CESifo WORKING PAPERS

8202
2020

April 2020

A New Measure of Foreign Rule Based on Genetic Distance

Dhammika Dharmapala

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 8202*

A New Measure of Foreign Rule Based on Genetic Distance

Abstract

The consequences for countries of past foreign rule are the subject of a vast literature across history and the social sciences. This paper constructs a novel measure of past foreign (or minority) rule - the genetic distance of a country's ruling elite in the year 1900 from the country's ethnic majority - by mapping historical information on these groups to existing data on bilateral genetic distances between countries and populations. This generates an "elite-population genetic distance" in 1900 (EPGD_1900) for each of 228 present-day countries and territories. While this continuous measure is positively correlated with existing dichotomous measures of foreign rule, it captures an additional dimension of variation absent from the existing measures. The paper documents robust conditional correlations between EPGD_1900 and current income levels, and between EPGD_1900 and current fiscal capacity (controlling for various relevant country characteristics, existing measures of foreign rule, the genetic distance of a country's ethnic majority from that of the UK, and continent fixed effects). In particular, both current GDP per capita and tax revenue as a percentage of GDP are substantially lower for countries and territories with higher EPGD_1900. While these relationships may be attributable to unobserved and persistent variation in state-building capabilities across societies, the results are robust to controlling for a widely-used index of state antiquity that measures the history of state-building capacity.

JEL-Codes: O100, H200.

Keywords: foreign rule, fiscal capacity, economic development.

Dhammika Dharmapala
University of Chicago Law School
USA - 60637, Chicago, IL
dharmap@uchicago.edu

Revised version: March 2020
First version: January 2015
I thank the Editor (Areendam Chanda), two anonymous referees, Dan Klerman, Holger Spamann and participants at the 2015 International Institute of Public Finance Annual Congress (especially my discussant Mohammad Reza Farzanegan) for helpful discussions and comments regarding an earlier version of this paper. I also thank Ray Mao for outstanding research assistance, and Ruud de Mooij, Thornton Matheson and Kelsey Moser for their help in providing the IMF dataset of fiscal variables. I also acknowledge the financial support of the Lee and Brena Freeman Faculty Research Fund at the University of Chicago Law School Any remaining errors or omissions are my own.

1) Introduction

The consequences for economic development and other present-day outcomes of the global history of foreign rule and colonization are the subject of a vast literature across history and the social sciences. For instance, within economics, Acemoglu, Johnson and Robinson (2001; hereafter AJR) use European settler mortality in the colonial era as a source of exogenous variation in the mode of colonization and especially in the extent of settlement. Their approach is based on the idea that higher settler mortality led to extractive colonial structures that tend to have a persistent negative impact on the quality of institutions (and hence on the level of economic development).[1] More generally, there is growing interest among economists in the role of history in determining contemporary institutional characteristics and economic outcomes (e.g. Nunn, 2009). One growing strand of this literature uses measures derived from genetic data to proxy for migrations and interactions among human populations in the distant past (e.g. Spolaore and Wacziarg, 2009 (hereafter SW); Ashraf and Galor, 2013).

In this paper, we construct a novel measure of the genetic distance between ruling elites and the majority of the population in the recent past (specifically, the year 1900, for reasons discussed below). This measure is computed using bilateral genetic distance data for countries and populations constructed by SW, which in turn is based on data from Cavalli-Sforza, Menozzi and Piazza (1994). As calculated in these studies, genetic distance provides a measure of the time that has elapsed since two populations became separated (i.e. since they last shared common ancestors). The genetic distance measure is based on random drift in neutral characteristics and not on traits selected for fitness. In sufficiently large populations, such random genetic drift occurs at a predictable rate, and thus genetic distance as measured in this way will be larger for populations that have been separated for longer periods of time. It is important to emphasize that, as in SW, genetic distance is intended to serve as a proxy for proximity in *cultural* traits among populations; as the measure is based on random genetic drift, there is no implication whatsoever that traits relevant to economic development are transmitted genetically. It is of course possible to directly measure cultural similarities between ruling elites and ethnic majorities. However, genetic distance is arguably more exogenous, in that it is determined by patterns of prehistoric migrations rather

[1] Other quasi-experimental approaches to the study of the consequences of foreign rule include Feyrer and Sacerdote (2009) and Iyer (2010), which are discussed in more detail in Section 4 below. Note that some aspects of the construction of the settler mortality measure have been questioned by Albouy (2012).

than by the potentially endogenous adoption by populations of the cultural traits of ruling elites. Moreover, Spolaore and Wacziarg (2015) report evidence suggesting that genetic distance can serve as a useful summary statistic for a wide variety of cultural traits that are transmitted across generations.

SW report the bilateral genetic distance between each pair of present-day countries. Using this data, we calculate the genetic distance between ruling elites and populations in 1900 by using historical sources to determine the ethnicity of ruling elites and of the majority of the population for each country at that time. We then map these ruling elites and populations to countries or populations for which genetic data exists, and compute the corresponding genetic distances. This provides a measure of the "closeness" of ruling elites and the populations they ruled in 1900 for 228 present-day countries and territories. The year 1900 is chosen for concreteness, and because for many countries the measure for 1900 is quite representative of foreign or minority rule during the late nineteenth century and the first half of the twentieth century. More recently, Spolaore and Wacziarg (2018) report the bilateral genetic distance between country-pairs using an alternative genomic dataset based on human microsatellite variation (from Pemberton *et al.* (2013)). We use this data to construct an alternative version of the genetic distance between ruling elites and populations in 1900, which leads to very similar results to those for the baseline measure.

The "elite-population genetic distance" in 1900 (EPGD_1900) has a number of notable advantages over other approaches to measuring foreign rule. Existing measures of foreign rule are typically dichotomous. In the analysis in Section 2.2 below, we focus on two such measures in relation to EPGD_1900: the classification of countries' former colonial status by AJR, and the coding of countries' colonial history in the Issue Correlates of War (ICOW) database. Unlike these indicator variables for foreign rule (or a series of indicator variables for specific imperial powers, as used, for instance, by Klerman *et al*. (2011)), EPGD_1900 is a continuous measure. It takes account not only of rule by (external) empires, but also of the role of traditionally dominant domestic minority ethnic groups (such as Baltic Germans in Estonia and Latvia, Finland-Swedes in Finland, and Franco-Mauritians in British-ruled Mauritius); it thus provides a common metric for such diverse phenomena as Portuguese rule in Mozambique, Manchu rule in China, and Ottoman rule in Albania. Because it incorporates information on non-European as well as

European empires and ruling elites, it embodies a wider perspective than many existing measures of colonial rule.[2]

Consequently, EPGD_1900 allows us to pose a somewhat different set of questions than those addressed in the prior literature. These relate not to the *net* effect of foreign rule, but to how the "foreignness" or "distance" of past foreign or domestic minority rule (holding everything else, including institutions, constant) is associated with contemporary outcomes. For example, it may be possible for British legal origins to have a positive *partial* effect on contemporary outcomes, but for higher genetic distance between British rulers and imperial subjects to simultaneously have a negative *partial* effect on the latter's contemporary outcomes. Moreover, as argued in Section 2 below, EPGD_1900 is quite distinct from measures of countries' internal ethnic or genetic diversity, such as the widely-used ethnolinguistic fractionalization index (e.g. Desmet *et al*., 2012) and the measure of countries' internal genetic heterogeneity constructed by Ashraf and Galor (2013). It is possible, for instance, for a very homogeneous society that was under foreign rule in 1900 to have a high EPGD_1900, and for a very heterogeneous society to be ruled by an elite that is representative of all groups.

We document that EPGD_1900 varies widely across countries and regions of the world, reflecting differing experiences with past foreign or minority rule. It is positively correlated with existing binary measures of foreign rule, as might be expected. However, the magnitude of the correlation coefficient is relatively modest: about 0.47 with the AJR measure and about 0.35 with the ICOW variable. As discussed in Section 2.2 below, a principal components analysis of the variation among these three measures identifies a primary component that explains about 60% of the variation and is related in similar fashion to all three measures. However, a second component that explains much of the remaining variation is positively related to EPGD_1900 but negatively related to the AJR and ICOW measures. Arguably, this may be due to EPGD_1900 being a continuous measure that captures not only foreign or colonial rule as conventionally understood, but also the role of dominant domestic minority groups. Overall, comparisons between EPGD_1900 and existing variables suggest strongly that – while there is some substantial overlap – EPGD_1900 includes an additional dimension of variation absent from the existing measures.

[2] For example, EPGD_1900 is, on average, quite high for countries in Asia and Africa that were not under European rule in 1900, due to the prominent role in many such countries of rule by (non-European) empires or dominant domestic minority ethnic groups.

To further illustrate the potential value of this new measure, Section 3 below documents a robust conditional correlation between EPGD_1900 and current GDP per capita. In particular, countries with higher values of EPGD_1900 tend to have lower income levels today. This holds both among formerly colonized and non-colonized subsamples of countries (as classified by AJR), though the estimated effect is stronger among the former. This relationship is robust to controlling for a wide variety of relevant country characteristics, the existing measures of foreign rule described above, the genetic distance of a country's ethnic majority from that of the UK, legal origins, a widely-used measure of state antiquity, and continent fixed effects. Moreover, the implied magnitude is quite substantial.

The new EPGD_1900 measure can also potentially shed light on a central question in comparative economic development: the causes and consequences of the extensive variation in governments' ability to raise tax revenue. A literature analyzing the determinants of cross-country variation in governments' "fiscal capacity" has emerged in recent years (e.g. Besley and Persson, 2009, 2013, 2014; Dincecco and Prado, 2010, 2012). This literature has identified past external military conflicts and the presence of inclusive political institutions as crucial factors associated with the growth of fiscal capacity (measured using various proxies such as the ratio of revenue to GDP). In this literature, an important consideration in whether a state makes investments in tax administration and fiscal capacity is the extent to which ruling elites internalize the benefits to the wider population of the provision of public goods. This research on fiscal capacity forms part of a broader literature on taxation and development, within which an important theme relates to institutional and historical constraints on tax policymaking (e.g. Dharmapala and Hines, 2009).

EPGD_1900 may be hypothesized to potentially affect fiscal capacity through two primary channels. One is that foreign (or other genetically distant) rulers may invest less in developing fiscal capacity, as they do not internalize the benefits of government spending on programs such as mass education to the same extent as would less genetically distant elites. Another is that tax compliance may be lower when rulers are foreign or more distant. Arguably, the institutional features that emerge in response to these mechanisms may persist over time, even when the ethnicity of ruling elites changes. Moreover, the legacy of past investment in fiscal capacity may have long-term consequences – for instance, Dincecco and Prado (2010) argue that fiscal capacity positively affects education expenditures, which in turn increases productivity and income per capita.

This paper uses a dataset on fiscal variables compiled by the IMF for 171 countries to construct a standard measure of fiscal capacity - revenue as a percentage of GDP averaged over the period 2004-2013. Section 3 below documents a robust conditional correlation between EPGD_1900 and current fiscal capacity. In particular, countries with higher values of EPGD_1900 tend to have lower levels of revenue as a percentage of GDP today. This relationship is robust to controlling for a wide variety of relevant country characteristics, the existing measures of foreign rule described above, the genetic distance of a country's ethnic majority from the UK, legal origins, a widely-used measure of state antiquity, the extent of past external military conflicts, and continent fixed effects. It is also robust to using the maximum value of revenue and individual income tax revenue (relative to GDP) over the 2004-2013 period as an alternative measure of fiscal capacity. The implied magnitude is substantial, although well within the range of variation observed in the data.

The cross-sectional nature of the analysis raises obvious concerns about unobserved heterogeneity across countries. Perhaps the most compelling endogeneity story involves unobserved and persistent variation in state-building capabilities across societies. Lower levels of state-building capability may have caused societies to fall under foreign rule (or the rule of a genetically distant domestic minority group) in 1900 and also to have limited fiscal capacity today. To control for societies' state-building capability, we use an index of state antiquity originally constructed by Bockstette *et al*. (2002) and Chanda and Putterman (2007). For each country, they assign a score for each 50-year period since the year 1 of the Common Era, based on the existence of a large-scale state and the extent of independent local control.

We use an updated and extended version of this index constructed by Borcan *et al*. (2018). In particular, our baseline results use an ancestry-adjusted version of the index that measures the state-building history of populations currently resident in a country; as described in Section 2 below, this takes account of population flows since 1500. The index discounts past state-building history at a 1% discount rate. The results are also essentially identical using a variant of the index (from Hariri (2012)) that only measures state-building history up to 1500. As noted earlier, our results for both GDP per capita and revenue are robust to the inclusion of these state antiquity measures. This allays to some degree the concerns about persistent variation in state-building capabilities across societies. Nonetheless, in view of the remaining concerns about unobserved heterogeneity across countries, the role of these cross-country correlations is not to establish any form of causal inference, but rather to suggest hypotheses that may be worthy of further investigation.

Section 2 describes the construction of the EPGD_1900 measure, and the other variables used in the analysis. Section 3 reports the results. Section 4 discusses their implications and limitations. Section 5 concludes.

2) Data

2.1) The Construction of the Genetic Distance Measure

This paper constructs a novel measure of the genetic distance between ruling elites and populations in 1900 for a sample of 228 present-day countries and territories.[3] The measure is based most proximately on the bilateral genetic distance data for country-pairs constructed by SW. They report a measure of the genetic distance between the populations of each pair of countries in the world. This variable is based on data from the classic study of human population genetics by Cavalli-Sforza, Menozzi and Piazza (1994), who report measures of the genetic distance between various human populations classified by ethnicity. SW transform this data from the ethnicity-pair level to the country-pair level, using information on the location of ethnic groups by country and the ethnic composition of countries. This transformation enables the use of these country-pair genetic distances in the analysis of economic and institutional variables that are typically defined only at the country level. In particular, SW study the role of genetic distance between countries – and especially the genetic distance between a country and the country or countries at the technological frontier - on the diffusion of economic development.

As calculated in SW, genetic distance (known as F_{ST} distance) provides a measure of the time that has elapsed since two populations became separated during the global migrations of modern humans out of their place of origin in Africa. In particular, the F_{ST} distance measure serves as a proxy for the length of time since these populations last shared common ancestors. The genetic distance measure is based on random drift in neutral characteristics and not on traits selected for fitness. In sufficiently large populations, such random genetic drift occurs at a predictable rate, and thus genetic distance as measured in this way will be larger for populations that have been separated for longer periods of time. As emphasized earlier, genetic distance is intended to serve

[3] Note, however, that due to data requirements for certain variables, the set of countries and territories in the sample are those that existed in 2004, and does not include new states created since then.

as a proxy for proximity in *cultural* traits among populations; as the measure is based on random genetic drift, there is no implication whatsoever (either in SW or in this paper) that traits relevant to economic development are transmitted genetically.

As noted above, SW report the bilateral genetic distance between each pair of present-day countries. In the units used in the SW dataset, these bilateral genetic distances vary from zero to 3375, with a mean of about 1169. Using this data, we calculate the genetic distance between ruling elites and populations in 1900 by using historical sources to determine the ethnicity of ruling elites and of the majority of the population for each country at that time. We begin by consulting the 11th edition of the *Encyclopaedia Britannica* - which was published in 1911 (and therefore written very close to the time period in which we are interested) and is now widely available online[4] - to make an initial determination of the political status of each country or territory in 1900. We then use a large number of country-specific (or, in some cases, region-specific or empire-specific) scholarly sources to determine the scholarly consensus regarding the ethnicity of each country's ruling elite and the ethnicity of its majority population in 1900. The Data Appendix below reports, for each of the 228 countries and territories in our sample, the ethnicity of the ruling elite in 1900 and characterizes the country's ethnic majority, as determined using these various sources. For some countries, there are reasonable alternative interpretations of its political status or of the identity of its ruling elite or ethnic majority. As discussed below in more detail, in these instances we use alternative characterizations of the ethnicity of the ruling elite in 1900 and of the country's ethnic majority to construct an alternative value of EPGD_1900 for that country that is used in robustness checks; these are also reported in the Data Appendix. The country-specific scholarly sources consulted in this process are listed in the Online Appendix that is available upon request.

