Infection, Genetics and Evolution 54 (2017) 299–307

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Research paper

Genetic adaptation to historical pathogen burdens☆

Johannes W. Fedderke[a,*], Robert E. Klitgaard[b], Valerio Napolioni[c]

[a] Pennsylvania State University, School of International Affairs, State College, PA, USA
[b] Claremont Graduate University, School of Social Sciences, Policy and Evaluation, Claremont, CA, USA
[c] Stanford University, School of Medicine, Palo Alto, CA, USA

ARTICLE INFO

Article history:
Received 29 March 2017
Received in revised form 7 July 2017
Accepted 13 July 2017
Available online 17 July 2017

Keywords:
Geography
Paleohistorical markers
Historical pathogen burdens
Genetic adaptation

ABSTRACT

Historical pathogen burdens are examined as possible triggers for genetic adaptation. Evidence of adaptation emerges for the acid phosphatase locus 1 (*ACP1*), interleukin-6 (*IL6*), interleukin-10 (*IL10*), human leukocyte antigen (*HLA*) polymorphisms, along with a measure of heterozygosity over 783 alleles. Results are robust to controlling for the physical and historical environment humans faced, and to endogeneity of the historical pathogen burden measure. The present study represents a proof-of-concept which may pave the way to the analysis of future aggregate measures coming from whole-genome sequencing/genotyping data.

1. Introduction

Over evolutionary time, exposure to diseases leads humans to adapt genetically. The non-random distribution of pathogens humans faced historically, resulted in a corresponding non-random distribution of human immunity to disease through genetic adaptation (Barnes et al., 2010; Diamond and Bellwood, 2003; Wolfe et al., 2007), since resistance is likely to have evolved through increased allelic variation of the major histocompatibility complex (MHC) in populations in response to the pathogens concerned (Gluckman et al., 2009; Karlsson et al., 2014; Vogel and Chakravartti, 1966; Vogel and Motulsky, 1997) . Examples include the evolved immunity to a range of temperate zone diseases amongst Europeans that decimated local populations during colonization (Diamond, 1997; Diamond and Bellwood, 2003; Dobyns, 1966; Wolfe et al., 2007). Selection at genes such as *G6PD*, *HBB* and *CD40LG*, variation in which confer protection against malaria, seems to have started within the past 10,000 years (Siddle and Quintana-Murci, 2014) , coinciding with the Neolithic period. Since genetic adaptation takes time, the emergence of pathogen resistance is likely to be dynamic, with morbidity and mortality rising before falling with the emergence of immunity (Cohen, 1989).

In this paper, we explore genetic adaptation responses to diseases that humans faced historically. It goes without saying that we are not suggesting that genetic adaptation is the only evolutionary response to disease pressure. A range of alternatives cover social organization, culture, institutions and (medical) technology, amongst others. See for instance Thornhill and Fincher (2014) and Hays (2009). These are not the focus of the present discussion. The novel feature of the paper is that it employs a large compilation of global phenotypes, tests for the historical pathogen intensity to genetic adaptation link at the country level of aggregation, subjecting the evidence to a range of statistical robustness tests.

Three testing strategies are employed to test the robustness of reported results.

(A) First, we report the strength of statistical association between historical pathogen burdens and a set of genetic markers for which the link is hypothesized on a priori grounds as ranging from direct, to indirect, to not relevant as a disease response at all.
(B) Second, we test for the robustness of the association between historical pathogen burdens and genetic markers, while controlling for a wide array of variables that measure global and local geographical conditions, as well as conditions that served as possible triggers for changes in human behavior

☆ Fedderke acknowledges the research support of Economic Research Southern Africa.
* Corresponding author.
E-mail address: jwf15@psu.edu (J.W. Fedderke).

http://dx.doi.org/10.1016/j.meegid.2017.07.017
1567-1348/

such as the transition to agriculture, that may have independently influenced genetic adaptation.

(C) Third, statistical adjustments for the possibility of possibility that adaptation may itself alter recorded intensity of disease, for instance due to improved immunity (i.e. the presence of reverse causation rendering estimated associations biased and inconsistent), also reinforced the statistical association between historical disease measures and genetic adaptation.

In undertaking these tasks, we have assembled a novel data set, that combines not only a range of genetic data based on a large compilation of global phenotypes, with a set of variables covering geographical conditions across all country level units of aggregation (up to 238 geographical sites across the globe), as well as variables that cover human development from the present to paleohistorical time periods.

2. Methods

2.1. Data — historical pathogen burden

Our measure of the historical pathogen burden faced by humans is derived from Murray and Schaller (2010). We employ the seven-disease index (*ms_7*) which covers leishmanias, schistosomes, trypanosomes, malaria, typhus, filariae, and dengue. The Murray and Schaller (2010) source also records a 9 disease (adds leprosy and tuberculosis) and 6 disease (drops malaria) index value. We choose the 7 disease index since it has much greater geographic coverage than the 9 disease index (224 geographic locations vs. 160), and greater pathogen coverage than the 6 disease index. The alternative measures contain little by way of different information from the 7 disease index, with correlations between the 7 and 9, and 7 and 6 disease indexes of 0.98 and 0.98.

2.2. Data — genetic markers

We consider a number of distinct genes that have been linked to disease adaptation. For some, the link between pathogens and genetic adaptation is direct (*ACP1, IL6, IL10, HLA*), for some indirect (*FAAH*, *SLC6A4*) or merely established by observed correlation (*Rhesus*), and for some there should be no association by construction (genetic distance). A more detailed overview on the role of the different genetic measures considered is provided in Supplementary material.

We have assembled data on the frequencies of *ACP1*A*, *ACP1*B*, and *ACP1*C* alleles (gene map locus on Chr. 2p25.3, OMIM*171500) in the populations of 121 countries. The data is a compilation of 153,090 global genotypes. This is a new data set, and therefore this paper represents the first time country-level *ACP1* frequencies have been incorporated into studies of historical pathogen burdens. In addition to *ACP1*, we have assembled new data on national allele frequencies of the Interleukin-6 (*IL6*, gene map locus on Chr. 7p15.3, OMIM*147620) (*IL-6*) -174G > C (rs1800795) and the Interleukin-10 (*IL10*, gene map locus on Chr. 1q32.1, OMIM*124092) (*IL10*) -1082G > A (rs1800796) polymorphisms. *ACP1*, *IL6 rs1800795* and *IL10 rs1800796* genotypes were retrieved through an extensive literature search carried out on PubMed and Google free search engines. A detailed explanation of the data retrieval and the definition of country-level estimates for *ACP1*, *IL6 rs1800795* and *IL10 rs1800796* allele frequencies are provided in Supplementary material.

