses. *** denotes statistical significance at the 1 percent level, ** at the 5 percent level, and * at the 10 percent level.

TABLE B.2: Genetic Diversity and Ethnolinguistic Heterogeneity across Countries in the Old World

	(1) OLS EG	(2) OLS EF-F	(3) OLS EF-A	(4) OLS ELF-D	(5) OLS POL-D	(6) OLS POL-ER	(7) OLS POL-RQ
Genetic diversity	14.750***	6.867***	3.612*	3.881**	6.886**	1.120**	1.137*
	(4.317)	(1.948)	(1.940)	(1.550)	(2.663)	(0.486)	(0.589)
Years since Neolithic	-0.071*	-0.022	-0.010	-0.018	-0.026	-0.007	-0.006
	(0.038)	(0.019)	(0.021)	(0.015)	(0.027)	(0.005)	(0.006)
Settlement duration	0.038	0.020	0.010	0.011	0.019	0.003	-0.000
	(0.023)	(0.013)	(0.014)	(0.007)	(0.012)	(0.002)	(0.004)
Duration as colony	0.019	0.018*	-0.004	0.015	0.030*	0.003	0.005
	(0.024)	(0.011)	(0.013)	(0.010)	(0.018)	(0.004)	(0.005)
Absolute latitude	-0.003	0.001	-0.003	0.005	0.010	0.003**	0.003**
	(0.015)	(0.006)	(0.006)	(0.005)	(0.009)	(0.001)	(0.001)
Mean land quality	-0.638**	-0.239	-0.249*	-0.184**	-0.322*	-0.011	-0.037
	(0.305)	(0.158)	(0.147)	(0.092)	(0.171)	(0.027)	(0.043)
Variation in land quality	1.015*	0.215	-0.128	-0.056	-0.093	-0.028	-0.007
	(0.575)	(0.290)	(0.293)	(0.205)	(0.373)	(0.067)	(0.100)
Mean elevation	-0.152	-0.085	-0.058	-0.019	-0.041	0.009	0.050**
	(0.154)	(0.079)	(0.075)	(0.062)	(0.114)	(0.016)	(0.022)
Variation in elevation	-0.327	0.146	0.185	0.312*	0.575*	0.050	-0.004
	(0.385)	(0.193)	(0.185)	(0.162)	(0.294)	(0.052)	(0.059)
Dispersion in elevation	0.172*	0.026	0.019	-0.063*	-0.110*	-0.016	-0.010
	(0.090)	(0.046)	(0.044)	(0.037)	(0.066)	(0.013)	(0.014)
Observations	118	118	118	118	118	106	106
Adjusted R^2	0.41	0.48	0.49	0.30	0.31	0.26	0.12

Notes: This table demonstrates, exploiting variations across countries in the Old World, the statistically significant positive relationships between genetic diversity (adjusted for post-1500 migrations) and various measures of contemporary ethnolinguistic heterogeneity, conditional on geographical and historical covariates. All regressions include controls for the percentage of arable land, distance to waterways, total land area, temperature, precipitation, the percentage of land in tropical and subtropical climate zones, disease richness, and island, landlocked, and continental fixed effects. Heteroskedasticity-robust standard errors are reported in parentheses. *** denotes statistical significance at the 1 percent level, ** at the 5 percent level, and * at the 10 percent level.

References

Ahlerup, Pelle, and Ola Olsson. 2012. "The Roots of Ethnic Diversity." *Journal of Economic Growth*, 17(2): 71–102.

Alesina, Alberto, Arnaud Devleeschauwer, William Easterly, Sergio Kurlat, and Romain Wacziarg. 2003. "Fractionalization." *Journal of Economic Growth*, 8(2): 155–194.

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

Desmet, Klaus, Ignacio Ortuño-Ortín, and Romain Wacziarg. 2012. "The Political Economy of Linguistic Cleavages." *Journal of Development Economics*, 97(2): 322–338.

Duclos, Jean-Yves, Joan Esteban, and Debraj Ray. 2004. "Polarization: Concepts, Measurement, Estimation." *Econometrica*, 72(6): 1737–1772.

Esteban, Joan, and Debraj Ray. 2011. "Linking Conflict to Inequality and Polarization." *American Economic Review*, 101(4): 1345–1374.

Esteban, Joan, Laura Mayoral, and Debraj Ray. 2012. "Ethnicity and Conflict: An Empirical Study." *American Economic Review*, 102(4): 1310–1342.

Fearon, James D. 2003. "Ethnic and Cultural Diversity by Country." *Journal of Economic Growth*, 8(2): 195–222.

Fincher, Corey L., and Randy Thornhill. 2008. "Assortative Sociality, Limited Dispersal, Infectious Disease and the Genesis of the Global Pattern of Religion Diversity." *Proceedings of the Royal Society B*, 275(1651): 2587–2594.

Gallup, John Luke, Jeffrey D. Sachs, and Andrew D. Mellinger. 1999. "Geography and Economic Development." *International Regional Science Review*, 22(2): 179–232.

Michalopoulos, Stelios. 2012. "The Origins of Ethnolinguistic Diversity." *American Economic Review*, 102(4): 1508–1539.

Montalvo, José G., and Marta Reynal-Querol. 2005. "Ethnic Polarization, Potential Conflict, and Civil Wars." *American Economic Review*, 95(3): 796–816.

New, Mark, David Lister, Mike Hulme, and Ian Makin. 2002. "A High-Resolution Data Set of Surface Climate Over Global Land Areas." *Climate Research*, 21(1): 1–25.

Nordhaus, William D. 2006. "Geography and Macroeconomics: New Data and New Findings." *Proceedings of the National Academy of Sciences*, 103(10): 3510–3517.

Putterman, Louis. 2008. "Agriculture, Diffusion, and Development: Ripple Effects of the Neolithic Revolution." *Economica*, 75(300): 729–748.

Putterman, Louis, and David N. Weil. 2010. "Post-1500 Population Flows and the Long Run Determinants of Economic Growth and Inequality." *Quarterly Journal of Economics*, 125(4): 1627–1682.

Ramachandran, Sohini, Omkar Deshpande, Charles C. Roseman, Noah A. Rosenberg, Marcus W. Feldman, and L. Luca Cavalli-Sforza. 2005. "Support from the Relationship of Genetic and Geographic Distance in Human Populations for a Serial Founder Effect Originating in Africa." *Proceedings of the National Academy of Sciences*, 102(44): 15942–15947.

Ramankutty, Navin, Jonathan A. Foley, John Norman, and Kevin McSweeney. 2002. "The Global Distribution of Cultivable Lands: Current Patterns and Sensitivity to Possible Climate Change." *Global Ecology and Biogeography*, 11(5): 377–392.

Reynal-Querol, Marta. 2002. "Ethnicity, Political Systems, and Civil Wars." *Journal of Conflict Resolution*, 46(1): 29–54.