We then map these ruling elites and populations to countries or populations for which genetic data exists in the SW dataset, and compute the corresponding genetic distances. For a country Y whose ruling elite's ethnicity is that associated with country X (or which was ruled by country X), EPGD_1900 is defined as follows:

$$\text{EPGD\_1900} = \text{diff}[X, Y], j \qquad (1)$$

[4] The 11th edition of the *Encyclopaedia Britannica* is available at various locations online, including for instance: https://catalog.hathitrust.org/Record/007910230

where "diff" refers to the F_{ST} genetic distance measure between country X and country Y in the SW dataset.[5] The term "*j*" refers to one of two alternative measures of the genetic distance between country X and country Y reported in SW: "dominant" or "weighted." The former involves using the genetic distance between the ethnic majority (i.e. numerically dominant) group of country X and the ethnic majority (i.e. numerically dominant) group of country Y. The latter takes account of the different ethnic groups (weighted by relative population size) within each country; for instance, if country X includes only one ethnic group (group A) and country Y includes group B (constituting 80% of the population) and group C (constituting 20% of the population), then the weighted distance measure between X and Y is a weighted sum of the genetic distances between groups A and B (with a weight of 0.8) and between groups A and C (with a weight of 0.2).

Typically, the "dominant" measure is the appropriate one for our purposes. For instance, if country Y in our example above was ruled by country X in 1900, we generally wish to set EPGD_1900 equal to the genetic distance between ethnic groups A and B (which is accomplished by using the "dominant" genetic distance between countries X and Y). However, for a relatively small number of countries that have ethnic majorities with highly heterogeneous ancestry, the "weighted" measure is arguably more appropriate. For example, suppose that instead of having two distinct ethnic groups B and C, country Y has an ethnic majority that is descended from both ethnic groups B and C. Then, we generally wish to set EPGD_1900 equal to the "weighted" genetic distance between countries X and Y (which represents the genetic distance between ethnic group A and a composite of ethnic groups B and C). The Online Appendix specifies for each of the 228 countries and territories in our sample precisely how EPGD_1900 is derived from the SW data in terms of Equation (1), and notes those instances (mostly for countries in Central and South America) in which the weighted rather than dominant measure is used.

The same procedure described above is used to construct an alternative version of EPGD_1900 from the more recent dataset constructed by Spolaore and Wacziarg (2018). This also reports the bilateral genetic distance between country-pairs, but uses an alternative genomic dataset based on human microsatellite variation derived from Pemberton *et al.* (2013). In the units used in the Spolaore and Wacziarg (2018) dataset, these bilateral genetic distances vary from zero to about

[5] Note that this distance is based on genetic studies on populations in the late twentieth century. However, this can be viewed as a reasonable proxy for the distance in 1900 because of the extremely slow rate at which random drift occurs.

0.1, with a mean of about 0.04. In some respects, the Pemberton *et al.* (2013) genetic data is more detailed than that used in the construction of the original SW dataset. However, it has somewhat more limited coverage than that of our baseline variable. Thus, we use the SW-based version of EPGD_1900 as our basic measure. However, the alternative version of EPGD_1900 derived from the Spolaore and Wacziarg (2018) data leads to very similar results (as shown in Table 7 and discussed in Section 3 below).

EPGD_1900 differs in significant respects from various other concepts in the literature. The ethnolinguistic fractionalization index (e.g. Desmet et al., 2012) has been widely used in the literature on comparative development. This, however, measures the ethnic and linguistic heterogeneity within a country, not the closeness of that country's population to foreign rulers. EPGD_1900 is also quite distinct from the measure of countries' internal genetic heterogeneity constructed by Ashraf and Galor (2013), who argue that internal genetic diversity (which tends to decline with geographical distance from modern humans' place of origin in Africa) has affected patterns of long-term economic development. Their measure can be viewed essentially as capturing heterogeneity *within* the ethnic majority, not the distance between the ethnic majority and a ruling elite. Nonetheless, both ethnolinguistic fractionalization and internal genetic heterogeneity are used as controls in the analysis below.

For many countries, the coding of EPGD_1900 is relatively straightforward. An example is Mozambique, which was under Portuguese rule in 1900 (e.g. Clarence-Smith, 1985). We code the ruling elite as consisting of Portuguese officials, and we use the country-pair genetic distance reported in SW for Mozambique and Portugal. As discussed above, the relevant measure for our purposes is the genetic distance between the ethnic majority of Portugal and the ethnic majority of Mozambique, and so we set Mozambique's EPGD_1900 equal to diff[Mozambique, Portugal], dominant (the "dominant" version of the genetic distance between Mozambique and Portugal.

However, there are some countries for which calculating EPGD_1900 is less straightforward. For instance, the political status of some countries in 1900 is ambiguous and not susceptible to straightforward interpretation and coding. Classifying ruling elites by ethnicity also requires a certain amount of judgment. In general, we rely on a variety of factors that are highlighted in the country-specific sources. This classification is not, for instance, based solely on the ethnicity of the individual ruler. For example, Bulgaria in 1900 had a monarch of German origin, but the vast majority of government officials consisted of ethnic Bulgarians, and so the ruling elite (as well as the population) is classified as Bulgarian (e.g. Jelavich, 1983).

Bosnia and Herzegovina provides a particularly powerful example of these difficulties. In 1900 Bosnia and Herzegovina was nominally an Ottoman province. However, it was under the *de facto* control of the Austro-Hungarian (Habsburg) Dual Monarchy (e.g. Okey, 2007). As the senior Habsburg officials governing this region who are listed in historical sources were predominantly ethnic Germans rather than ethnic Hungarians, we use the genetic distance between Austria and Yugoslavia reported by SW to calculate EPGD_1900 for Bosnia and Herzegovina. In view of the ambiguity in political status, we also construct an alternative measure (used in robustness checks) that classifies the ruling elite as Ottoman Turkish and sets EPGD_1900 for Bosnia and Herzegovina equal to the genetic distance between Turkey and Yugoslavia (as reported in the SW dataset). For the Austro-Hungarian Dual Monarchy more generally (e.g. Sugar, 1963; Taylor, 1976), we assign the present-day Czech Republic and Slovenia to the Austrian Empire (and so use the genetic distance of these countries' ethnic majorities to Austria) and assign Croatia and Slovakia to the Kingdom of Hungary (and so use the genetic distance of these countries' ethnic majorities to Hungary).

In general, in the face of such ambiguities, the baseline classifications used to construct EPGD_1900 reflect an interpretation of political realities that stresses local control and *de facto* rather than *de jure* rule (as indicated by the example of Bosnia and Herzegovina). Another example is Egypt, which in 1900 was nominally under Ottoman rule, but was under *de facto* British control (e.g. Tignor, 1966). We use the UK-Egypt genetic distance as Egypt's EPGD_1900, while also constructing an alternative measure used in robustness checks that relies on the Turkey-Egypt genetic distance. The general principle we use also implies that for the baseline EPGD variable we classify the elites of Estonia and Latvia as Baltic-German and that of Finland as Finland-Swedish; robustness checks use an alternative measure that classifies the ruling elite as Russian as all three countries were part of the Russian Empire in 1900 (e.g. Kappeler, 2001).

Countries and territories that had the status of imperial "protectorates" in 1900 raise particularly difficult issues of interpretation and judgment. For instance, British protectorates had assigned to them an official known as a "Resident" whose *de facto* role varied considerably, from being essentially a diplomat to being in effect the administrator of the protectorate (e.g. Onley, 2007). Here again, we use the general principle that emphasizes *de facto* control, but with context-specific judgments and with the use of alternative classifications in robustness checks. For example, the Gulf states of Bahrain, Kuwait and the United Arab Emirates are classified as being under local rule in 1900 in the baseline analysis, but an alternative classification used in robustness checks classifies them as being under the rule of their British Residents. A similar approach is used for Malaysia (then primarily the Federated Malay States), Brunei, and present-day Botswana (then known as the Bechuanaland Protectorate).

Where political or ethnic patterns in 1900 were particularly complex or heterogeneous across regions, we use the ruling elite and ethnic majority in and around the country's present-day capital as proxies. For example, India in 1900 consisted of several provinces under direct British rule, along with approximately 600 polities (often referred to as Princely States) under local rule but British suzerainty (e.g. Ramusack, 2004). In view of this complexity, we focus on the situation in the present-day capital city of New Delhi and its environs, and thus use the UK-India genetic distance to compute India's EPGD_1900.

In a few instances, it is not possible to match ruling elites or ethnic majorities to the SW dataset. In these cases, we use proxies based on populations included in the SW dataset that country-specific scholarly sources indicate are of similar ethnicity. Note, however, that this issue primarily affects smaller countries, and these tend to also have missing values of the other variables used in the regression analysis. Thus, the countries for which proxies are needed play a very limited role in the regression analysis described below. For the vast majority of countries, the ethnic majority in the latter part of the twentieth century (to which the SW data pertain) is identical to the ethnic majority in 1900. However, for a small number of countries in our sample, historical sources indicate that the ethnic majority in 1900 was different to that in the latter part of the twentieth century.[6] In these instances, we use the ethnic majority in 1900 (rather than the current ethnic majority) to construct EPGD_1900. The results are robust to excluding these countries from the analysis.

While it would be possible to compute similar measures for other years, the year 1900 is chosen for concreteness. For many countries, the measure for 1900 provides a representative picture of foreign or minority rule during the late nineteenth century and the first half of the

[6] These countries are Brazil, Guyana and Israel.

twentieth century.[7] This period witnessed the maximal extension of European empires. As a consequence, EPGD_1900 is positively correlated with genetic distance from Northwest Europe. This may potentially confound the investigation of how EPGD_1900 is related to current outcomes. To control for this, we also compute the genetic distance between each country's ethnic majority and (the ethnic majority of) the UK (using the SW dataset). The UK is the most common imperial ruler in our sample, and genetic distance to the UK is highly correlated with genetic distance from Northwest Europe in general. The regression analysis below uses genetic distance to the UK as a control variable.

2.2) EPGD_1900 and Existing Measures of Foreign Rule

Table 1 reports descriptive statistics for EPGD_1900 and various related variables. The global mean of EPGD_1900 is around 839 for the 228 countries and territories that we code (using the same units as in the SW dataset). Several variants of the measure – involving alternative coding of the ruling elites of Estonia, Finland and Latvia, of the ruling powers for several countries in the Eastern Mediterranean and Southeast Europe (Cyprus, Egypt and Bosnia and Herzegovina), of the ruling powers in the Gulf region, and of the ethnic majorities in some southern African countries (all detailed extensively in the Online Appendix) – are also reported. Note that the results described in Section 3 below are robust to using any of these variants of EPGD_1900.

The variation in the measure across countries is mapped in Figure 1. As is evident from the map, EPGD_1900 varies substantially across countries and across continents, reflecting varying patterns of past foreign rule. As shown in Table 1, EPGD_1900 is highest on average for Africa and lowest for Europe, with Asia being below, and the Americas and Oceania above, the global mean. Countries' genetic distance to the UK is slightly larger on average than their EPGD_1900, with a mean of about 916. The alternative version of EPGD_1900 using genomic data from microsatellite variation has a mean of about 0.024 (using the same units as in the Spolaore and Wacziarg (2018) data), while countries' genetic distance to the UK is about 0.029 on average.

This subsection seeks to compare EPGD_1900 to two existing influential measures of foreign rule. One is AJR's binary classification of the former colonial status of 161 countries that

[7] If political or military control changed during the year 1900 (as with the French conquest of Chad) we use the regime prevailing at the end rather than the beginning of 1900.

can be matched to our sample countries.[8] About two thirds of these countries are ex-colonies according to their coding. The other is derived from the ICOW colonial history dataset constructed by Paul Hensel,[9] which ". . . seeks to identify colonial or other dependency relationships for each state over the past two centuries." The ICOW data reports the former colonial ruler for each of 192 countries and territories that can be matched to our sample countries. The ICOW dataset makes a determination as to the "primary" former ruler in cases where a country had multiple foreign rulers at different times. It also reports former colonies' dates of independence. We construct a binary ICOW "foreign rule" variable by coding all countries that have a former colonial ruler in the ICOW dataset as ex-colonies. About 85% of the 192 countries in the ICOW data are ex-colonies under this definition.

It is potentially of interest to consider how EPGD_1900 varies across former rulers. Using the ICOW definition of the primary former ruler, countries formerly ruled by France have the highest mean value of EPGD_1900, with somewhat (but not dramatically) lower values for countries formerly under British or Spanish rule (as shown in Table 1).[10] Historically, European colonization is often viewed as having occurred in different waves (for instance, in the sixteenth century versus the nineteenth century). We proxy for these waves by using the ICOW dataset's information on the date of independence. For the 30 countries that became independent prior to 1900 (primarily in the Americas, representing the earlier wave of colonization), the mean EPGD_1900 is around 434, and is substantially lower than that for countries that became independent after 1900 (representing a later wave of colonization); the latter group of 133 countries has a mean EPGD_1900 of over 1000. The interpretation of this difference, however, is complicated by the possibility that it may potentially be affected by more extensive European settlement in the countries that experienced the earlier wave of colonization.

An important theme in the scholarly literature on colonialism is the distinction between direct and indirect modes of colonial rule. There is no standard source for data on the extent of direct versus indirect rule across all former colonies. However, Lange (2004) constructs an index of the importance of legal and administrative institutions characteristic of indirect rule for 33 formerly British-ruled territories. This index represents the number of court cases recognizing

[8] This data is available at: https://economics.mit.edu/faculty/acemoglu/data/ajr2001
[9] This data s available at: https://www.paulhensel.org/icowcol.html
[10] Note that the ICOW definition of the primary former foreign ruler may refer to rule at any time over the last two centuries, and does not necessarily entail that this former ruler ruled the country in 1900.

customary law in 1955 relative to the total number of court cases in that territory in that year. For the 12 countries with an index of zero (indicating direct rule), the mean EPGD_1900 is just under 1000. For the 21 countries with an index greater than zero (indicating varying degrees of indirect rule), the mean EPGD_1900 is about 1200. This suggests that indirect rule was more likely to be used with more genetically distant populations; however, the difference in the mean EPGD_1900 is fairly small in relation to the overall variation in the measure.

A central question regarding EPGD_1900 is whether it embodies substantial incremental information relative to the existing binary measures of foreign rule. Table 2 presents a correlation matrix for EPGD_1900 and the AJR and ICOW measures. It is reasonable to expect that these should all be positively correlated, as they seek to measure related phenomena; indeed Table 2 shows positive correlation coefficients. However, the magnitude of the correlation between EPGD_1900 and the other two measures is relatively modest: about 0.47 for the AJR measure and about 0.35 for the ICOW measure. This suggests that there is considerable variation in EPGD_1900 that is not fully captured by existing binary measures (for instance, because EPGD_1900 represents a continuous measure of the foreignness of foreign rule and because it also captures the role of dominant domestic minorities as well as colonial rule).