From Ashraf and Galor (2013) we employ a measure of genetic heterozygosity in countries adjusted for ancestry (*pdiv_aa*). It measures the expected heterozygosity between two randomly selected people in the country in question, after adjusting for ancestry. It is based on two sources. The first comes from data about heterozygosity in 53 ethnic groups in the HGDP-CEPH Human Genome Diversity Cell Line Panel in a sample of 21 countries. Second, Ashraf and Galor (2013) build on work by Ramachandran et al. (2005), which shows that this heterozygosity is highly correlated with the migratory distance of these 53 groups from East Africa ($r = 0.92$). This robust association between genetic diversity and migratory distance before the Common Era is used to obtain predicted values of genetic diversity for an extended sample of 145 countries.

A further measure of genetic diversity is provided by Cook (2015), in the form of the human leukocyte antigen (*HLA*) system, a highly polymorphic genetic cluster located on the sixth chromosome, responsible for the location of foreign proteins in order to direct an immune response to identified pathogens.

The prevalence of the *rs324420 A allele* in the *FAAH* gene (gene map locus on Chr. 1p33, OMIM*602935) (Minkov and Bond, 2016), and the *5-HTTLPR Short allele* (*SLC6A4*S*) in the serotonin-transporter gene (*SLC6A4*, gene map locus on Chr. 17q11.2, OMIM*182138) are obtained from Chiao and Blizinsky (2010) and Minkov et al. (2014). The *rs324420 A*allele is involved in the hydrolysis of anandamide, a substance that enhances sensory pleasure and helps reduce pain (Minkov and Bond, 2016), and may thus represent an adaptation to the impact of disease. The *5-HTTLPR S allele* shows significant geographic variation, with higher East Asian than European frequencies, and since carries of the *S allele* produce significantly less *5-HTT mRNA* and protein, generating higher concentrations of serotonin in the synaptic cleft relative to carriers of the *Long allele*, which in turn is associated with increased negative emotion, heightened anxiety, harm avoidance, fear conditioning, attention bias toward negative information, and depression in the face of environmental risk factors (Chiao and Blizinsky, 2010). The implication is that the frequency of the *SLC6A4*S* may map into social and personality traits, including individualism vs. collectivism, IQ, risk acceptance, and long- or short-term orientation (Chiao and Blizinsky, 2010; Minkov et al., 2014).

The Rhesus factor (*RHD*, gene map locus on 1p36.11, OMIM*111680) polymorphism measures derive from Flegr (2016), for both the frequency of Rhesus negative homozygotes (*rhdneg*) and Rhesus positive heterozygotes (*rhdhetero*). Flegr (2016) reports that the burden associated with many diseases correlated with the frequencies of particular Rhesus genotypes in a country and that the direction of the relation was nearly always the opposite for the frequency of Rhesus negative homozygotes (*rhdneg*) and that of Rhesus positive heterozygotes (*rhdhetero*).

Indexes on genetic distance between human populations come from Spolaore and Wacziarg (2016), for the plurality population group in a country (*gdist_plu*), and a weighted genetic distance in which each population group is represented by population weight (*gdist_w*). Spolaore and Wacziarg (2016) follow Cavalli-Sforza et al. (1994) in using measures of F_{ST} distance, based on indices of heterozygosity, the probability that two alleles at a given locus selected at random from two populations will be different. F_{ST} takes a value equal to zero if and only if the allele distributions are identical across the two populations, whereas it is positive when the allele distributions differ. A higher F_{ST} is associated with larger differences. The computation of genetic distances concentrates on neutral characteristics that are not affected by strong directional selection, but only by random drift.

2.3. Data — geography

Data on mean elevation, its variability, as well as the climatic zones of countries on the Köppen-Geiger classification system is derived from Center for International Earth Science Information Network (CIESIN), Columbia University (2007). For climatic conditions, we include a set of 5 climatic controls, which represent aggregations of the 45 climate type classification under Köppen-Geiger: the percentage of territory that is tropical (*P_Tropical*), mondane

(*P_Mondane*), temperate (*P_Temperate*), continental (*P_Continental*) or dry (*P_Dry*). We also controlled for the diversity of climatic conditions in each country, by means of a climate Gini coefficient (our computation). The coefficient is distributed over the 0–1 interval, with a value of 0 occurring where all climatic zones constitute an equal proportion of the land area, while →1 reflects ever greater lack of proportionality of the climatic zones in a countries' land area.

Data on absolute longitude and latitude, mean precipitation, its variability, terrain roughness, soil quality, mean temperature and its variability, distance from water sources (oceans, lakes, rivers), and the nature of the vegetation are derived from Nordhaus and Chen (2016) and Nordhaus (2006). Note that this data is scaled from country one degree grid cells, weighted relative to country total area. Data on global temperatures for the past 11,300 years, obtained from paleotemperature readings from 73 globally distributed sites, with sampling distributions ranging from 20 to 500 years, and median resolution of 120 years, is derived from Marcott et al. (2013). We employ both first and second moment measures. Population size and density from 10,000 BCE to 2000 CE is obtained from Klein Goldewijk et al. (2010).

Data on the number of domesticable animal and plant species and the continental landmass size and orientation from Olsson and Hibbs (2005) and Hibbs and Olsson (2004). Coverage was extended for domesticable continental species to all countries in the data base one the basis of the Olsson and Hibbs (2005) and Hibbs and Olsson (2004) classification, while continental landmass and axis orientation was extended to island locations on the basis of Center for International Earth Science Information Network (CIESIN), Columbia University (2007) data. Finally, as a measure of timing associated with human settlement to that measuring the timing of the Neolithic transition, we also consider a measure of the duration of human settlement (Ahlerup and Olsson, 2012) .

The entire dataset including all the variables used for the analysis is reported as Supplementary data.

3. Expected results

We anticipate robust evidence of adaptation in the *ACP1*, *IL6*, *IL10,* and *HLA* genes. The measure of heterozygotic genetic diversity (*pdiv_aa*) may also show statistical association with historical pathogen burdens. Given the indirect link between the *FAAH*, and *SLC6A4* genes and pathogen burdens, and the correlative evidence in favor of an association between pathogen burdens and the *RHD* gene, in these instances we anticipate weaker, or less robust statistical association with measures of historical pathogen burdens.

By way of a counterfactual test, since the measures of genetic distance (*gdist*) by construction are independent of adaptation to pathogen burdens, these variables should not report statistical association with historical disease burdens.

Reported associations should be robust to controlling for the influence of geography, and paleohistorical features of the environment humans faced, and to allowing for the possibility that recorded pathogen intensity is itself determined by genetic adaptation (endogeneity).

4. Estimation methodology

Our empirical results below confirm the presence of strong correlations between measures of historical disease environments and the distributions of certain genes. However, correlations cannot confirm that the association is not simply a reflection of other environmental factors which may have triggered genetic adaptation that happen to be correlated with historical pathogen burdens. Nor do simple correlations allow for correction for the impact of possible of reverse causation, that genetic adaptation can itself come to impact recorded pathogen burdens. To allow for these concerns, our empirical methodology controls for an array of other variables that potentially affect genetic adaptation, and corrects for the possibility of reverse causation by means of instrumental variables estimation.