To explore this question further, we undertake a principal components analysis of the variation represented by EPGD_1900, the AJR variable, and the ICOW measure. In general, a principal components analysis identifies a set of uncorrelated linear combinations (components) of the variables representing the common underlying variation. Typically, principal components analysis is used to reduce dimensionality in a setting where a large number of variables measure related phenomena. This is not of course the aim in our setting; however, identifying the underlying components can shed light on the structure of the data. The principal components analysis indicates the existence of a primary component (Component 1) that accounts for 63% of the variation in the three variables. Moreover, as shown in Figure 2, Component 1 is related positively (and about equally, at a level of about 0.5 or 0.6) to all three measures. However, there is also a second component (Component 2) identified by the analysis, explaining about 22% of the variation. Unlike Component 1, Component 2 loads very differently for EPGD_1900 compared to the other two measures: as shown in Figure 2, it is positively (and quite highly, at a level of about 0.77) related to EPGD_1900, whereas it is negatively related to the AJR and ICOW measures.

Principal components analysis does not in itself provide any guidance as to the substantive interpretation of the various components. However, it may be reasonable to view Component 1 as capturing the variation across countries associated with foreign rule ("external" colonialism), while Component 2 may capture variation associated with countries that did not experience foreign rule but were dominated by domestic minority elites ("internal" colonialism, captured by EPGD_1900 but not by the other measures). This, however, is to some extent speculative.

Overall, though, it appears that - while there is substantial overlap among the three measures - EPGD_1900 captures an incremental dimension of variation absent from the existing measures. It should also be emphasized that the regression analysis below controls for the AJR and ICOW measures. Thus, the estimated relationship between EPGD_1900 and current GDP per capita (fiscal capacity) should be interpreted as, in effect, the relationship between this incremental variation embodied in EPGD_1900 and current GDP per capita (fiscal capacity). Moreover, the robust conditional correlation that is documented in Section 3 below between foreign or minority rule (as measured by EPGD_1900) and current GDP per capita (and between EPGD_1900 and fiscal capacity) does not exist for either the AJR or ICOW measures of foreign rule.

2.3) Outcome and Control Variables

The primary outcome variables used in the analysis are GDP per capita and revenue as a percentage of GDP (used as a proxy for fiscal capacity). The GDP per capita measure is obtained from Dharmapala and Hines (2009) and is for 2004, measured in US dollars in purchasing power parity (PPP) terms. Dharmapala and Hines (2009) primarily use the World Bank's World Development Indicators (WDI) database for this measure. However, for countries and territories for which GDP data are missing in WDI (typically, smaller jurisdictions), they use estimates of GDP per capita (also in thousands of US$, in PPP terms, for 2004 or the nearest available year) from the CIA's World Factbook.[11] As shown in Table 3, the mean GDP per capita in our sample is about $11,000. Figure 3 depicts a bivariate plot of the relationship between EPGD_1900 and GDP per capita in 2004. Given the regional distribution of EPGD_1900 in Figure 1, the negative

[11] The year 2004 corresponds to the beginning of the period over which the fiscal capacity measures are averaged (2004-2013). However, the precise year chosen is unlikely to be important, as cross-sectional differences in such country characteristics tend to be highly stable over short periods of time.

relationship is perhaps unsurprising. The regression analysis in Section 3 below explores the robustness of this relationship to adding a very extensive set of controls.

The measures of fiscal capacity are based on an IMF dataset that represents an updated version of that in Keen and Mansour (2010). It reports fiscal variables for IMF member countries, and covers up to 171 countries over the period 1960-2013 (although with many missing observations). The dataset reports government revenue (including social contributions) as a percentage of GDP and tax revenue (excluding social contributions) as a percentage of GDP. It also reports various components of revenue. Although this dataset is available as a panel, the basic research questions addressed in this paper can in practice only be addressed cross-sectionally.[12] Thus, we compute current measures of fiscal capacity, notably the mean of revenue as a percentage of GDP over the 10-year period 2004-2013, using all available data (so that the mean revenue variable is nonmissing whenever at least one year of data is available over 2004-2013). In addition, we compute the maximum value of revenue over the 2004-2013 period as an alternative measure of fiscal capacity (since the maximum may possibly reflect a government's ability to tax better than does the mean); the results below are robust to using maximum rather than mean revenue measures. The revenue variables include social contributions as well as taxes; however, the results are very similar when tax revenue (excluding social contributions) is used instead.

Descriptive statistics for these measures of fiscal capacity are shown in Table 3. The primary measure – the mean of revenue as a percentage of GDP over 2004-2013 – has a mean of about 31%. The alternative measure - the maximum of revenue as a percentage of GDP over 2004-2013 - has a mean of about 37%. Figure 4 shows a bivariate plot of the relationship between EPGD_1900 and revenue as a percentage of GDP (averaged over 2004-2013); a negative relationship is evident. Section 3 below describes the regression analysis that we use to investigate rather more systematically the relationship in Figure 4.

The regression analysis uses an extensive set of control variables (descriptive statistics for which are reported in Table 3). These include population size, expressed in thousands, from Dharmapala and Hines (2009), who primarily use the World Bank's World Development Indicators (WDI) database for this measure. However, for countries and territories for which

---

[12] In principle, it would be possible to construct a longitudinal variable similar to EPGD_1900 over time. While some countries have experienced changes in the ethnicity of ruling elites – due for instance to decolonization – over the period covered by the dataset, the revenue data does not have sufficient coverage across countries over a long time span to make this approach viable.

population data is missing in WDI (typically, smaller jurisdictions), they use estimates of population (also in thousands, for 2004 or the nearest available year) from the CIA's World Factbook. A variety of geographical variables are used. These are the absolute value of countries' latitude, from Gallup, Sachs and Mellinger (1999), an indicator variable for landlocked countries, and the distance by air from the nearest of New York, Rotterdam and Tokyo (both extended versions of variables constructed by Gallup, Sachs and Mellinger (1999), obtained from Dharmapala and Hines (2009)). The distance by air variable is measured in kilometers, and represents "the smallest distance of the country's capital city to one of the following three cities: New York, Rotterdam, or Tokyo." (Gallup, Sachs and Mellinger, 1999, fn. 13, pp. 4–5). An indicator variable for islands is obtained from Dharmapala and Hines (2009), who code this using information from the CIA's World Factbook.

Population density in 1500 is a widely-used measure that serves as a proxy for countries' economic development prior to the era of European expansion. The natural logarithm of this variable is obtained from Borcan *et al*. (2018). Ethnolinguistic fractionalization is a variable that is widely used in the literature to measure the internal diversity of countries' populations. It is computed as the probability that two randomly chosen individuals in a given country speak a different language. Recently, Desmet *et al*. (2012) have extended and updated this variable. They report a series of measures of ethnolinguistic fractionalization and polarization, based on a linguistic tree that depicts the genealogical relationships among nearly 7000 languages. In particular, they report measures of fractionalization at different levels of aggregation (for instance, at the level of language families rather than of individual languages). In the regression analysis below, we use their measure of fractionalization at the highest level of aggregation.[13]

Another measure of internal diversity – based on genetic rather than linguistic data – is constructed by Ashraf and Galor (2013). They draw on measures of "expected heterozygosity" constructed by population geneticists that represent the probability that two randomly selected individuals in a population differ genetically with respect to a spectrum of traits. They aggregate these measures to the country level to construct an index of the internal genetic diversity of a country, which tends to decline with geographical distance from modern humans' place of origin in Africa. Ashraf and Galor (2013) argue that their measure – which essentially captures genetic heterogeneity within a country's ethnic majority - has affected patterns of long-term economic

[13] This is the variable that is labeled "ELF1" in their dataset, which is available at: http://faculty.smu.edu/kdesmet/

development in a nonlinear (inverted U-shaped) manner. Thus, the regression analysis described below controls for a quadratic function of their measure.

As genetic distance between countries may be correlated with the geographical distance between them, we construct a measure of the geographical distance between each country and the country (if any) that ruled it in 1900. This distance is based on the Centre d'Etudes Prospectives et D'Informations Internationale (CEPII) dataset (see Mayer and Zignago (2011)). The CEPII dataset reports the distance between each bilateral country-pair (X, Y) based on the bilateral distance between the largest city in country X and the largest city in country Y, with the distance between the cities being weighted by the fraction of the population of each country that lives in that city. The variable is measured in kilometers. If a country had no foreign ruler in 1900, our geographical distance measure equals zero. Note that this includes countries dominated by a domestic minority ethnic group (i.e. such countries are coded as having a geographical distance of zero if they had no foreign ruler in 1900). The underlying presumption is that dominant domestic minorities (such as Baltic Germans in Estonia) are as "indigenous" to their country as is the ethnic majority (for example, ethnic Estonians).

Importantly, our analysis includes an updated version of the index of state antiquity constructed by Bockstette *et al*. (2002) and Chanda and Putterman (2007). For each country, they assign a score for each 50-year period from the year 1 of the Common Era to 1950 based on the existence of a large-scale state and the extent of independent local control. We use an updated and extended version of this index constructed by Borcan *et al*. (2018) to control for societies' state-building capability. Our baseline analysis uses an ancestry-adjusted version of the index that measures the state-building history of populations currently resident in a country, and so takes account of population flows since 1500. For instance, in this version of the index, the state-building history of Australia primarily reflects the state-building history of its European settlers rather than its indigenous population (and is thus close to the index value for the UK). The index discounts past state-building history at a 1% discount rate.

A potential concern with controlling for state-building capacity using the index described above is that post-1500 values of the index may partly reflect state-building activities conducted by European colonizers or other foreign rulers. To address this, Hariri (2012) uses a version of the state antiquity index that only represents state history up to 1500. To replicate the Hariri (2012) version of the index, we use the original data on each 50-year period constructed by Bockstette *et* *al*. (2002) and Chanda and Putterman (2007),[14] and omit the 50-year periods since 1500 (thereby avoiding the potential influence on the index of state-building activities by European colonial powers). The results below are essentially identical when using the Hariri (2012) variant of the index.

We also include as a control the World Bank governance index constructed by Kauffmann *et al*. (2005). They use an unobserved components weighting procedure to construct aggregate country scores for the following elements of country-level governance: Voice and Accountability, Political Stability, Government Effectiveness, Regulatory Quality, Rule of Law, and Control of Corruption. Each of these measures takes values from approximately -2.5 to 2.5, with higher values indicating better governance). Each index is normalized so that the mean across all countries is 0 and the standard deviation is 1. We use the (unweighted) sum of these six measures for each country in 2004 as our measure of the quality of governance. We also include indicator variables for each of four historical legal origins – the British, French, German and Scandinavian legal families identified by scholars of comparative law - reported in La Porta *et al*. (2008). As is well-known, a very large literature has investigated the claim that these historical legal origins are associated with cross-country differences in a wide variety of legal and economic outcomes.

A longstanding theme in the study of the history of taxation is the role of external military conflicts in stimulating the development of state-building and fiscal capacity. Dincecco and Prado (2012) use historical sources to construct a series of variables that provide quantitative measures of countries' military history over the period 1816-1913. These include the number of wars fought, the number of deaths in war, and the number of years during which the country was engaged in war. The country coverage of these variables is relatively limited, with observations on 96 countries. Finally, our regression specifications include continent fixed effects, for Africa, the Americas, Asia, Europe and Oceania, based on the classification in Borcan *et al*. (2018).

3) Results

3.1) Income Levels

---

[14] This data is available at: http://www.econ.brown.edu/fac/louis_putterman/antiquity%20index.htm. Note that the units for the Hariri (2012) version are different from the units in the Borcan *et al*. (2018) version of the index due to the use of a scaling factor in the latter.

The basic specification for analyzing the relationship between EPGD_1900 and current income levels involves an OLS regression of the natural logarithm of the GDP per capita variable described in Section 2.3 on the natural logarithm of (1 + EPGD_1900). The use of the log of GDP per capita is quite standard, and we use the log of EPGD_1900 in the regression analysis to account for the skewness in the variable, adding a constant of 1 before taking the logarithm in order not to omit countries for which EPGD_1900 = 0. Admittedly, adding an arbitrary constant potentially affects the interpretation of the magnitude of the relationship, but this is of limited concern in an exploratory analysis such as this where the estimated magnitudes are primarily illustrative. Note, however, that the basic results are fairly similar using the level rather than log of EPGD.

The basic results are presented in Table 4 in stepwise fashion. Column 1 reports the simple bivariate relationship corresponding to Figure 3, and shows that the negative relationship evident in Figure 3 is highly significant (note that all of the regression analysis reports robust standard errors). Column 2 adds continent fixed effects; while its magnitude is somewhat smaller, the relationship remains negative and highly significant. Apart from Column 1 of Tables 4 and 5, all regressions reported in the paper include continent fixed effects. Thus, the estimates represent *within-continent* relationships (i.e. EPGD_1900 and current income levels are negatively related within each continent). This implies that the estimated relationship is not simply an artifact of the distribution of EPGD_1900 across continents (with, for instance, relatively low values in Europe and relatively high values in Africa).

As discussed in Section 2.2, EPGD_1900 is intended to capture elements of foreign or minority rule that go beyond existing binary classifications of (primarily European) colonization. Thus, any effect it may have should, in principle, apply both in formerly colonized countries and elsewhere. Column 3 of Table 4 reports the results from the same specification as in Column 2, restricting the sample to those countries that are classified by AJR as ex-colonies. Column 4 of Table 4 reports the corresponding results for the subsample of countries that are classified by AJR as not being ex-colonies. The negative relationship is stronger in the subsample of colonized countries, with a larger coefficient that is statistically significant at the 1% level.[15] However, it is also quite strong among non-colonized countries (and is statistically significant at the 5% level, even though there are fewer than half the number of observations as in the colonized subsample).

[15] This relationship among the colonized subsample is also robust to controlling for the log of population density in the year 1500 (a widely-used proxy for the level of economic development in the precolonial era).

Arguably, this reflects the role of dominant domestic minorities or of rule by foreign non-European rulers among the countries in the non-colonized subsample.[16]

Column 5 of Table 4 adds a set of particularly important control variables, all of which can be regarded as essentially exogenous or predetermined with respect to current income levels. These include the log of one plus the genetic distance of a country's ethnic majority from that of the UK. This controls for the potentially confounding effect of cultural factors specific to the UK or to Northwest Europe more generally, as discussed in Section 2.1. The two existing measures of foreign rule analyzed in Section 2.2 – the AJR and ICOW classifications of ex-colonies – are also added to the specification. Their inclusion entails that the estimated relationship between EPGD_1900 and current income levels is attributable to the *incremental* variation in EPGD_1900 relative to the AJR and ICOW measures – that is, to the information it embodies beyond the binary classification of countries into the colonized and non-colonized (for instance, to its measurement of the extent of the "foreignness" of foreign rule, or to the role of dominant domestic minorities).

The inclusion of the state antiquity index described in Section 2.3 is an attempt to address a set of concerns discussed earlier about persistent variation in state-building capabilities across societies. The index provides a proxy for societies' state-building capabilities, at least to the extent that these capabilities have been realized over recorded history in the form of historically-attested state structures. The log of population density in the year 1500 is used (as in much of the prior literature on comparative development) to represent the level of economic development before the era of European empires. Together, these two controls allay to some degree the concern that the relationship between EPGD_1900 and current outcomes is confounded by long-run historical differences across countries in the ability to develop state capacity or economic activity. A set of geographical variables that are relevant for economic development are also included. These are the absolute value of a country's latitude, indicator variables for landlocked countries and for islands, and the distance by air from the nearest of New York, Rotterdam and Tokyo. The log of population is also included to control for potential effects of country size.

The results in Column 5 of Table 4 show that the negative relationship between EPGD_1900 and current income levels is robust to the inclusion of all these control variables (although the sample size falls from 227 to 142 due to missing observations for these controls).

---

[16] For instance, among countries outside Europe that are classified by AJR as never having been colonies, the mean EPGD_1900 is 519, which is quite high in relation to the overall sample mean of 839.

The coefficient is about the same size as that in Column 2 of Table 4, and it remains statistically significant at the 1% level. The robustness of the result to the inclusion of the state antiquity index is particularly important. It suggests that the basic relationship is not being driven by variation in state-building capabilities across societies, at least to the extent that such capabilities are susceptible to quantitative measurement. Moreover, although this is not shown in Table 4, the basic result is virtually identical when controlling instead for a variant of the state antiquity index that only measures state-building history up to 1500 (based on Hariri (2012)); this alternative version of the index avoids the potential influence of state-building activities by European powers since 1500.

3.2) Fiscal Capacity

The basic specification for analyzing the relationship between EPGD_1900 and current fiscal capacity involves an OLS regression of the mean over 2004-2013 of government revenue as a percentage of GDP (the variable described in Section 2.3 above) on the log of (1 + EPGD_1900). The independent variable of interest is thus identical to that described in Section 3.1. The basic results are presented in Table 5 in a stepwise fashion that is closely analogous to that used in Table 4. Column 1 reports the simple bivariate relationship corresponding to Figure 4, and shows that the negative relationship evident in Figure 4 is highly significant. Column 2 adds continent fixed effects; while its magnitude is somewhat smaller, the relationship remains negative and highly significant. All subsequent regressions reported in the paper include continent fixed effects, implying that EPGD_1900 and current fiscal capacity are negatively related *within* each continent.

Column 3 of Table 5 reports the results from the same specification as in Column 2, restricting the sample to those countries that are classified by AJR as ex-colonies. Column 4 of Table 5 reports the corresponding results for the subsample of countries that are classified by AJR as not being ex-colonies. The negative relationship between EPGD_1900 and current fiscal capacity is clearly stronger in the subsample of colonized countries, with a larger coefficient that is statistically significant at the 1% level. In the non-colonized subsample, the estimated coefficient is also negative, albeit somewhat smaller, and is of borderline statistical significance (even though the number of observations in the non-colonized subsample is about half that in the colonized subsample). Thus, the negative relationship between EPGD_1900 and current fiscal capacity that holds among countries generally is by no means absent (even though weaker) among non-colonized countries (as defined by AJR).

Column 5 of Table 5 adds the same set of particularly important control variables as in the specification in Column 5 of Table 4. Especially noteworthy is the inclusion of the state antiquity index. As discussed earlier, this seeks to address what is perhaps the most compelling endogeneity story with respect to the relationships estimated in Table 4 and Table 5: the possibility that unobserved and persistent variation in state-building capabilities across societies leads those countries with weaker capacity both to fall under foreign rule (or the rule of a genetically distant domestic minority) in 1900 and to have lower levels of fiscal capacity today. Moreover, it is possible that imperial powers are generally more likely to colonize countries that are less distant (in genetic and geographical terms). If so, then even among countries that experienced foreign rule, colonization by genetically distant powers (and hence higher values of EPGD_1900) may indicate particularly low levels of fiscal and military capacity, and thus possibly the existence of omitted variables that affect these forms of state capacity.

The results in Column 5 of Table 5 show that the negative relationship between EPGD_1900 and current fiscal capacity is robust to the inclusion of all these control variables (although the sample size falls from 171 to 131 due to missing observations for these controls). The coefficient is quite similar in size to that in Column 2 of Table 4, and it remains statistically significant at the 1% level. As discussed in Section 3.1, the robustness of the result to the inclusion of the state antiquity index is particularly noteworthy. It suggests that the negative relationship between EPGD_1900 and current fiscal capacity is not attributable to variation in state-building capabilities across societies, at least to the extent that these capabilities are captured by the index. Moreover, the basic result is virtually identical when controlling instead for a variant of the state antiquity index (due to Hariri (2012)) that measures state-building history up to 1500.

3.3) Robustness Checks and Extensions

The results in Column 5 of Table 4 (Table 5) reveal a robust negative relationship between EPGD_1900 and current income levels (current fiscal capacity). However, there are a number of other potentially important factors that may confound this type of relationship. These concerns are addressed in the more extensive regression specifications reported in Table 6. For instance, genetic distances between countries may be correlated with the geographical distance between them. Thus, we add to the basic specification the geographical distance variable described in Section 2.3, based on information from the CEPII dataset. This represents the geographical distance from the country's foreign ruler (if any) in 1900, with the variable taking on the value zero for countries that had no foreign ruler in 1900.

Preferences for redistribution may affect countries' observed fiscal capacity, especially when it is measured by revenue as a percentage of GDP. These preferences may themselves be influenced by the internal diversity of the country's population. While this is conceptually distinct from what EPGD_1900 seeks to measure, it is nonetheless a potential concern because internal diversity and EPGD_1900 may be correlated. Thus, we add the ethnolinguistic fractionalization measure constructed by Desmet *et al*. (2012) and described in Section 2.3 to control for the diversity of countries' populations. It is also possible that EPGD_1900 may be correlated with another measure of the internal genetic diversity of countries' populations constructed by Ashraf and Galor (2013) and described in Section 2.3, although again EPGD_1900 is conceptually distinct from their variable. As Ashraf and Galor (2013) argue that their measure affects long-term economic development in an inverse U-shaped, the regression specifications reported in Table 6 control for a quadratic function of the Ashraf and Galor (2013) measure of internal genetic diversity.

It is evident from Column 1 of Table 6 that the strongly negative relationship between EPGD_1900 and current GDP per capita is robust to the inclusion of these additional controls. All of the controls used hitherto are arguably exogenous or at least predetermined with respect to current income levels (and indeed with respect to current fiscal capacity). However, it is possible that current GDP per capita is in part determined by the quality of a country's governance institutions (even though it is also possible that some omitted variable may determine both GDP per capita and governance quality, or even that higher income levels create a demand for better governance and hence affect governance quality), While mindful of its potential endogeneity, we thus add the World Bank governance index described in Section 2.3 to the specification in Column 2 of Table 6. In addition, we include in that specification a series of indicator variables for countries' legal origins, from La Porta *et al*. (2008) (who argue that historical legal origins are associated with the quality of current legal and governance institutions.

The results reported in Column 2 of Table 6 show that the strongly negative relationship between EPGD_1900 and current GDP per capita is robust to the inclusion of these measures of governance quality.[17] The magnitude of the coefficient (about -0.08) is slightly smaller but comparable to that in prior specifications. This estimated coefficient implies that at the mean GDP per capita in our sample (of about $11,000) moving from an EPGD_1900 of zero to the maximum EPGD_1900 in our sample (that for Mozambique) is associated with GDP per capita that is about $5300 lower. This may appear to be a quite large magnitude, but it should be borne in mind that it applies to the maximal variation observed in the sample.

Column 3 of Table 6 reports the results from a model of revenue as a percentage of GDP that adds (to the basic specification in Column 5 of Table 4) the additional variables in Column 1 of Table 6 (geographical distance to a foreign ruler, ethnolinguistic fractionalization, and internal genetic diversity). It also adds the military history variables constructed by Dincecco and Prado (2012): the number of external wars fought, the number of deaths in war, and the number of years during which the country was engaged in war over 1816-1913. These variables capture the idea – widespread in the study of the history of taxation - that external military conflicts have historically played a central role in the development of fiscal capacity. The limited coverage of the military history variables leads to a substantial reduction in the sample size (to 82 countries). However, the strongly negative relationship between EPGD_1900 and current government revenue as a percentage of GDP is robust to the inclusion of these additional controls.

Column 4 of Table 6 adds to the model of revenue as a percentage of GDP the World Bank governance index (notwithstanding its potential endogeneity with respect to government revenue), the log of GDP per capita in 2004 (which may also be potentially endogenous), and legal origins. Again, the basic result is quite robust to the inclusion of these controls.[18] The magnitude of the coefficient (about -1.2) is quite comparable to that in prior specifications. This estimated coefficient implies that moving from an EPGD_1900 of zero to the maximum value in the sample (that for Mozambique) is associated with a nine percentage point lower revenue as a percentage of GDP (relative to a mean revenue of 31% of GDP). While quite substantial, this is well within the range of variation observed in the data – for example, Denmark's government revenue is 56% of GDP while Guatemala's government revenue is 12% of GDP.

---

[17] This relationship also appears to be fairly persistent, at least since the latter part of the twentieth century. The results are very similar when the log of GDP per capita in 1975 (from La Porta *et al*. (2008)) is used as the dependent variable instead of the log of GDP per capita in 2004.

[18] This relationship also appears to be fairly persistent, at least since the latter part of the twentieth century. The results are very similar when the government revenue as a percentage of GDP in 1975 (also from the IMF database) is used as the dependent variable instead of government revenue as a percentage of GDP over 2004-2013.

The results above use the EPGD_1900 measure constructed from the SW dataset of bilateral genetic distances between countries. Recently, Spolaore and Wacziarg (2018) construct a new dataset of these bilateral genetic distances using an alternative genomic dataset based on human microsatellite variation (from Pemberton *et al.* (2013)). As described in Section 2.1, we use this newer data to construct an alternative version of EPGD_1900 and of countries' genetic distance to the UK. Table 7 reports results from models for GDP per capita and revenue as a percentage of GDP that are identical to those in Table 6, except that these alternative versions of EPGD_1900 and genetic distance to the UK are used. The strongly negative relationship between EPGD_1900 and current income levels is robust to using this alternative genomic data, as is the strongly negative relationship between EPGD_1900 and current fiscal capacity.

4) Discussion

The regression analysis reported in Section 3 highlights the robustness of the relationship between EPGD_1900 and current income levels and between EPGD_1900 and current fiscal capacity. Importantly, these relationships are attributable to the incremental variation embodied in EPGD_1900 relative to existing (binary) measures of foreign rule, because the analysis controls for the AJR and ICOW measures. Moreover, the robust conditional correlations documented in this paper - between EPGD_1900 on the one hand and current income levels and fiscal capacity on the other - do not exist for the AJR and ICOW measures of foreign rule. This is evident from Columns 5 of Tables 4 and 5 as well as from Table 6, where the coefficients on the AJR and ICOW variables are not statistically significant. These coefficients are also insignificant when EPGD_1900 is omitted from these specifications (while either the AJR variable, the ICOW variable, or both, are included). Thus, the paper's result that past foreign (or minority) rule is robustly associated with lower levels of current income and fiscal capacity is novel, in the sense that it is not apparent when using existing binary measures of past foreign rule.

Clearly, however, it cannot be claimed that EPGD_1900 was randomly assigned to countries in 1900. Thus, the issue of how to interpret the observed relationships between our proxy for foreign or minority rule in 1900 and countries' contemporary outcomes remains open. As discussed previously, the most compelling alternative interpretation relates to persistent heterogeneity in the unobserved state-building capacity of different societies: this characteristic may explain both why a society succumbed to foreign rule in the recent past and why it exhibits lower levels of economic development and fiscal capacity today. That the observed relationships are robust to controlling for the comprehensive and widely-used state antiquity index (which captures the observable component of each country's state-building capacity) is noteworthy. It leaves open the possibility, however, that there may be unobserved elements of state-building capacity for which we cannot control.

The role of cross-country correlations is not, of course, to establish causal inference, but to suggest hypotheses that may be worthy of further investigation, if possible by means of quasi-experiments. In this instance, there are powerful theoretical and conceptual grounds for believing that the types of factors measured by EPGD_1900 may potentially affect fiscal capacity. Besley and Persson (2009) develop a model in which governments choose to make investments in state capacity that have long-term consequences. These investments include the establishment of revenue-raising institutions and administrative structures, which they term "fiscal capacity investments." In their framework, governments choose larger fiscal capacity investments when political institutions are more inclusive (in the sense of reflecting the interests of a larger fraction of the population).

This idea can be extended straightforwardly to a context in which rulers may be more or less "distant" from the populations they govern, and where ethnic similarity serves a similar function to inclusive political institutions in determining the extent to which these populations' interests are internalized by rulers. Thus, foreign (or other genetically distant) rulers may invest less in developing fiscal capacity, as they do not internalize the benefits of government spending on programs such as mass education to the same extent as would less genetically distant elites. Admittedly, the types of fiscal capacity investments that form the basis of modern fiscal systems were mostly made after 1900, even in developed countries. However, the institutional features that emerge in response to higher values of EPGD_1900 may persist over time, even when the ethnicity of ruling elites changes, and become manifest in later decisions about building state capacity.[19]

[19] Brambor *et al*. (2020) construct an index of an aspect of state capacity that is quite different from our focus on government revenue - the ability of a state to collect information through mechanisms such as a census. Their index of state informational capacity is described in their codebook as: "An aggregate index of information capacity. It is based on a hybrid two-parameter and graded Item Response Model (IRT) that is based on five component indicators – when the country first established a statistical agency . . ., whether the country had in place a civil register . . . and a population register . . . and the graded indexes of census ability . . . and yearbook ability" (see: http://www.stanceatlund.org/information-capacity-dataset.html). Values of their index for 2004 have a strong bivariate negative relationship with EPGD_1900. This relationship is not statistically significant when an extensive set of controls is included; however, the number of countries is quite small (the sample is 66 for the bivariate relationship and substantially smaller when controls are included), so it is difficult to reach any firm conclusions.

A contrary intuition is that foreign (or other genetically distant) rulers may over-invest in revenue-raising ability as a mechanism to extract wealth, as they do not internalize the burdens of taxation experienced by the majority population to the same extent as would less genetically distant elites. Indeed, a very early study of the fiscal consequences of colonialism (Naoroji, 1901) formulates a "fiscal drain" theory which suggests that the population of a country under foreign rule may be taxed excessively to pay high salaries to foreign officials from the ruling country and to support military ventures abroad that primarily benefit the foreign ruler. Ultimately, whether this effect exists (and is larger or smaller than the "inclusiveness" effect described earlier) is an empirical question.

In either event, an alternative channel through which EPGD_1900 may affect fiscal capacity is the willingness of a population to comply with taxation (which is sometimes termed "tax morale"). The willingness to comply may be more limited when rulers are foreign or otherwise genetically distant (perhaps especially in a framework taxation is used as a mechanism to extract wealth, but also more generally). The nineteenth-century statesman Lord Randolph Churchill expressed this idea as follows:

> "The position of India in relation to taxation and the sources of public revenue is very peculiar . . . [due to] the character of the Government which is in the hands of foreigners who hold all the principal administrative offices . . . The impatience of the new taxation, which will have to be borne wholly as a consequence of the foreign rule imposed on the country . . . would constitute a political danger . . . which those responsible for that Government have long regarded as of the most serious order."[20]

Moreover, it is possible that resistance to taxation that initially emerges under foreign rule may persist over time, even when the ethnicity of ruling elites changes; this may constrain the extent to which a country can make subsequent investments in fiscal capacity.

To attempt to test the idea that EPGD_1900 may lower tax morale, we collect data from the 2010-2014 wave of the World Values Survey on the question:[21] "Please tell me for each of the following actions whether you think it can always be justified, never be justified, or something in between: Cheating on taxes if you have a chance." We compute for each of the 57 countries for with WVS data on this question the fraction of respondents who state that "cheating on taxes is never justified." This measure of current tax morale is negatively related to EPGD_1900, although

[20] Quoted in Naoroji (1901, p. xii).
[21] See http://www.worldvaluessurvey.org/WVSContents.jsp.

it is not robustly statistically significant. In view of the small sample size, however, it is difficult to reach any firm conclusions.

The potential impact of past foreign or minority rule on economic development more generally may occur via effects on fiscal capacity. For instance, Dincecco and Prado (2010, 2012) argue that greater fiscal capacity enables more investment in education and thereby increases productivity and income in the long term. Alternatively, there may exist other independent channels. For instance, the extractive institutions emphasized by AJR as being detrimental for long-run development may be more likely to be established when EPGD_1900 is larger.