To explore the association between our genetic markers and historical pathogen burdens, our baseline ordinary least squares (OLS) estimation specification is given by:

$$G_i = \beta_0 + \beta_{MS} P_i + \varepsilon_i \tag{1}$$

where G_i denotes out set of genetic markers (*ACP1*A*, *ACP1*B*, *IL6*, *IL10, HLA, pdiv_aa, FAAH*, *SLC6A4, RHD, gdist_plu, gdist_w*), P_i our measure of historical pathogen burdens, and ε_i a Gaussian error for country *i*.

To allow for the influence of geography, and paleohistorical features of the environment humans faced, and for endogeneity of the pathogen burden measure given the possibility that $cov(P_i, \varepsilon_i) \neq 0$ thus rendering OLS estimation under specification Eq. (1) biased and inconsistent, estimation is by instrumental variables (IV):

$$G_i = \beta_0 + \beta_{MS} \hat{P}_i + \sum_j \beta_j X_{i,j} + \eta_i \tag{2}$$

$$P_i = \pi_0 + \sum_k \pi_k Z_{i,k} + \sum_j \pi_j X_{i,j} + \nu_i \tag{3}$$

where notation is defined as above, $X_{i,j}$ denotes a set of *j* exogenous geographical and paleohistorical controls, $Z_{i,k}$ a set of *k* exogenous instruments, and η_i, ν_i, two Gaussian error terms for country *i*. $\hat{P}_i$ denotes the conditional mean values obtained from the reduced form first stage regression Eq. (3). Standard errors reported for the second stage regression Eq. (2) are corrected for impact of the reduced form estimation.

Legitimate IV estimation requires instrument *strength* ($r_{P,Z} \to 1$) and *validity* ($r_{Z,\varepsilon} \to 0$). Under instrument *validity*, since $r_{Z,\varepsilon} \to 0$, the conditional mean values, $\hat{P}_i$, obtained from the reduced form Eq. (3), come to satisfy the $cov(P_i, \varepsilon_i) = 0$ condition for consistency of the least squares estimator. Provided only that instrument strength is satisfied, since $r_{P,Z} \to 1$, the conditional mean values, $\hat{P}_i$, employed in the second stage regression Eq. (2), have not lost the information contained in the historical pathogen burden measure, P_i.

Instrument *strength* is readily confirmed from the reduced form estimation, or direct consideration of the strength of association between the endogenous regressor, P_i, and the set of instruments, Z_i. Instrument *validity* requires instruments to be orthogonal to the second stage population error structure, satisfying the exclusion restriction that the instruments do not have a direct impact on the second stage regression dependent variable. We present evidence in support of both requirements of legitimate IV estimation.

5. Results: genetic adaptation to historical pathogen burdens

The patterns of association specified in Section 3 are plausible on first examination of the data. For the *ACP1*, *IL6* and *IL10* genes, strong bivariate associations with historical pathogen burdens are borne out — see Fig. 1 and the associated Pearson product-moment correlations recorded in the Figure caption. For the genetic heterogeneity measures (*HLA*, *pdiv_aa*), the association with historical pathogen burdens is both considerably weaker, and subject to the impact of strong outliers — see Fig. 2 and associated Pearson correlations. For the *FAAH* , *SLC6A4*S*, and Rhesus factor polymorphism genes associations are very weak — Figs. 2 and 3 and associated Pearson correlations. For the genetic distance variables, association with historical pathogen burdens is entirely absent — Fig. 3 and associated Pearson correlations.

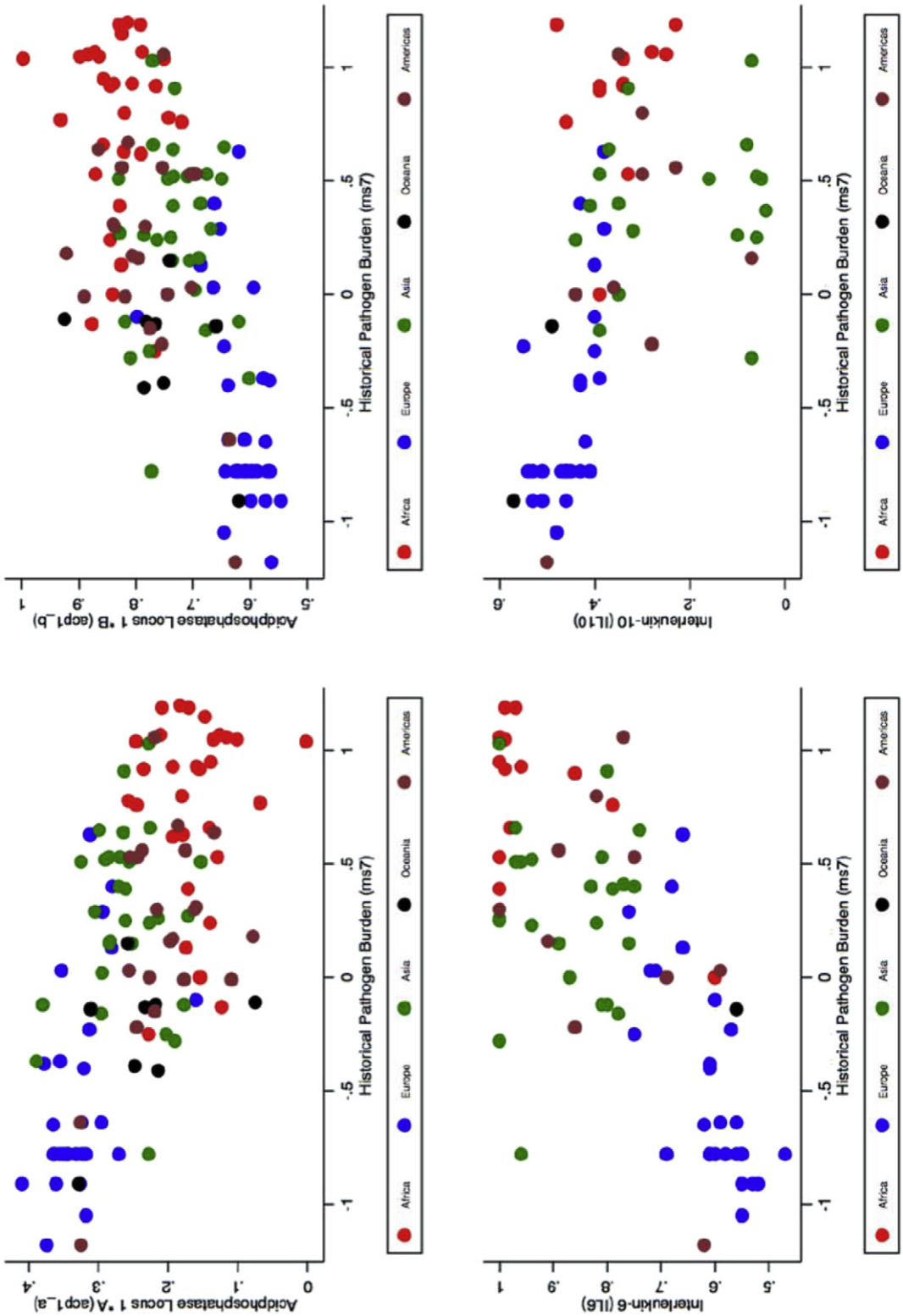

Fig. 1. Historic pathogen burdens with the *ACP1*, *IL6*, *IL10* genes. Correlation of 7-disease index with *ACP1*A* ($r = -0.61$), *ACP1*B* ($r = 0.64$), *IL6* ($r = 0.73$), and *IL10* (-0.50).