There is considerable evidence – both quantitative and qualitative – that foreign rule (and its degree of "foreignness") affects important outcomes in various specific contexts. Iyer (2010) analyzes the impact of direct British rule in India, relative to rule by British-protected Indian rulers of the Princely States. This analysis uses as a source of exogenous variation the "Doctrine of Lapse" – a policy implemented in the mid-nineteenth century under which the British annexed to the territory under their direct rule any Princely State in which the ruler died without a natural heir. The "Doctrine of Lapse" thus provides a highly credible source of identification (albeit one that is very specific to mid-nineteenth-century India). Iyer (2010) finds that, when the selection bias associated with the annexation of wealthier regions is controlled for using exogenous variation across Princely States in the existence of a natural heir, direct British rule led to lower levels of public goods provision (including lower levels of education, health and public infrastructure).

Feyrer and Sacerdote (2009) use a dataset of islands and argue that wind patterns that affect maritime travel provide a source of exogenous variation in the timing of European colonization. They find that islands that experienced longer periods of colonial rule have higher income levels, a result that appears to be in tension with the idea that EPGD_1900 is negatively associated with current income levels. However, their instrument addresses the *duration* of colonial rule among islands (all of which were under European rule by 1900), rather than our question about the effect of its "foreignness." Notwithstanding this difference, it is worth noting that the negative relationship between EPGD_1900 and current income reported in Tables 4 and 6 does *not* hold for the subsample consisting only of islands. There are 18 countries and territories that are islands (as defined in Dharmapala and Hines (2009)) with the required data for the specification in Column 5 of Table 4; the relationship between EPGD_1900 and current GDP per capita is positive but statistically insignificant for these islands. This suggests that islands are not necessarily representative of the wider sample of countries in terms of the association between past foreign rule and current income levels.

There is an abundance of qualitative literature in history, area studies, and the social sciences claiming that foreign rule led to deleterious consequences in various particular contexts. Among many such examples is the characterization of the impact of Ottoman rule in the Balkans by Vucinich (1962), which highlights cultural differences between the Ottoman rulers and their Balkan subjects and concludes that (p. 616): “The social and psychological effects of Ottoman rule are apparent in . . . technical and intellectual conservatism, attachment of low social value to work, suspicion of government, economic wastefulness and inefficiency.” These types of claims (which may or may not be justified in any given context) can be viewed as specific instances of the more general relationships analyzed in this paper.

Even if foreign rule is associated with identifiable current outcomes, it may seem that any potential policy implications would be severely limited. While it is certainly true that past history cannot be changed, a greater understanding of the historical development of current institutions and outcomes may nonetheless help to shape policy. In this vein, a potential source of quasi-exogenous variation that might be explored in further investigating our hypothesis arises from several historical instances of the arbitrary territorial division of ethnic groups, often into one segment ruled by a colonial power, and another that remained under some degree of self-rule. Such episodes provide opportunities for using a spatial discontinuity design around the borders of these historical partitions. Such further investigation is arguably warranted by the strength of the robust conditional correlations documented in this paper.

5) Conclusion

The impact of past foreign rule on countries’ current outcomes has been the subject of extensive discussion across many academic disciplines. This paper constructs a novel measure of past foreign or minority rule, based on the genetic distance between ruling elites and the majority of the population in the recent past. It combines historical information on the ethnicity of ruling elites and populations in the year 1900 with existing data on bilateral genetic distances between countries and populations. This measure of “elite-population genetic distance” in 1900 (EPGD_1900) is constructed for 228 present-day countries and territories. The paper documents a strong negative conditional correlation between this measure and current income levels, and between this measure and current fiscal capacity, controlling for various relevant country characteristics, existing measures of foreign rule, the genetic distance of a country's ethnic majority to the UK, and continent fixed effects. The most compelling endogeneity story involves unobserved and persistent variation in state-building capabilities across societies. However, the relationship is robust to controlling for a widely-used index of state antiquity. The paper concludes that the conditional correlation is sufficiently robust to warrant further analysis, using possible sources of exogenous variation that arise in certain specific settings.

References

Acemoglu, D., S. Johnson and J. A. Robinson "The Colonial Origins of Comparative Development: An Empirical Investigation" *American Economic Review*, 91, 1369-1401.

Albouy, D. Y. (2012) "The Colonial Origins of Comparative Development: An Empirical Investigation: Comment" *American Economic Review*, 102, 3059-3076.

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

Besley, T. and T. Persson (2009) "The Origins of State Capacity: Property Rights, Taxation and Politics" *American Economic Review*, 99, 1218-1244.

Besley, T. and T. Persson (2013) "Taxation and Development" in A. Auerbach, R. Chetty, M. Feldstein and E. Saez (eds.) *Handbook of Public Economics*, Vol. 5, North-Holland: Amsterdam and Oxford, pp. 51-110.

Besley, T. and T. Persson (2014) "Why Do Developing Countries Tax So Little?" *Journal of Economic Perspectives*, 28, 99-120.

Bockstette, V., Chanda, A., and L. Putterman (2002) "States and Markets: The Advantage of an Early Start" *Journal of Economic Growth*, 7, 347-369.

Borcan, O., Olsson, O. and L. Putterman (2018) "State history and economic development: Evidence from six millennia" *Journal of Economic Growth*, 23(1), 1-40.

Brambor, T., Goenaga, A., Lindvall, J. and J. Teorell (2020) "The lay of the land: Information capacity and the modern state" *Comparative Political Studies*, 53(2), 175-213.

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

Chanda, A. and L. Putterman (2007) "Early Starts, Reversals and Catch-up in the Process of Economic Development" *Scandinavian Journal of Economics*, 109, 387-413.

Clarence-Smith, W.G. (1985) *The Third Portuguese Empire, 1825-1975: A Study in Economic Imperialism*, Manchester University Press.

Desmet, K., Ortuño-Ortín, I. and R. Wacziarg (2012) "The political economy of linguistic cleavages" *Journal of Development Economics*, 97(2), pp.322-338.

Dharmapala, D. and J. R. Hines, Jr. (2009) "Which Countries Become Tax Havens?" *Journal of Public Economics*, 93, pp. 1058-1068.

Dincecco, M. and J. M. Prado (2010) "Fiscal Capacity, Education and Development" Working Paper.

Dincecco, M. and J. M. Prado (2012) "Warfare, Fiscal Capacity and Performance" *Journal of Economic Growth*, 17, 171-203.

Feyrer, J. and B. Sacerdote (2009) "Colonialism and Modern Income: Islands as Natural Experiments" *Review of Economics and Statistics*, 91, 245-262.

Gallup, J. D., J. D. Sachs and A. Mellinger (1999) "Geography and Economic Development" CID Working Paper No. 1.

Hariri, J. G. (2012) "The autocratic legacy of early statehood" *American Political Science Review*, 106(3), 471-494.

Iyer, L. (2010) "Direct versus Indirect Colonial Rule in India: Long-term Consequences" *Review of Economics and Statistics*, 92, 693-713.

Jelavich, B. (1983) *History of the Balkans*, Cambridge University Press.

Kappeler, A. (2001) *The Russian Empire: A Multi-Ethnic History*, Routledge.

Kaufmann, D., A. Kraay and M. Mastruzzi (2005) "Governance Matters IV: Governance Indicators for 1996-2004" World Bank working paper.

Keen, M. and M. Mansour (2010) "Revenue Mobilization in Sub-Saharan Africa: Challenges from Globalization I—Trade Reform" *Development Policy Review*, 28, 552–71.

Klerman, D. M., P. G. Mahoney, H. Spamann and M. I. Weinstein (2011) "Legal Origin or Colonial History?" *Journal of Legal Analysis*, 3, 379-409.

La Porta, R., Lopez-de-Silanes, F. and A. Shleifer (2008) "The economic consequences of legal origins" *Journal of Economic Literature*, 46(2), 285-332.

Mayer, T. and S. Zignago (2011) "Notes on CEPII's distances measures: The GeoDist database" CEPII Working Paper No. 2011 – 25

Naoroji, D. (1901) *Poverty and Un-British Rule in India*. Swan Sonnenschein & Co: London.

Nunn, N. (2009) "The Importance of History for Economic Development" *Annual Review of Economics*, 1, 65-92.

Okey, R. (2007) *Taming Balkan Nationalism: The Habsburg 'Civilizing Mission' in Bosnia 1878-1914*, Oxford University Press.

Onley, J. (2007) *The Arabian Frontier of the British Raj: Merchants, Rulers, and the British in the Nineteenth-century Gulf*, Oxford University Press.

Pemberton, T. J., DeGiorgio, M., & Rosenberg, N. A. (2013) "Population structure in a comprehensive genomic data set on human microsatellite variation" *G3—Genes/Genomes/Genetics*, 3, 903–919.

Ramusack, B. N. (2004) *The New Cambridge history of India III. 6. The Indian Princes and Their States*, Cambridge University Press, Cambridge.

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

Spolaore, E. and R. Wacziarg (2015) "Ancestry, Language and Culture" CESifo Working Paper No. 5388.

Spolaore, E. and R. Wacziarg (2018) "Ancestry and development: New evidence" *Journal of Applied Econometrics*, 33(5), 748-762.

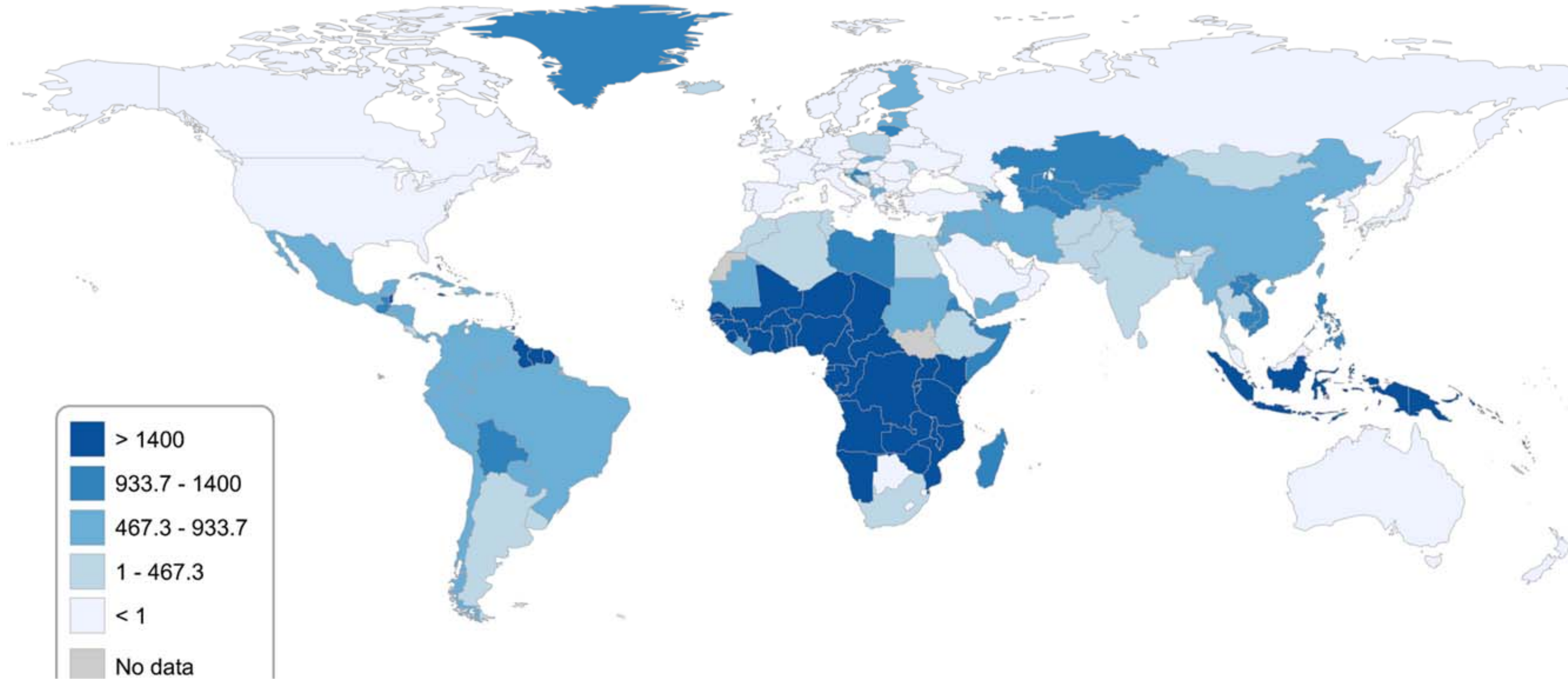
Sugar, P.F. (1963) "The nature of the non-Germanic societies under Habsburg rule" *Slavic Review*, 22(1), 1-30.

Taylor, A.J.P. (1976) *The Habsburg monarchy, 1809-1918: A history of the Austrian Empire and Austria-Hungary*, University of Chicago Press.

Tignor, R.L. (1966) *Modernization and British colonial rule in Egypt, 1882-1914*, Princeton University Press.

Vucinich, W.S. (1962) "The nature of Balkan society under Ottoman rule" *Slavic Review*, 21(4), 597-616.

Figure 1: Map of EPGD_1900

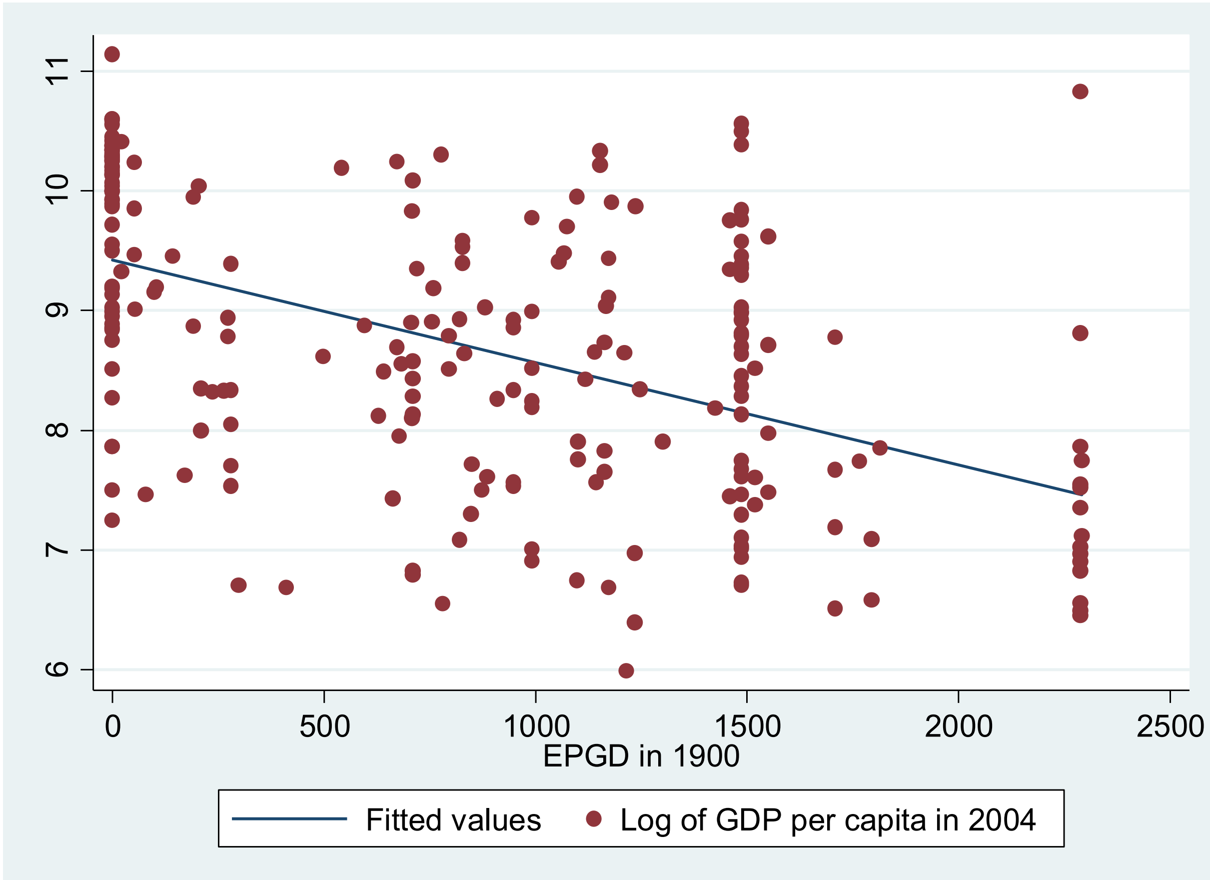


Note: This map depicts the variation in the values of EPGD_1900 across the countries in the sample. EPGD_1900 is a measure of the genetic distance between a country's ruling elite and its ethnic majority in the year 1900, based on bilateral genetic distances between country-pairs reported in Spolaore and Wacziarg (2009).

Figure 2: The Relationship Between EPGD_1900 and Existing Measures of Foreign Rule

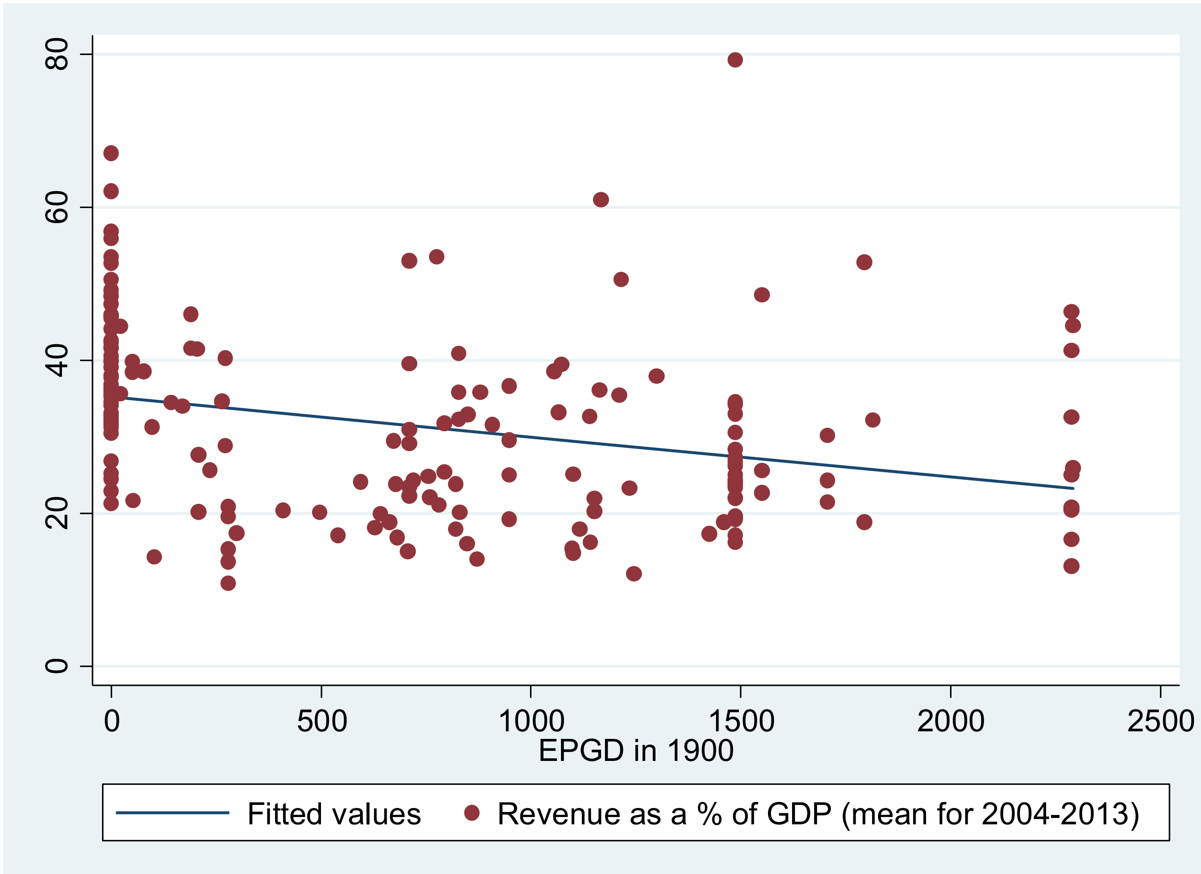
Note: This graph depicts component loadings from a principal components analysis of three different measures of past foreign rule. EPGD_1900 is a measure of the genetic distance between a country's ruling elite and its ethnic majority in the year 1900, based on bilateral genetic distances between country-pairs reported in Spolaore and Wacziarg (2009). The AJR ex-colony variable is a binary classification of countries formerly under foreign rule from Acemoglu *et al*. (2001). The ICOW measure is also a binary classification of countries formerly under foreign rule, based on the Issue Correlates of War (ICOW) colonial history database constructed by Paul Hensel.

Figure 3: EPGD_1900 and Current GDP per capita



Note: This bivariate plot depicts the relationship between EPGD_1900 - a measure of the genetic distance between a country's ruling elite and its ethnic majority in the year 1900, based on bilateral genetic distances between country-pairs reported in Spolaore and Wacziarg (2009) - and the log of GDP per capita (in US$ measured in PPP terms) in 2004 (from Dharmapala and Hines (2009)).

Figure 4: EPGD and Current Government Revenue as a % of GDP



Note: This bivariate plot depicts the relationship between EPGD_1900 - a measure of the genetic distance between a country's ruling elite and its ethnic majority in the year 1900, based on bilateral genetic distances between country-pairs reported in Spolaore and Wacziarg (2009) - and government revenue as a percentage of GDP averaged over 2004-2013 (calculated based on the IMF's database on government revenues).

Table 1: Descriptive Statistics for EPGD_1900 and Related Variables

		Mean	Standard Deviation	Number of Observations
EPGD_1900 and Variants	EPGD_1900	839.4802	700.5765	228
	EPGD_1900 with alternative values for the Baltic region	842.8361	701.0391	228
	EPGD_1900 with alternative values for some Eastern Mediterranean countries	846.7916	696.8572	228
	EPGD_1900 with alternative values for the Gulf region	850.0109	691.696	228
	EPGD_1900 with alternative values for Southern Africa	869.4934	713.0723	228
EPGD_1900 by Continent	Africa	1409.37	734.638	56
	Americas	1017.374	516.8813	52
	Asia	524.6939	449.4694	49
	Europe	166.1034	329.7734	47
	Oceania	1085.688	508.6373	24
Genetic Distance to the UK		915.8584	706.2877	228
Alternative Measure using Microsatellite Variation	EPGD_1900	.0236721	.0216022	193
	Genetic distance to the UK	.0288951	.0202098	189
Existing Measures of Foreign Rule	AJR Ex-Colony Variable	.6708075	.4713862	161
	ICOW Measure	.8489583	.3590257	192
EPGD_1900 by Primary Former Ruler, Wave of Colonization, and Indirect Rule	British (ICOW definition)	921.3351	804.5552	59
	French (ICOW definition)	1351.097	605.8128	24
	Spanish (ICOW definition)	779.1528	478.2433	20
	Early Colonization (ICOW)	434.3129	402.0137	30
	Later Colonization (ICOW)	1028.936	721.7973	133
	Direct British Rule (Lange, 2004)	999.1667	613.4574	12
	Indirect British Rule (Lange, 2004)	1208.287	899.4674	21

Note: This table reports descriptive statistics for EPGD_1900 (the measure of the genetic distance between a country's ruling elite and its ethnic majority in the year 1900, based on bilateral genetic distances between country-pairs reported in Spolaore and Wacziarg (2009)) and for several related variables. The variants of EPGD_1900 involve alternative classifications of the ruling elites of certain countries in the Baltic region, the Eastern Mediterranean, the Gulf, and southern Africa, as detailed in the Data Appendix. The classification of countries by continent is based on Borcan *et al*. (2018). Genetic distance to the UK measures the bilateral genetic distance between the population of a country and that of the UK, based on the Spolaore and Wacziarg (2009) data. The alternative genetic distance measures using microsatellite variation are based on bilateral genetic distances between country-pairs reported in Spolaore and Wacziarg (2018). The AJR ex-colony variable is a binary classification of countries formerly under foreign rule from Acemoglu *et al*. (2001). The ICOW measure is also a binary classification of countries formerly under foreign rule, based on the Issue Correlates of War (ICOW) colonial history database constructed by Paul Hensel. The "primary" foreign ruler and the wave of colonization (based on the year of independence) is from the ICOW colonial history database. The classification of direct versus indirect British rule is from Lange (2004), who constructs an index that represents the number of court cases recognizing customary law in 1955 relative to the total number of court cases in that territory in that year.

Table 2: Correlation Matrix for EPGD_1900 and Existing Measures of Foreign Rule

	EPGD_1900	AJR Ex-Colony Variable	ICOW Measure
EPGD_1900	1		
AJR Ex-Colony Variable	0.4711	1	
ICOW Measure	0.3462	0.5225	1

Note: This table reports a correlation matrix for three measures of past foreign rule. EPGD_1900 is a measure of the genetic distance between a country's ruling elite and its ethnic majority in the year 1900, based on bilateral genetic distances between country-pairs reported in Spolaore and Wacziarg (2009). The AJR ex-colony variable is a binary classification of countries formerly under foreign rule from Acemoglu *et al*. (2001). The ICOW measure is also a binary classification of countries formerly under foreign rule, based on the Issue Correlates of War (ICOW) colonial history database constructed by Paul Hensel.

Table 3: Descriptive Statistics for Other Variables

		Mean	Standard Deviation	Number of Observations
Outcome Variables	GDP per capita (PPP; thousands of US$)	11089.5	11577.17	227
	Revenue as a % of GDP (mean over 2004-2013)	31.20292	12.17263	171
	Revenue as a % of GDP (max over 2004-2013)	36.75789	17.19132	171
Country Characteristics	Log of Population (in thousands)	14.74616	2.68306	228
	Log of Population Density in 1500	0.9053197	1.460571	154
	Ethnolinguistic Fractionalization	0.1577771	0.18014	223
	Geographical Distance to Foreign Ruler	4241.793	4700.735	219
	Internal Genetic Diversity (Ashraf and Galor, 2013)	0.7008106	0.0557446	207
	Latitude (Absolute Value)	0.2802764	0.1888412	207
	Landlocked (=1)	0.1798246	0.3848862	228
	Island (=1)	0.1798246	0.3848862	228
	Distance by Air from New York, Rotterdam or Tokyo	4174.434	2595.052	228
State Antiquity	State Antiquity Index (Borcan et al., 2018)	0.2210557	0.1663684	152
	State Antiquity in 1500 (Hariri, 2012)	462.2249	454.4507	151
Governance Index		0.0041308	0.9280716	209
Legal Origins	British	0.3333333	0.4725473	207
	French	0.5458937	0.4990963	207
	German	0.0966184	0.2961536	207
	Scandinavian	0.0241546	0.1539012	207
Military History	War Deaths, 1816-1913	0.1019138	0.2646616	96
	Number of Wars, 1816-1913	3.572917	5.687928	96
	War Years, 1816-1913	15.04167	22.24169	96

Continent	Africa	0.245614	0.4313978	228
Dummies	Americas	0.2280702	0.4205113	228
	Asia	0.2149123	0.4116652	228
	Europe	0.2061404	0.4054225	228
	Oceania	0.1052632	0.3075674	228

Note: This table reports descriptive statistics for the outcome and control variables used in the regression analysis. The outcome variables are GDP per capita and revenue as a percentage of GDP. GDP per capita is measured in thousands of U.S. dollars in purchasing power parity (PPP) terms for 2004 (from Dharmapala and Hines (2009)). Government revenue as a percentage of GDP (calculated based on the IMF's database on government revenues) is averaged over 2004-2013. An alternative measure – the maximum value of government revenue as a percentage of GDP over 2004-2013 – is used in robustness checks. Control variables used in the regression analysis are the following. Population is measured in thousands of residents in 2004 (from Dharmapala and Hines (2009)). The log of population density in 1500 is from Borcan *et al.* (2018). Ethnolinguistic fractionalization (a measure of the probability that two randomly chosen individuals in a given country speak a different language) is from Desmet *et al*. (2012). Geographical distance to a foreign ruler in 1900 is constructed using the Centre d'Etudes Prospectives et D'Informations Internationale (CEPII) dataset described in Mayer and Zignago (2011). Internal genetic diversity is a measure of the probability that two randomly selected individuals in a population differ genetically with respect to a spectrum of traits, from Ashraf and Galor (2013). The absolute value of countries' latitude is from Gallup, Sachs and Mellinger (1999). The indicator variable for landlocked countries, and the distance by air (measured in kilometers) from the nearest of New York, Rotterdam and Tokyo are from Dharmapala and Hines (2009) and represent extended versions of variables constructed by Gallup, Sachs and Mellinger (1999). The indicator variable for islands is from Dharmapala and Hines (2009). The index of state antiquity is from Borcan *et al*. (2018) and represents an updated version of the measure constructed by Bockstette *et al*. (2002) and Chanda and Putterman (2007) by assigning countries a score for each 50-year period of history based on the existence of a large-scale state and the extent of independent local control. The baseline analysis uses an ancestry-adjusted version of the index that discounts past state-building history at a 1% discount rate. An alternative version used in robustness checks only includes state history up to 1500, and is based on Hariri (2012). The governance index (for 2004) is constructed by Kaufmann *et al.* (2005), taking values roughly in the (-2.5, 2.5) interval, with a zero mean and unit variance in the whole sample, with higher values corresponding to better governance. Indicators for countries' legal origins are from La Porta *et al*. (2008). The number of wars fought, the number of deaths in war, and the number of years during which the country was engaged in war (all for external wars over 1816-1913) are from Dincecco and Prado (2012). The classification of countries by continent is based on Borcan *et al*. (2018).