Note that this evidence conforms to the anticipated strength of association noted in Section 3.

The same inference follows from the OLS regression results reported in Table 1, reporting results from the estimation of Eq. (1) of Section 4.

Results confirm that *ACP1*A* and *IL10 -1082*G* are statistically significantly negatively associated with historical pathogen burdens, while for *ACP1*B* and *IL6-174*G* the association is statistically significant and positive – columns 1–4 of Table 1. What is more, historical pathogen burdens account for 40–50 % of the variation in

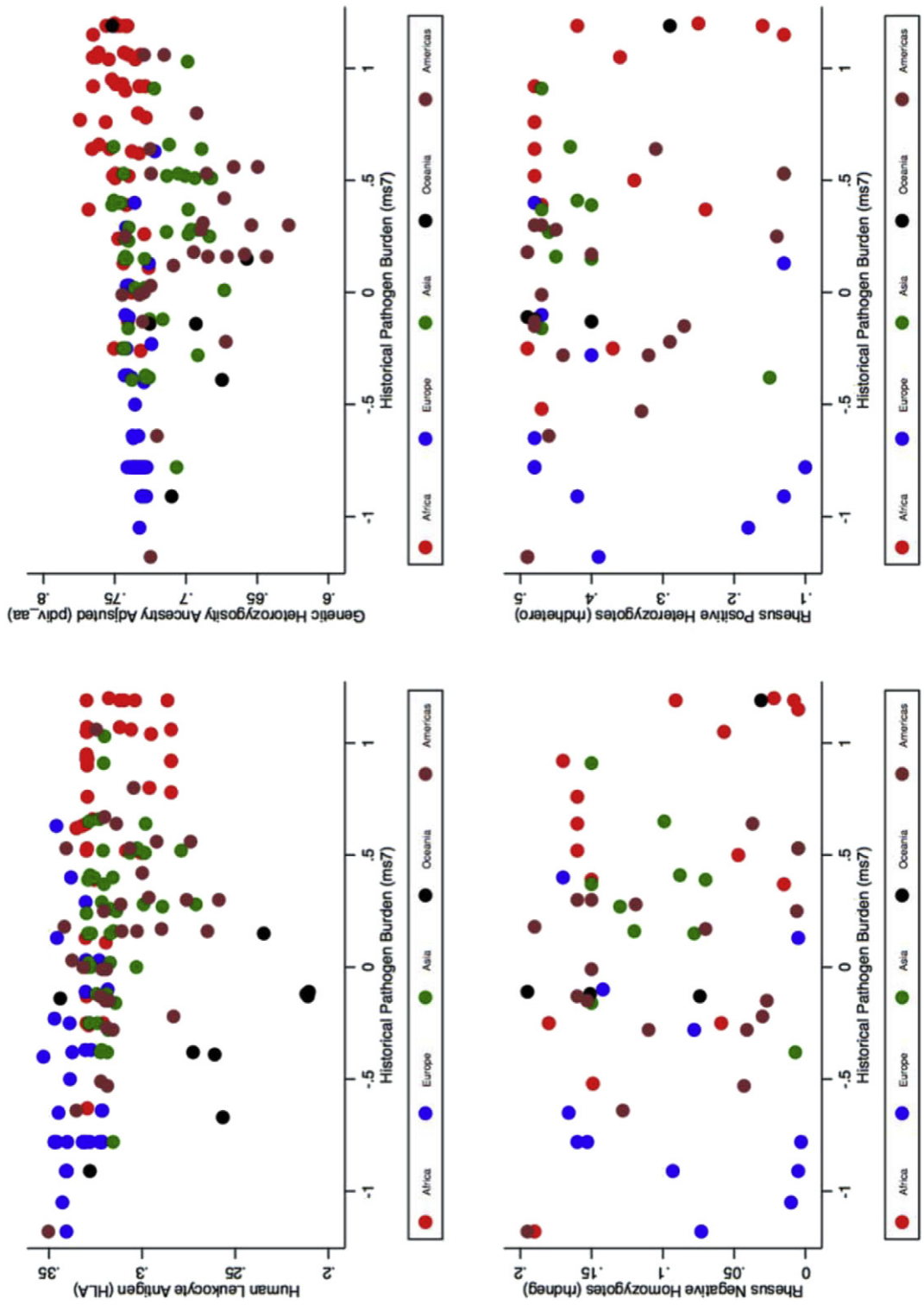

Fig. 2. Historic pathogen burdens with the genetic diversity measures, and the RHDNEG, RHDNHETERO genes. Correlation of 7-disease index with *HLA* ($r = -0.16$), pdiv_aa ($r = 0.10$), RHDNEG ($r = -0.16$), and RHDHETERO (-0.11).

the frequency of the country-level genetic markers (variation at the country level of aggregation will return higher levels of goodness-of-fit than at the individual level of aggregation). By contrast, the *FAAH*, *SLC6A4 5-HTTLPR*S*, *rhdneg*, *rhdhetero* genes (columns 9 & 10, 7 & 8), and the genetic distance measures (*gdist_w, gdist_plu,*columns 11 & 12) all prove to be statistically insignificantly associated with the historical pathogen burden measures, and they fail to account for any variation in the genetic variables ($R^2 \approx 0$). The *HLA* heterogeneity measure (column 5) does prove to be statistically significantly and negatively associated with the historical pathogen burden, but much more weakly than for the *ACP1*, *IL6* and *IL10* genes, with only 2% of the genetic variation accounted for by the historical pathogen burden — surprisingly since the *HLA* polymorphism is explicitly presented as a disease adaptation mechanism. Moreover, we find the

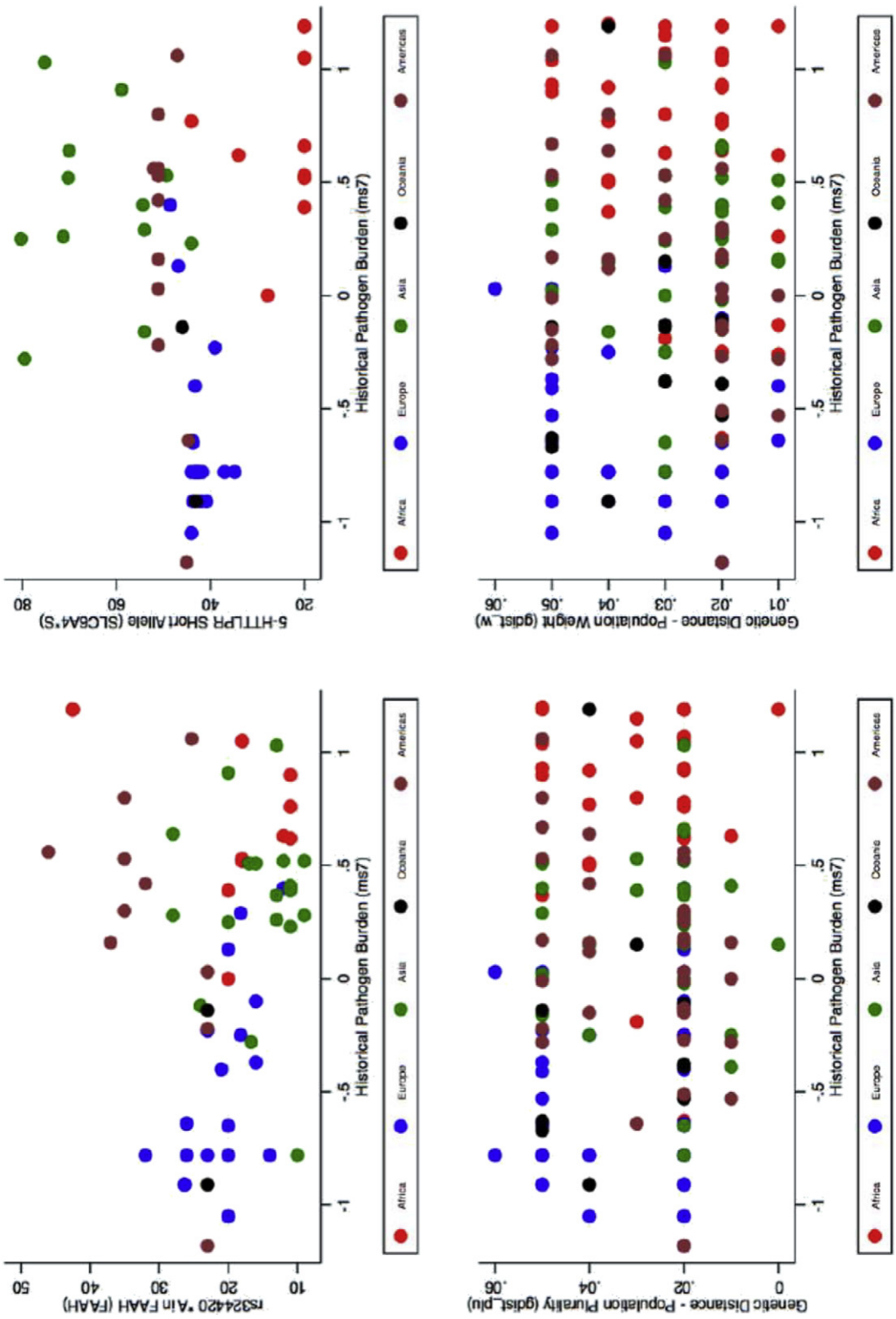

Fig. 3. Historic pathogen burdens with the *FAAH*, *SLC6A4*S*, genes, and the genetic distance measures. Correlation of 7-disease index with *FAAH* ($r = 0.05$), *SLC6A4*S* ($r = -0.08$), GDIST_W ($r = -0.05$), and GDIST_PLU ($r = 0.05$).

HLA polymorphism *declining* rather than increasing in the intensity of historical pathogen burdens, contradicting the prior expectation that the polymorphism should increase in response to pathogen exposure as an immunity-response. The *pdiv_aa* measure of genetic heterogeneity (column 6) is insignificantly though positively associated with the 7-disease index of historical pathogen burdens, with only 3% of the variation in heterogeneity is accounted for by the disease burden measure.

The evidence is thus consistent with the expectations specified in Section 3. The projection of a strong association with historical

Table 1
Genetic adaptation to historical pathogen burden.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	ACP1*A	ACP1*B	IL6	IL10	HLA	pdiv_aa	rhdneg	rhdhetero	FAAH	SLC6A4*S	gdist_w	gdist_plu
ms_7	−0.0789***	0.102 ***	0.190***	−0.108***	−0.00656*	0.00461	−0.0155	−0.0209	0.716	−1.642	−0.000989	0.0128
	(−8.40)	(9.09)	(9.26)	(−4.69)	(−2.02)	(1.34)	(−1.21)	(−0.82)	(0.42)	(−0.60)	(−0.60)	(0.66)
Constant	0.253***	0.722***	0.760***	0.360***	0.316***	0.726***	0.0986***	0.379***	21.28***	44.55***	0.0304***	0.0411**
	(41.75)	(100.18)	(57.55)	(23.24)	(159.13)	(330.82)	(12.05)	(23.38)	(18.75)	(23.23)	(29.15)	(3.33)
Observations	120	120	77	67	167	164	61	61	65	59	162	162
Adjusted R^2	0.369	0.407	0.527	0.241	0.018	0.005	0.008	−0.005	−0.013	−0.011	−0.004	−0.004

t statistics are in parentheses.
* $p < 0.05$.
** $p < 0.01$.
*** $p < 0.001$.

pathogen burdens is confirmed for the *ACP1*, *IL6*, and *IL10* genes, with historical pathogen burdens proving statistically significant for each gene, and accounting for roughly half the variation in the four genetic variables. Associations of historical pathogen burdens with the *FAAH*, *SLC6A4*S*, *rhdneg*, *rhdhetero* genes are weak, with historical pathogen burdens proving statistically significant, and accounting for none of the variation in the genetic measure. Evidence also confirms historical pathogen burdens' statistically insignificance for all the genetic distance measures, nor does it account for any variation in the genetic distance variables.

The only countervailing evidence emerges for the genetic heterogeneity variables. For the *HLA* measure, historical pathogen burdens are statistically significant, but predict a decline in diversity with rising pathogen burdens, rather than an increase as hypothesized. What is more, only 2% of the variation in *HLA* is accounted for by historical disease burdens. For the heterozygotic diversity measures, historical pathogen burdens are statistically insignificant for the ancestry adjusted (*pdiv_aa*) measure, and again the proportion of diversity accounted for remains small (3%).

Nonetheless results are reassuring: strong statistical association with historical pathogen burdens emerges for genes where a mechanism of adaptation has been suggested; where such mechanism are not identified, the association is weak or altogether absent.

5.1. Robustness controlling for geographical and paleohistorical determinants of genetic adaptation and allowing for endogeneity of historical pathogen burdens

Are our results robust to controlling for additional geographical and paleohistorical environmental factors, and when we allow for the possibility that genetic adaptation may itself come to influence recorded intensity of pathogen burdens, generating reverse causality from the genetic measures to the historical pathogen burden measure? Both concerns carry the same statistical consequence: bias and inconsistency of parameter estimates, such that any inference is tainted.

We address these statistical concerns by the estimation of the system (Eqs. (2), (3)) specified in Section 4.