Table 4: The Relationship between EPGD_1900 and Current GDP per Capita

	(1) Full Sample	(2) Full Sample	(3) Colonized Countries	(4) Non-Colonized Countries	(5) Full Sample
	Dependent Variable: Log of GDP per capita				
Log of EPGD_1900	**-0.210*** **	**-0.145*** **	**-0.211*** **	**-0.102** **	**-0.138*** **
	(0.0208)	**(0.0227)**	**(0.0341)**	**(0.0391)**	**(0.0275)**
Log of Genetic Distance					-0.0362
from the UK					(0.0403)
Ex-Colony (AJR)					0.137
					(0.297)
Ex-Colony (ICOW)					-0.267
					(0.194)
Log of Population					-0.0999*
					(0.0496)
State Antiquity					0.513
					(0.457)
Log of Population Density					-0.0677
in 1500					(0.0635)
Latitude (Absolute Value)					1.248
					(0.760)
Landlocked (=1)					-0.482***
					(0.148)
Island (=1)					0.312
					(0.238)
Distance by Air					-0.000
					(0.000)
Continent Fixed Effects?	N	Y	Y	Y	Y
Constant	9.779***	10.09***	10.01***	7.233***	11.12***
	(0.113)	(0.112)	(0.211)	(0.261)	(1.107)
Observations	227	227	108	53	142
R-squared	0.279	0.484	0.528	0.501	0.736

Note: This table reports the results of regression models of the log of GDP per capita, measured in thousands of US dollars in purchasing power parity (PPP) terms for 2004 (from Dharmapala and Hines (2009)). EPGD_1900 is a measure of the genetic distance between a country's ruling elite and its ethnic majority in the year 1900, based on bilateral genetic distances between country-pairs reported in Spolaore and Wacziarg (2009). Genetic distance to the UK measures the bilateral genetic distance between the population of a country and that of the UK, based on the Spolaore and Wacziarg (2009) data. The AJR ex-colony variable is a binary classification of countries formerly under foreign rule from Acemoglu *et al*. (2001). The ICOW measure is also a binary classification of countries formerly under foreign rule, based on the Issue Correlates of War (ICOW) colonial history database constructed by Paul Hensel. Population is measured in thousands of residents in 2004 (from Dharmapala and Hines (2009)). The log of population density in 1500 is from Borcan *et al.* (2018). The absolute value of countries' latitude is from Gallup, Sachs and Mellinger (1999). The index of state antiquity is from Borcan *et al*. (2018) and represents an updated version of the measure constructed by Bockstette *et al*. (2002) and Chanda and Putterman (2007) by assigning countries a score for each 50-year period of history based on the existence of a large-scale state and the extent of independent local control. The baseline analysis uses an ancestry-adjusted version of the index that discounts past state-building history at a 1% discount rate. The indicator variable for landlocked countries, and the distance by air (measured in kilometers) from the nearest of New York, Rotterdam and Tokyo are from Dharmapala and Hines (2009) and represent extended versions of variables constructed by Gallup, Sachs and Mellinger (1999). The indicator variable for islands is from Dharmapala and Hines (2009). The classification of countries by continent is based on Borcan *et al*. (2018). Robust standard errors are in parentheses; * significant at 10%; ** significant at 5%; *** significant at 1%

Table 5: The Relationship between EPGD_1900 and Current Fiscal Capacity

	(1) Full Sample	(2) Full Sample	(3) Colonized Countries	(4) Non-Colonized Countries	(5) Full Sample
	Dependent Variable: Revenue as a % of GDP				
Log of EPGD_1900	**-1.805*** **	**-1.332*** **	**-2.087*** **	**-0.663***	**-1.075*** **
	(0.268)	**(0.326)**	**(0.488)**	**(0.331)**	**(0.404)**
Log of Genetic Distance					0.258
from the UK					(0.491)
Ex-Colony (AJR)					5.279+
					(3.130)
Ex-Colony (ICOW)					-3.376+
					(1.951)
Log of Population					-1.452***
					(0.537)
State Antiquity					6.405
					(6.145)
Log of Population Density					-2.377***
in 1500					(0.762)
Latitude (Absolute Value)					14.85*
					(7.066)
Landlocked (=1)					-2.667
					(1.772)
Island (=1)					-3.241
					(2.015)
Distance by Air					-0.000241
					(0.00046)
Continent Fixed Effects?	N	Y	Y	Y	Y
Constant	40.03***	43.95***	48.05***	25.42***	64.96***
	(1.464)	(1.414)	(1.930)	(2.202)	(10.37)
Observations	171	171	97	51	131
R-squared	0.204	0.300	0.224	0.618	0.598

Note: This table reports the results of regression models of government revenue as a percentage of GDP (calculated based on the IMF's database on government revenues), averaged over 2004-2013. EPGD_1900 is a measure of the genetic distance between a country's ruling elite and its ethnic majority in the year 1900, based on bilateral genetic distances between country-pairs reported in Spolaore and Wacziarg (2009). Genetic distance to the UK measures the bilateral genetic distance between the population of a country and that of the UK, based on the Spolaore and Wacziarg (2009) data. The AJR ex-colony variable is a binary classification of countries formerly under foreign rule from Acemoglu *et al*. (2001). The ICOW measure is also a binary classification of countries formerly under foreign rule, based on the Issue Correlates of War (ICOW) colonial history database constructed by Paul Hensel. Population is measured in thousands of residents in 2004 (from Dharmapala and Hines (2009)). The log of population density in 1500 is from Borcan *et al.* (2018). The absolute value of countries' latitude is from Gallup, Sachs and Mellinger (1999). The index of state antiquity is from Borcan *et al.* (2018) and represents an updated version of the measure constructed by Bockstette *et al.* (2002) and Chanda and Putterman (2007) by assigning countries a score for each 50-year period of history based on the existence of a large-scale state and the extent of independent local control. The baseline analysis uses an ancestry-adjusted version of the index that discounts past state-building history at a 1% discount rate. The indicator variable for landlocked countries, and the distance by air (measured in kilometers) from the nearest of New York, Rotterdam and Tokyo are from Dharmapala and Hines (2009) and represent extended versions of variables constructed by Gallup, Sachs and Mellinger (1999). The indicator variable for islands is from Dharmapala and Hines (2009). The classification of countries by continent is based on Borcan *et al.* (2018). Robust standard errors are in parentheses; * significant at 10%; ** significant at 5%; *** significant at 1%

Table 6: Robustness Checks

	(1)	(2)	(3)	(4)
	Dependent Variable: Log of GDP per capita		Dependent Variable: Revenue as a % of GDP	
Log of EPGD_1900	**-0.142*** **	**-0.0846*** **	**-1.351** **	**-1.172** **
	(0.0295)	**(0.0233)**	**(0.545)**	**(0.510)**
Log of Genetic Distance	-0.0548	0.0110	-0.324	0.0557
from the UK	(0.0391)	(0.0280)	(0.585)	(0.606)
Ex-Colony (AJR)	0.152	-0.0771	5.387	1.779
	(0.323)	(0.223)	(4.481)	(4.200)
Ex-Colony (ICOW)	-0.203	0.0424	-5.778**	-1.604
	(0.183)	(0.124)	(2.576)	(3.038)
Log of Population	-0.114**	-0.0286	-1.604	-0.863
	(0.0510)	(0.0394)	(1.034)	(1.041)
State Antiquity	0.574	0.601	-8.432	-6.361
	(0.533)	(0.379)	(7.638)	(8.395)
Log of Population Density	-0.0437	-0.0646	-1.720**	-1.648*
in 1500	(0.0613)	(0.0438)	(0.846)	(0.877)
Latitude (Absolute Value)	1.527**	0.730	3.578	-4.629
	(0.698)	(0.506)	(9.552)	(10.24)
Landlocked (=1)	-0.469***	-0.317**	-0.601	1.741
	(0.156)	(0.129)	(2.615)	(2.753)
Island (=1)	0.205	0.0759	-3.891	-3.878
	(0.205)	(0.146)	(3.759)	(3.598)
Distance by Air	-0.00008*	-0.00008**	-0.00112	-0.00114
	(0.00004)	(0.00004)	(0.00136)	(0.00153)
Geographical Distance from	0.000006	0.000004	0.000170	5.98e-05
Foreign Ruler	(0.00002)	(0.00002)	(0.000535)	(0.000564)
Ethnolinguistic Fractionalization	0.584	0.396	3.029	0.896
	(0.362)	(0.307)	(7.526)	(8.708)
Internal Genetic Diversity	-34.51	-59.36	-1,958	-2,369*
	(62.67)	(49.61)	(1,286)	(1,335)
Internal Genetic Diversity Squared	16.60	39.76	1,418	1,749*
	(45.72)	(35.82)	(925.9)	(953.4)
Governance Index		0.619***		0.644
		(0.0759)		(2.466)
Log of GDP per capita				1.031
				(2.334)
Continent Fixed Effects?	Y	Y	Y	Y
Legal Origins?	N	Y	N	Y
Military History Variables?	N	N	Y	Y
Constant	26.71	31.04*	753.2*	861.8*
	(21.89)	(17.47)	(446.1)	(473.8)
Observations	141	141	82	82
R-squared	0.759	0.846	0.707	0.737

Note: Columns 1 and 2 of this table report the results of regression models of the log of GDP per capita, measured in thousands of US dollars in purchasing power parity (PPP) terms for 2004 (from Dharmapala and Hines (2009)). Columns 3 and 4 of this table report the results of regression models of government revenue as a percentage of GDP (calculated based on the IMF's database on government revenues), averaged over 2004-2013. EPGD_1900 is a measure of the genetic distance between a country's ruling elite and its ethnic majority in the year 1900, based on bilateral genetic distances between country-pairs reported in Spolaore and Wacziarg (2009). Genetic distance to the UK measures the bilateral genetic distance between the population of a country and that of the UK, based on the Spolaore and Wacziarg (2009) data. The AJR ex-colony variable is a binary classification of countries formerly under foreign rule from Acemoglu *et al*. (2001). The ICOW measure is also a binary classification of countries formerly under foreign rule, based on the Issue Correlates of War (ICOW) colonial history database constructed by Paul Hensel. Population is measured in thousands of residents in 2004 (from Dharmapala and Hines (2009)). The log of population density in 1500 is from Borcan *et al.* (2018). The absolute value of countries' latitude is from Gallup, Sachs and Mellinger (1999). The index of state antiquity is from Borcan *et al*. (2018) and represents an updated version of the measure constructed by Bockstette *et al*. (2002) and Chanda and Putterman (2007) by assigning countries a score for each 50-year period of history based on the existence of a large-scale state and the extent of independent local control. The baseline analysis uses an ancestry-adjusted version of the index that discounts past state-building history at a 1% discount rate. The indicator variable for landlocked countries, and the distance by air (measured in kilometers) from the nearest of New York, Rotterdam and Tokyo are from Dharmapala and Hines (2009) and represent extended versions of variables constructed by Gallup, Sachs and Mellinger (1999). The indicator variable for islands is from Dharmapala and Hines (2009). Ethnolinguistic fractionalization (a measure of the probability that two randomly chosen individuals in a given country speak a different language) is from Desmet *et al*. (2012). Geographical distance to a foreign ruler in 1900 is constructed using the Centre d'Etudes Prospectives et D'Informations Internationale (CEPII) dataset described in Mayer and Zignago (2011). Internal genetic diversity is a measure of the probability that two randomly selected individuals in a population differ genetically with respect to a spectrum of traits, from Ashraf and Galor (2013). The military history variables are the number of wars fought, the number of deaths in war, and the number of years during which the country was engaged in war (all for external wars over 1816-1913), from Dincecco and Prado (2012). Indicators for countries' legal origins are from La Porta *et al*. (2008). The classification of countries by continent is based on Borcan *et al*. (2018). Robust standard errors are in parentheses; * significant at 10%; ** significant at 5%; *** significant at 1%

Table 7: Results using an Alternative Measure of Genetic Distance

	(1)	(2)	(3)	(4)
	Dependent Variable: Log of GDP per capita		Dependent Variable: Revenue as a % of GDP	
Log of EPGD_1900	**-25.42*** **	**-18.71*** **	**-483.6*** **	**-438.8** **
(Alternative Measure)	**(7.741)**	**(5.548)**	**(177.2)**	**(169.8)**
Log of Genetic Distance	-5.517	1.090	350.2*	398.9**
from the UK (Alternative Measure)	(7.713)	(5.193)	(176.7)	(169.8)
Ex-Colony (AJR)	0.242	-0.110	4.722	0.114
	(0.337)	(0.216)	(4.531)	(4.510)
Ex-Colony (ICOW)	-0.434**	-0.0425	-6.474**	-0.657
	(0.188)	(0.133)	(2.721)	(3.093)
Log of Population	-0.0960**	-0.0228	-0.889	-0.185
	(0.0468)	(0.0353)	(0.964)	(0.899)
State Antiquity	0.491	0.458	-3.167	-2.027
	(0.570)	(0.375)	(8.452)	(9.184)
Log of Population Density	-0.0329	-0.0530	-1.553*	-1.409
in 1500	(0.0634)	(0.0435)	(0.911)	(0.891)
Latitude (Absolute Value)	0.827	0.0557	6.195	-2.730
	(0.701)	(0.481)	(10.22)	(10.68)
Landlocked (=1)	-0.387**	-0.277**	1.383	4.139
	(0.163)	(0.126)	(2.543)	(2.626)
Island (=1)	0.206	0.0135	-2.792	-3.455
	(0.226)	(0.157)	(3.101)	(3.170)
Distance by Air	-6.75e-05	-7.81e-05*	-0.00146	-0.00134
	(4.55e-05)	(3.96e-05)	(0.00105)	(0.00120)
Geographical Distance from	1.96e-05	2.44e-05	0.000832*	0.000734*
Foreign Ruler	(2.20e-05)	(1.50e-05)	(0.000433)	(0.000418)
Ethnolinguistic Fractionalization	0.666*	0.473	10.26	6.859
	(0.342)	(0.300)	(6.591)	(7.393)
Internal Genetic Diversity	27.50	-38.77	-1,989*	-2,517**
	(64.63)	(47.49)	(1,068)	(1,100)
Internal Genetic Diversity Squared	-29.54	25.27	1,485*	1,918**
	(47.19)	(34.33)	(767.2)	(777.9)
Governance Index		0.652***		0.969
		(0.0751)		(2.314)
Log of GDP per capita				1.906
				(2.107)
Continent Fixed Effects?	Y	Y	Y	Y
Legal Origins?	N	Y	N	Y
Military History Variables?	N	N	Y	Y
Constant	6.948	24.68	719.5*	854.1**
	(22.27)	(16.82)	(371.1)	(397.2)
Observations	138	138	81	81
R-squared	0.745	0.854	0.725	0.776

Note: Columns 1 and 2 of this table report the results of regression models of the log of GDP per capita, measured in thousands of US dollars in purchasing power parity (PPP) terms for 2004 (from Dharmapala and Hines (2009)). Columns 3 and 4 of this table report the results of regression models of government revenue as a percentage of GDP (calculated based on the IMF's database on government revenues), averaged over 2004-2013. EPGD_1900 is a measure of the genetic distance between a country's ruling elite and its ethnic majority in the year 1900, based on bilateral genetic distances between country-pairs reported in Spolaore and Wacziarg (2018) that use the alternative genetic distance measures using microsatellite variation. Genetic distance to the UK measures the bilateral genetic distance between the population of a country and that of the UK, based on the Spolaore and Wacziarg (2018) data using the alternative genetic distance measures using microsatellite variation. The AJR ex-colony variable is a binary classification of countries formerly under foreign rule from Acemoglu *et al*. (2001). The ICOW measure is also a binary classification of countries formerly under foreign rule, based on the Issue Correlates of War (ICOW) colonial history database constructed by Paul Hensel. Population is measured in thousands of residents in 2004 (from Dharmapala and Hines (2009)). The log of population density in 1500 is from Borcan *et al.* (2018). The absolute value of countries' latitude is from Gallup, Sachs and Mellinger (1999). The index of state antiquity is from Borcan *et al*. (2018) and represents an updated version of the measure constructed by Bockstette *et al*. (2002) and Chanda and Putterman (2007) by assigning countries a score for each 50-year period of history based on the existence of a large-scale state and the extent of independent local control. The baseline analysis uses an ancestry-adjusted version of the index that discounts past state-building history at a 1% discount rate. The indicator variable for landlocked countries, and the distance by air (measured in kilometers) from the nearest of New York, Rotterdam and Tokyo are from Dharmapala and Hines (2009) and represent extended versions of variables constructed by Gallup, Sachs and Mellinger (1999). The indicator variable for islands is from Dharmapala and Hines (2009). Ethnolinguistic fractionalization (a measure of the probability that two randomly chosen individuals in a given country speak a different language) is from Desmet *et al*. (2012). Geographical distance to a foreign ruler in 1900 is constructed using the Centre d'Etudes Prospectives et D'Informations Internationale (CEPII) dataset described in Mayer and Zignago (2011). Internal genetic diversity is a measure of the probability that two randomly selected individuals in a population differ genetically with respect to a spectrum of traits, from Ashraf and Galor (2013). The military history variables are the number of wars fought, the number of deaths in war, and the number of years during which the country was engaged in war (all for external wars over 1816-1913), from Dincecco and Prado (2012). Indicators for countries' legal origins are from La Porta *et al*. (2008). The classification of countries by continent is based on Borcan *et al*. (2018). Robust standard errors are in parentheses; * significant at 10%; ** significant at 5%; *** significant at 1%