The first concern is addressed by controlling for an array of geographical and historical variables at the country level of aggregation (the $\sum_j \beta_j X_{i,j}$ component of specification Eq. (2)). In particular, we control for absolute latitude (*ABSLAT*), ultraviolet radiation intensity (*uvr*), the number of domesticable species that humans encountered in different locations (*Diamond*), mean elevation (*MEANELEV*), roughness of terrain (*ROUGH*), mean precipitation (*AVPREC*), mean (*AVTEMP*) and standard deviation (*SDTEMP*) of temperature, distance to navigable water (*DWater*), mean (*mt8_10kBC*) and standard deviation (*sd8_10kBC*) of temperature in the early Holocene, and population size (*p10kBC*) and density (*pd10kBC*) in 10,000 BCE, and then a set of measures of the proportion of the land area of countries that fall into the five principal Köppen-Geiger classifications, tropical, montane, temperate, continental and dry (*P_Tropical, P_Montane, P_Temperate, P_Continental, P_Dry*).

To address the second concern, we estimate by means of instrumental variables (IV). Our instruments include the length of human occupation (*origtime*), absolute longitude (*ABSLONG*) and latitude (ABSLAT), the axis rotation (*axis*) and size (*size*) of the continental landmass of a country, and a measure of the variability of climate in a country on the Köppen-Geiger classification (*Climate_Gini*). For legitimate IV estimation, we require instrument *strength* and *validity*. Instrument *strength* is confirmed by the regression reported in Table 2, which confirms not only the statistical significance of the instruments with respect to the historical pathogen burden measure, but that they account for approximately 73% of the variation in historical pathogen burdens. Instrument *validity* requires instruments to be orthogonal to the second stage population error structure, and to satisfy the exclusion restriction that the instruments do not have a direct impact on the second stage regression dependent variable. While our instruments might impact genetic adaptation, as argued by Diamond (1997) this would have been *through* exposure to pathogen burdens. Moreover, in each instance we allow for additional geographical and environmental variables that are highly correlated with the instruments in the second stage regression, to capture any direct effect on genetic adaptation. Thus, in the case of the axis rotation and landmass size of continents, we allow the

Table 2
Regression of historical pathogen burden measure on instrumental variables.

	(1)
	ms7
origtime	0.00000961**
	(3.12)
origtime_sqr	−5.60e-11**
	(−2.73)
ABSLONG	−0.00211*
	(−2.30)
ABSLAT	−0.0258***
	(−11.31)
axis	−0.278***
	(−3.59)
size	0.00765*
	(2.30)
Climate_Gini	−1.159*
	(−2.37)
Constant	1.336***
	(5.71)
Observations	168
Adjusted R^2	0.728

t statistics are in parentheses.
* $p < 0.05$.
** $p < 0.01$.
*** $p < 0.001$.

for the number of domesticable species to impact genetic adaptation directly. In the case of the climate measures, we control for the proportion of territory under the alterative Köppen-Geiger classifications. Finally, note that Angrist et al. (1996) demonstrate that the exclusion restriction comes to be satisfied where the goodness of fit between instruments and endogenous regressor is high. Since the association between the historical pathogen burden measure and the instruments reported in Table 2 confirms an adj-R^2 of 0.73, this requirement is met.

Second stage estimation results are reported in Table 3. We restrict the analysis to those genetic markers for which there is evidence of a response to historical pathogen burdens, and for which we have sufficient observations to render the expansion of independent variables and the use of the instrumental variables technique credible — unfortunately this eliminates the *IL6* and *IL10* adaptations. We make one exception by also including the ancestry adjusted measure of genetic diversity (pdiv_aa) despite the insignificant result of Table 1. We do so since the ancestry unadjusted form of the variable is statistically significantly associated with historical pathogen burdens, suggesting the possibility of a link (not reported in Table 1, given the focus on the ancestry adjusted format of the *pdiv* variable).

Table 3
Genetic adaptation to historical pathogen burden.

	(1)	(2)	(3)	(4)
	ACP1*A	ACP1*B	HLA	pdiv_aa
	(IV)	(IV)	(IV)	(IV)
ms_7	-0.0900***	0.0998***	0.0168*	0.0466***
	(−3.43)	(3.45)	(2.12)	(5.36)
uvr	0.000259	0.00000857	−0.000268**	−0.000387***
	(0.82)	(0.02)	(−2.81)	(−4.19)
Diamond	0.0281***	−0.0334***	−0.000886	−0.00306
	(4.06)	(−4.39)	(−0.38)	(−1.26)
MEANELEV	2.94e-10	−3.83e-10	1.26e-11	−2.10e-10
	(0.79)	(−0.94)	(0.09)	(−1.49)
ROUGH	−0.162**	0.174**	0.00947	0.0411*
	(−2.86)	(2.79)	(0.81)	(2.28)
AVPREC	0.00000969	−0.00000555	−0.0000151***	−0.0000251***
	(0.69)	(−0.36)	(−3.66)	(−5.13)
AVTEMP	0.00588*	−0.00795*	−0.000555	−0.00125*
	(1.97)	(−2.42)	(−1.00)	(−2.01)
SDTEMP	−0.00481	0.00933*	−0.000245	0.000154
	(−1.14)	(2.00)	(−0.18)	(0.11)
DWater	0.0000221	−0.0000294	−0.00000396	−0.00000467
	(1.08)	(−1.30)	(−0.63)	(−0.74)
mt8_10kBC	0.142**	−0.154**	−0.0123	−0.0454*
	(3.20)	(−3.15)	(−0.72)	(−2.51)
sd8_10kBC	−1.712*	1.481	−0.0806	0.0803
	(−2.34)	(1.84)	(−0.28)	(0.26)
sd8_10kBC_sqr	4.394*	−3.780	0.183	−0.176
	(2.29)	(−1.79)	(0.24)	(−0.22)
pd10kBC	0.0300	−0.0416	−0.00457	−0.0132
	(1.16)	(−1.45)	(−0.46)	(−1.28)
p10kBC	0.000153**	−0.000176**	−0.0000335	−0.000107***
	(2.71)	(−2.84)	(−1.47)	(−4.50)
P_Tropical	−0.000745	0.000700	−0.0000859	−0.00123
	(−0.64)	(0.55)	(−0.81)	(−0.66)
P_Montane	0.00102	−0.00140	−0.000234	−0.00173
	(0.82)	(−1.01)	(−1.12)	(−0.93)
P_Temperate	−0.000455	0.000230	−0.0000901	−0.00135
	(−0.39)	(0.18)	(−0.79)	(−0.73)
P_Continental	−0.000167	−0.000397	−0.000321*	−0.00135
	(−0.14)	(−0.29)	(−2.18)	(−0.73)
P_Dry	−0.00100	0.000927	−0.000176	−0.00123
	(−0.85)	(0.71)	(−1.62)	(−0.67)
Constant	0.281	0.690***	0.418***	0.968***
	(1.76)	(3.93)	(12.91)	(5.05)
Observations	116	116	158	157
Adjusted R^2	0.580	0.665	0.259	0.325

t statistics are in parentheses.
* $p < 0.05$.
** $p < 0.01$.
*** $p < 0.001$.

In the presence of the array of geographical and historical variables, and under the IV-estimation, we find that the measure of historical pathogen burdens maintains its statistical significance for the four genetic measures. The *ACP1*B*, and *pdiv_aa* measures maintain their positive association with the instrumented measure of the historical pathogen burden, and *ACP1*A* its negative association. The ancestry adjusted heterogeneity measure now gains statistical significance. For the *HLA* polymorphism, under the instrumentation strategy we now find the theoretically mandated *positive* and statistically significant response to rising historical pathogen burdens.

Note that the statistically significant genetic response to historical pathogen burdens is maintained even in the presence of the wide range of geographical controls. Of the geographical controls, particularly the measure of the number of domesticable species (Diamond), terrain roughness, mean temperature, and temperature variation in the early Holocene, as well as population size in 10,000BC show signs of an independent impact on the genetic markers.

The evidence is thus consistent with the existence of genetic adaptation to historical disease burdens, robust to controlling for the impact of influence of geography, and paleohistorical features of the environment humans faced, consistent with the requirement that the association between pathogens and genetic adaptation is not statistically spurious. What is more, the association between historical pathogen burdens and genetic adaptation proves robust to allowing for the endogeneity of the pathogen burden measure.

6. Conclusions and evaluation

We explored evidence in support of human adaptation mechanisms in the face of historical pathogen burdens. The novel feature of the paper is that it employs a large compilation of global phenotypes to date, tests for the historical pathogen intensity to genetic adaptation link at the country level of aggregation, subjecting the evidence to a range of statistical robustness tests. Specifically, it examines the pathogen burden to genetic adaptation association across cases in which the strength of association should range from strong, to weak at best, to absent, to allow for the counterfactual case.

In addition, our testing strategy allows for both omitted variables bias and endogeneity of historical pathogen burdens.

We report evidence consistent with genetic adaptation to historical disease burdens. The evidence for genetic adaptation is not only statistically significant, but robust to controlling for a wide range of additional measures of the physical and historical environment humans have faced, and for possible reverse causality from genetic adaptation to historical pathogen burdens.

Importantly, strong statistical association with historical pathogen burdens emerges for genes where a mechanism of adaptation has been suggested. Where such mechanisms are not identified, the association is weak. For genetic markers constructed to ensure independence of the genetic measure from disease adaptation, we confirm the absence of any association. Thus evidence of adaptation emerges for the acid phosphatase locus 1 (*ACP1*) soluble genetic polymorphism, the interleukin-6 (*IL6*) G-allele and interleukin-10 (*IL10*) G-allele, the human leukocyte antigen (HLA), and the *pdiv_aa* measure of genetic diversity. No evidence of a statistically significant response to historical pathogen burdens emerges for measures of genetic distance between human populations, in the *rs324420 A* allele in the *FAAH* gene, nor in the *5-HTTLPR Short* allele (*SLC6A4*S*) in the serotonin-transporter gene.

Limitations attach to country level analyses arising a range of data quality issues beyond the usual loss of information due to aggregation. Theses arise from incomplete geographical coverage

across the full set of countries, introducing the potential of selection effect biases, as well as questions surrounding their construction in some instances (such as the inferential source of the *pdiv_aa* data for many of the sample points). To minimize such problem, we decided to focus on a few, widely-studied genetic variants with established functional effects, to guarantee a sufficient coverage at global level by determining their respective country-level estimates, instead of using a genome-wide approach. The present study represents a proof-of-concept which may pave the way to the analysis of future aggregate measures coming from whole-genome sequencing/genotyping data. Big efforts are being made in genotyping several populations across the world [e.g. POPRES Nelson et al., 2008, ALFRED Rajeevan et al. (2012), Haplotype Refence Consortium McCarthy et al., 2016, Simons Genome Diversity Project Mallick et al., 2016]. We are confident that soon we will be able to analyze the country-level aggregates of allele frequencies for the whole genome. However, this effort will require a very careful harmonization of genetic data, through imputation and population structure analyses, allowing a proper handling of genetic data at country-level.

On the other hand, our evidence suggests that useful insight under the application of appropriate statistical techniques is nonetheless feasible.

In the specified set of genetic dimensions there does therefore appear to be support for adaptation to historical pathogen burdens, which is robust to controlling for the physical and historical environment humans faced, and to endogeneity of the historical pathogen burden measure.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.meegid.2017.07.017.

References

Ahlerup, P., Olsson, O., 2012. The roots of ethnic diversity. J. Econ. Growth 17, 71–102.

Angrist, J.D., Imbens, G.W., Rubin, D.B., 1996. Identification of Causal Effects using Instrumental Variables. J. Am. Stat. Assoc. 91 (434), 444–455.

Ashraf, Q., Galor, O., 2013. The "Out of Africa" hypothesis, human genetic diversity, and comparative economic development. Am. Econ. Rev. 103 (1), 1–46.

Barnes, I., Duda, A., Pybus, O., Thomas, M., 2010. Ancient urbanisation predicts resistance to tuberculosis. Evolution 65 (3), 842–848.

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

Center for International Earth Science Information Network (CIESIN), Columbia University, 2007. National Aggregates of Geospatial Data: Population, Landscape and Climate Estimates, V.2 (PLACE II). CIESIN, Columbia University, Palisades, NY. Available at: http://sedac.ciesin.columbia.edu/place/.

Chiao, J., Blizinsky, K., 2010. Culture-gene coevolution of individualism-collectivism and the serotonin transporter gene. Proc. R. Soc. B 277, 529–537. http://dx.doi.org/10.1098/rspb.2009.1650.

Cohen, M.N., 1989. Health and the Rise of Civilization. Yale University Press, New Haven.

Cook, C.J., 2015. The natural selection of infectious disease resistance and its effect on contemporary health. Rev. Econ. Stat. 97 (4), 747–757.

Diamond, J., 1997. Guns, Germs and Steel. W.W.Norton.

Diamond, J., Bellwood, P., 2003. Farmers and their languages: the first expansions. Science 300, 597–603.

Dobyns, H., 1966. An appraisal of techniques with a new hemispheric estimate. Curr. Anthropol. 7, 395–416.

Flegr, J., 2016. Heterozygote advantage probably maintains rhesus factor blood group polymorphism: ecological regression study. PLoS ONE 11 (1), e0147955. http://dx.doi.org/10.1371/journal.pone.0147955.

Gluckman, P., Beedle, A., Hanson, M., 2009. Principles of Evolutionary Medicine. University Press, Oxford.

Hays, J.N., 2009. The Burdens of Disease: Epidemics and Human Response in Western History. Rutgers University Press.

Hibbs, D.A., Olsson, O., 2004. Geography, biogeography, and why some countries are rich and others are poor. Proc. Natl. Acad. Sci. 101, 3715–3720.

Karlsson, E.K., Kwiatkowski, D.P., Sabeti, P.C., 2014. Natural selection and infectious disease in human populations. Nat. Rev. Genet. 15, 379–393. http://dx.doi.org/10.1038/nrg3734.