Data Appendix: Values of EPGD_1900

Country	Ethnicity of the Ruling Elite in 1900	Ethnic Majority in 1900	EPGD_1900	Alternative Measure
Afghanistan	Pashtun	Tajik and Hazara (in and around Kabul)	410.5	.00907
Albania	Ottoman Turkish	Albanian	794	.00689
Algeria	French	North African	273	.01411
American Samoa	US officials, predominantly of British or other Northern European descent	Samoan	991	
Andorra	Catalan and French	Catalan	0	
Angola	Portuguese	African	2292	.05145
Anguilla	British	Of African descent	1487	
Antigua and Barbuda	British	Of African descent	1487	.05131
Argentina	Predominantly of Spanish descent	Of Spanish, Italian, other European and indigenous (Quechua, Guarani, Mapuche) descent	142.667	.0097582
Armenia	Russian	Armenian	208	.01312
Aruba	Dutch	Of Spanish, indigenous (Arawak) and African descent	672.23	
Australia	British and Irish	British and Irish	0	0
Austria	German	German	0	0
Azerbaijan	Russian	Azeri Turkish	949	.01312
Bahamas, The	British	Of African descent	1487	.05131
Bahrain	Arab	Arab	0	0
Alternative coding	British	Arab	273	
Bangladesh	British	Bengali	280	.0165
Barbados	British	Of African descent	1487	.05131
Belarus	Russian	Belarussian	0	0
Belgium	Flemish and Walloon	Flemish and Walloon	0	0
Belize	British	Of African descent	1487	.05738
Benin	French	African	1487	.04773
Bermuda	British	Of African descent	1487	
Bhutan	Of Tibetan origin	Of Tibetan origin	0	0
Bolivia	Predominantly of Spanish descent	Of indigenous (Quechua, Aymara) descent	1300	.06006
Bosnia and Herzegovina	German	Bosnian	191	.00607
Alternative coding	Ottoman Turkish	Bosnian	794	
Botswana	Tswana	Tswana	0	0
Alternative coding	British	Tswana	2288	
Brazil	Predominantly of Portuguese descent	Of Portuguese, African and indigenous (Tupi-Guarani) descent	879.875	.0244738
British Virgin Islands	British	Of African descent	1487	
Brunei	Malay	Malay	0	0
Alternative coding	British	Malay	1275	
Bulgaria	Bulgarian	Bulgarian	0	0
Burkina Faso	French	African	1487	.04773

Burundi	German	Hutu and Tutsi	1708	.04487
Cambodia	French	Khmer	1100	.03496
Cameroon	German	African	1708	.04321
Canada	British and French	British and French	0	0
Cape Verde	Portuguese	Of African and Portuguese descent	1139.83	.032653
Cayman Islands	British	Of African descent	1487	
Central African Republic	French	African	2288	.04797
Chad	French	African	1767	.04608
Channel Islands	Channel Islander	Channel Islander	0	0
Chile	Predominantly of Spanish descent	Of Spanish and indigenous (Mapuche) descent	720.349	.0333169
China	Manchu	Chinese	498	.00745
Colombia	Predominantly of Spanish descent	Of Spanish, indigenous (Chibcha) and African descent	595.201	.0256232
Comoros	French	Comorian	2288	.04321
Congo, Dem. Rep.	Flemish and Walloon	African	2288	.05145
Congo, Rep.	French	African	2288	.04321
Cook Islands	British	Polynesian	991	
Costa Rica	Predominantly of Spanish descent	Of Spanish and indigenous (Chibcha) descent	103.534	.0043142
Cote d'Ivoire	French	African	1487	.04773
Croatia	Hungarian	Croatian	1055	.01123
Cuba	US officials, predominantly of British or other Northern European descent	Of Spanish, African and indigenous (Taino) descent	707.766	.0244444
Cyprus	British	Greek	204	.00607
Alternative coding	Ottoman Turkish	Greek	794	
Czech Republic	German	Czech	0	0
Denmark	Danish	Danish	0	0
Djibouti	French	Somali and Afar	1163	.0344
Dominica	British	Of African descent	1487	.05131
Dominican Republic	Predominantly of Spanish descent	Of Spanish, African and indigenous (Taino) descent	705.889	.0245193
Ecuador	Predominantly of Spanish descent	Of Spanish and indigenous (Quechua) descent	909.418	.0409002
Egypt, Arab Rep.	British	Egyptian	236	.01411
Alternative coding	Ottoman Turkish	Egyptian	710	
El Salvador	Predominantly of Spanish descent	Of Spanish and indigenous (Pipil) descent	682.547	.0314372
Equatorial Guinea	Spanish	African	2288	.05219
Eritrea	Italian	Tigrinya	1234	.03351
Estonia	Baltic German	Estonian	828	.00946
Alternative coding	Russian	Estonian	1066	
Ethiopia	Amhara	Oromo (in and around Addis Ababa)	298.104	.0128577
Faeroe Islands	Danish	Faeroese	0	
Falkland Islands	British	British	0	
Fiji	British	Fijian	1550	.05395
Finland	Finland-Swedish	Finnish	776.86	.0088813
Alternative coding	Russian	Finnish	1066	
France	French	French	0	0

French Guiana	French	Of African descent	1487	
French Polynesia	French	Polynesia	991	
Gabon	French	African	2288	.05219
Gambia, The	British	African	1487	.04773
Georgia	Russian	Georgian	208	.01312
Germany	German	German	0	0
Ghana	British	African	1487	.05378
Gibraltar	British	Gibraltarian	51	
Greece	Greek	Greek	0	0
Greenland	Danish	Greenlander (Inuit)	1180	
Grenada	British	Of African descent	1487	.05131
Guadeloupe	French	Of African descent	1487	
Guam	US officials, predominantly of British or other Northern European descent	Chamorro	1098	
Guatemala	Predominantly of Spanish descent	Maya	1246	.05738
Guinea	French	African	1487	.04773
Guinea-Bissau	Portuguese	African	1794	.04773
Guyana	British and Dutch	Of African descent	1487	.05131
Haiti	Of French and African descent	Of African descent	662.511	.0190745
Honduras	Predominantly of Spanish descent	Of Spanish and indigenous (Lenca) descent	678.078	.0308787
Hong Kong, China	British	Chinese	1152	.03964
Hungary	Hungarian	Hungarian	0	0
Iceland	Danish	Icelandic	21	0
India	British	Hindustani (in and around the capital New Delhi)	280	.01631
Indonesia	Dutch	Indonesian	1425	.03496
Iran, Islamic Rep.	Qajar (Azeri Turkish)	Persian	821	.01814
Iraq	Ottoman Turkish	Iraqi	710	.0087
Ireland	British and Anglo-Irish	Irish	0	0
Isle of Man	Manx	Manx	0	
Israel	Ottoman Turkish	Palestinian	710	.00413
Italy	Italian	Italian	0	0
Jamaica	British	Of African descent	1487	.05131
Japan	Japanese	Japanese	0	0
Jordan	Ottoman Turkish	Jordanian	710	.00771
Kazakhstan	Russian	Kazakh	949	.01312
Kenya	British	African	2288	.04456
Kiribati	British	Micronesian	1174	.05395
Korea, Dem. Rep.	Korean	Korean	0	0
Korea, Rep.	Korean	Korean	0	0
Kuwait	Arab	Arab	0	0
Alternative coding	British	Arab	236	
Kyrgyz Republic	Russian	Kyrgyz	949	.01312
Lao PDR	French	Lao	1143	.03496
Latvia	Baltic German	Latvian	828	.00946
Alternative coding	Russian	Latvian	1066	
Lebanon	Ottoman Turkish	Syro-Lebanese	710	.0087
Lesotho	Southern Sotho	Southern Sotho	0	0
Alternative coding	British	Southern Sotho	2288	

Liberia	Americo-Liberian (of African and European descent)	African	780.675	.036144
Libya	Ottoman Turkish	North African	1167	.01602
Liechtenstein	German	German	0	0
Lithuania	Russian	Lithuanian	1066	0
Luxembourg	Luxembourgish	Luxembourgish	0	0
Macao, China	Portuguese	Chinese	1236	
Macedonia, FYR	Ottoman Turkish	Macedonian	794	.00689
Madagascar	French	Malagasy	1098	.0344
Malawi	British	African	2288	.05145
Malaysia	Malay (Sultanate of Selangor around the capital Kuala Lumpur)	Malay	0	0
Alternative coding	British	Malay	1275	
Maldives	Dhivehi	Dhivehi	0	0
Alternative coding	British	Dhivehi	556	
Mali	French	African	1487	.05131
Malta	British	Maltese	51	.00607
Marshall Islands	German	Micronesian	1518	.05395
Martinique	French	Of African descent	1487	
Mauritania	Arab origin (Emirate of Trarza in and around the capital Nouakchott)	Of North African and West African descent	848.816	.0268928
Mauritius	Franco-Mauritian	Indo-Mauritian	280	.01631
Mayotte	French	Comorian	2288	.04321
Mexico	Predominantly of Spanish descent	Of Spanish and indigenous (primarily Mexica) descent	758.801	.0349402
Micronesia, Fed. Sts.	German	Micronesian	1518	.05395
Moldova	Russian	Moldovan	77	.01123
Monaco	Monegasque	Monegasque	0	0
Mongolia	Manchu	Mongolian	170	.0126
Montserrat	British	Of African descent	1487	
Morocco	Alaouite dynasty (Arab origin)	North African	263	0
Mozambique	Portuguese	African	2292	.05145
Myanmar	British	Bamar	873	.04462
Namibia	German	Ovambo, Herero, Nama, others	1708	.05145
Nauru	German	Micronesian	1518	.05395
Nepal	Nepali	Newar (in and around Kathmandu)	847	.02755
Netherlands	Dutch	Dutch	0	0
Netherlands Antilles	Dutch	Of African descent	1459	
New Caledonia	French	Melanesian	1550	
New Zealand	British and Maori	British and Maori	0	0
Nicaragua	Predominantly of Spanish descent	Of Spanish and indigenous descent	628.603	.0274453
Niger	French	African	1487	.05061
Nigeria	British (in and around the former capital Lagos)	African	1487	.04943
Niue	British	Polynesian	991	
Norfolk Island	British	Of British and Polynesian descent	495.5	

Northern Mariana Islands	German	Micronesian	1174	.05395
Norway	Norwegian	Norwegian	0	0
Oman	South Arabian	South Arabian	0	0
Alternative coding	British	South Arabian	236	
Pakistan	British	Panjabi (in and around Islamabad)	280	.01208
Palau	German	Micronesian	1174	.05395
Panama	Predominantly of Spanish descent	Of Spanish and indigenous (Chibcha) descent	755.291	.0318792
Papua New Guinea	British (in and around the capital Port Moresby)	Papuan	1816	.05395
Paraguay	Predominantly of Spanish descent	Of Spanish and indigenous (Guarani) descent	641.737	.0324896
Peru	Predominantly of Spanish descent	Of Spanish and indigenous (Quechua) descent	830.945	.0349134
Philippines	US officials, predominantly of British or other Northern European descent	Filipino	1117	.03496
Poland	Russian (in and around the capital Warsaw)	Polish	51	0
Portugal	Portuguese	Portuguese	0	0
Puerto Rico	US officials, predominantly of British or other Northern European descent	Of Spanish, indigenous (Taino) and African descent	707.766	
Qatar	Arab	Arab	0	0
Alternative coding	Ottoman Turkish	Arab	710	
Reunion	French	Of French, African, Malagasy and Asian descent	1163	
Romania	Romanian	Romanian	0	0
Russian Federation	Russian	Russian	0	0
Rwanda	German	Hutu and Tutsi	1708	.04487
Samoa	German	Samoan	1210	.0627
San Marino	Italian	Italian	0	0
Sao Tome and Principe	Portuguese	African	1794	
Saudi Arabia	Arab	Arab	0	0
Alternative coding	Ottoman Turkish	Arab	710	
Senegal	French	African	1487	.04884
Serbia and Montenegro	Serbian	Serbian	0	0
Seychelles	Franco-Seychellois	Of French, African, Malagasy and Asian descent	1073.24	.0344
Sierra Leone	British	African	1487	.04884
Singapore	British	Chinese	1152	.03964
Slovak Republic	Hungarian	Slovak	828	.00946
Slovenia	German	Slovenian	191	.00607
Solomon Islands	British	Melanesian	1550	.06976
Somalia	Italian (in and around the capital Mogadishu)	Somali	1234	.03351
South Africa	British	Primarily of Dutch descent (in and around the	21	0

		administrative capital Pretoria)		
Alternative coding	British	Northern Sotho (in the wider Tshwane region around Pretoria)	2288	
Spain	Spanish	Spanish	0	0
Sri Lanka	British	Sinhalese and Tamil	280	.0165
St. Helena	British	Of British, African, Malagasy and Asian descent	1163	
St. Kitts and Nevis	British	Of African descent	1487	.05131
St. Lucia	British	Of African descent	1487	.05131
St. Pierre and Miquelon	French	French	0	0
St. Vincent and the Grenadines	British	Of African descent	1487	.05131
Sudan	British	Of Arab and Nubian descent	885.03	.0276366
Suriname	Dutch	Of African descent	1487	.04943
Swaziland	Swazi	Swazi	0	0
Alternative coding	British	Swazi		
Sweden	Swedish	Swedish	0	0
Switzerland	Swiss-German and Swiss-French	Swiss-German and Swiss-French	0	0
Syrian Arab Republic	Ottoman Turkish	Syro-Lebanese	710	.0087
Taiwan	Japanese	Chinese and Taiwanese Aboriginal	541	.00745
Tajikistan	Uzbek (Emirate of Bukhara in and around the capital Dushanbe)	Tajik	821	.01814
Tanzania	German	African	2288	
Thailand	Thai and Thai-Chinese	Thai	52.5	.005915
Timor-Leste	Portuguese	Timorese	1215	.0627
Togo	German	African	1459	
Tokelau	British	Polynesian	991	
Tonga	Tongan	Tongan	0	0
Trinidad and Tobago	British and French	Of African descent	1487	.05131
Tunisia	French	North African	273	.01411
Turkey	Ottoman Turkish	Ottoman Turkish	0	0
Turkmenistan	Russian	Turkmen	949	.01312
Turks and Caicos Islands	British	Of African descent	1487	.05131
Tuvalu	British	Polynesian	991	
Uganda	British	African	2288	.04321
Ukraine	Russian	Ukrainian	0	0
United Arab Emirates	Arab	Arab	0	0
Alternative coding	British	Arab	236	
United Kingdom	British	British	0	0
United States	Predominantly of British or other Northern European descent	Predominantly of British or other Northern European descent	0	0

Uruguay	Predominantly of Spanish descent	Of Spanish, Italian, other European and indigenous descent	98.5461	.0043452
Uzbekistan	Russian (in and around the capital Tashkent)	Uzbek	949	.01312
Vanuatu	French (in and around the capital Port Vila)	Melanesian	1550	.0627
Venezuela, RB	Predominantly of Spanish descent	Of Spanish, indigenous (Arawak) and African descent	672.684	.0236698
Vietnam	French	Vietnamese	1100	.03496
Virgin Islands (U.S.)	Danish	Of African descent	1459	
Wallis and Futuna	French	Polynesian	991	
West Bank and Gaza	Ottoman Turkish	Palestinian	710	.00771
Yemen, Rep.	Ottoman Turkish	South Arabian	710	
Zambia	British	African	2288	.05145
Zimbabwe	British	African	2288	.05145