Klein Goldewijk, K., Beusen, A., Janssen, P., 2010. Long term dynamic modeling of global population and built-up area in a spatially explicit way, HYDE 3.1. Holocene 20 (4), 565–573. http://dx.doi.org/10.1177/0959683609356587.

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., Skoglund, P., Lazaridis, I., Sankararaman, S., Fu, Q., Rohland, N., Renaud, G., Erlich, Y., Willems, T., Gallo, C., Spence, J.P., Song, Y.S., Poletti, G., Balloux, F., van Driem, G., de Knijff, P., Romero, I.G., Jha, A.R., Behar, D.M., Bravi, C.M., Capelli, C., Hervig, T., Moreno-Estrada, A., Posukh, O.L., Balanovska, E., Balanovsky, O., Karachanak-Yankova, S., Sahakyan, H., Toncheva, D., Yepiskoposyan, L., Tyler-Smith, C., Xue, Y., Abdullah, M.S., Ruiz-Linares, A., Beall, C.M., Di Rienzo, A., Jeong, C., Starikovskaya, E.B., Metspalu, E., Parik, J., Villems, R., Henn, B.M., Hodoglugil, U., Mahley, R., Sajantila, A., Stamatoyannopoulos, G., Wee, J.T., Khusainova, R., Khusnutdinova, E., Litvinov, S., Ayodo, G., Comas, D., Hammer, M.F., Kivisild, T., Klitz, W., Winkler, C.A., Labuda, D., Bamshad, M., Jorde, L.B., Tishkoff, S.A., Watkins, W.S., Metspalu, M., Dryomov, S., Sukernik, R., Singh, L., Thangaraj, K., Pääbo, S., Kelso, J., Patterson, N., Reich, D., 2016. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature, 13 538 (7624), 201–206.

Marcott, S.A., Shakun, J.D., Clark, P.U., Mix, A.C., 2013. A reconstruction of regional and global temperature for the past 11,300 years. Science 339, 1198–1201. http://dx.doi.org/10.1126/science.1228026.

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K., Luo, Y., Sidore, C., Kwong, A., Timpson, N., Koskinen, S., Vrieze, S., Scott, L.J., Zhang, H., Mahajan, A., Veldink, J., Peters, U., Pato, C., van Duijn, C.M., Gillies, C.E., Gandin, I., Mezzavilla, M., Gilly, A., Cocca, M., Traglia, M., Angius, A., Barrett, J.C., Boomsma, D., Branham, K., Breen, G., Brummett, C.M., Busonero, F., Campbell, H., Chan, A., Chen, S., Chew, E., Collins, F.S., Corbin, L.J., Smith, G.D., Dedoussis, G., Dorr, M., Farmaki, A.E., Ferrucci, L., Forer, L., Fraser, R.M., Gabriel, S., Levy, S., Groop, L., Harrison, T., Hattersley, A., Holmen, O.L., Hveem, K., Kretzler, M., Lee, J.C., McGue, M., Meitinger, T., Melzer, D., Min, J.L., Mohlke, K.L., Vincent, J.B., Nauck, M., Nickerson, D., Palotie, A., Pato, M., Pirastu, N., McInnis, M., Richards, J.B., Sala, C., Salomaa, V., Schlessinger, D., Schoenherr, S., Slagboom, P.E., Small, K., Spector, T., Stambolian, D., Tuke, M., Tuomilehto, J., Van den Berg, L.H., Van Rheenen, W., Volker, U., Wijmenga, C., Toniolo, D., Zeggini, E., Gasparini, P., Sampson, M.G., Wilson, F., Frayling, T., de Bakker, P.I., Swertz, M.A., McCarroll, S., Kooperberg, C., Dekker, A., Altshuler, D., Willer, C., Iacono, W., Ripatti, S., Soranzo, N., Walter, K., Swaroop, A., Cucca, F., Anderson, C.A., Myers, R.M., Boehnke, M., McCarthy, M.I., Durbin, R., 2016. A reference panel of 64,976 haplotypes for genotype imputation. Haplotype Ref. Consortium. Nat. Genet. 48 (10), 1279–1283.

Minkov, M., Blagoev, V., Bond, M., 2014. Improving research in the emerging field of cross-cultural sociogenetics: the case of serotonin. J. Cross-Cult. Psychol. 1–19. http://dx.doi.org/10.1177/0022022114563612.

Minkov, M., Bond, M., 2016. A genetic component to national differences in happiness. J. Happiness Stud. 1–20. http://link.springer.com/article/10.1007/s10902-015-9712-y.

Murray, D.R., Schaller, M., 2010. Historical prevalence of infectious diseases within 230 geopolitical regions: a tool for investigating origins of culture. J. Cross-Cult. Psychol. 41 (1), 99–108.

Nelson, M.R., Bryc, K., King, K.S., Indap, A., Boyko, A.R., Novembre, J., Briley, L.P., Maruyama, Y., Waterworth, D.M., Waeber, G., Vollenweider, P., Oksenberg, J.R., Hauser, S.L., Stirnadel, H.A., Kooner, J.S., Chambers, J.C., Jones, B., Mooser, V., Bustamante, C.D., Roses, A.D., Burns, D.K., Ehm, M.G., Lai, E.H., 2008. The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. Am. J. Hum. Genet. 83 (3), 347–358.

Nordhaus, W.D., 2006. Geography and macroeconomics: new data and new findings. Proc. Natl. Acad. Sci. 103 (10), 3510–3517.

Nordhaus, W.D., Chen, X., 2016. Global Gridded Geographically Based Economic Data (G-Econ). NASA Socioeconomic Data and Applications Center (SEDAC, Palisades, NY. http://doi.org/10.7927/H42V2D1C, Version 4.

Olsson, O., Hibbs, D., 2005. Biogeography and long-run economic development. Eur. Econ. Rev. 49 (4), 909–938.

Rajeevan, H., Soundararajan, U., Kidd, J.R., Pakstis, A.J., Kidd, K.K., 2012. ALFRED: an allele frequency resource for research and teaching. Nucleic Acids Res. 40 (Database issue).D1010–5.

Ramachandran, S., et al. 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc. Natl. Acad. Sci. 102 (44), 15942–15947.

Siddle, K.J., Quintana-Murci, L., 2014. The Red Queen's long race: human adaptation to pathogen pressure. Curr. Opin. Genet. Dev. 29, 31–38.

Spolaore, E., Wacziarg, R., 2016. Ancestry and Development: New Evidence, Mimeo: Tufts University.

Thornhill, R., Fincher, C.J., 2014. The Parasite–Stress Theory of Value and Sociality: Infectious Disease, History and Human Values Worldwide. Springer, New York.

Vogel, F., Chakravartti, M.R., 1966. ABO blood groups and smallpox in a rural population of West Bengal and Bihar (India). Humangenetik 3 (2), 166–180.

Vogel, F., Motulsky, A., 1997. Human Genetics: Problems and Approaches. 2nd, Springer, Berlin.

Wolfe, N.D., Dunavan, C.P., Diamond, J., 2007. Origins of major human infectious diseases. Nature 447, 279–283. http://dx.doi.org/10.1038/nature05775.