2012

Investigating the role of genetic variation in long run economic outcomes

Charles Justin Cook
Louisiana State University and Agricultural and Mechanical College

INVESTIGATING THE ROLE OF GENETIC VARIATION IN LONG RUN ECONOMIC OUTCOMES

A Dissertation

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

in

The Department of Economics

by

Charles Justin Cook
B.S., Louisiana State University, 2006
M.S., Louisiana State University, 2009
August 2012

For Mom and Elizabeth

Acknowledgments

I would like to take this opportunity to thank my advisor Dr. Areendam Chanda. When I came to Dr. Chanda talking about differences in milk consumption, he could have easily directed me to some other area of study; instead, he encouraged me to pursue my interests and showed me the true importance of my topic. For this, I am truly indebted. In addition to the wide variety of encouragement offered by Dr. Chanda, I would also like to thank him for the time he has spent in preparing comments, having detailed discussions about my work, and providing opportunities for me to meet other researchers with similar interests. If not for his initial and continued support, this dissertation would not be possible.

I would also like to thank my committee members. Dr. McMillin thank you for always having an open door; you have given me constant advice in regards to both teaching and research. Thanks to Dr. Unel for a spirited introduction to economic growth. Without your excellent teaching, advice, and book recommendations, my research would be drastically different.

Thanks to Dr. Mocan for research suggestions, encouragement, and introducing me to a number of influential researchers.

Special thanks to all office and class mates who have provided me with numerous ideas and discussions. In particular, I am grateful to Emre, Elif, Colin, Bibhu, Dan, Matt, and Christian.

Finally, but not least, I would like to thank my family. My beautiful wife Elizabeth has supported me both emotionally and financially through this process; you may not have always listened, but you allowed me an opportunity to express my thoughts, which was of invaluable help. Thanks and love to my mother, who has supported me my entire life. It is for these two influential women, for which this dissertation is dedicated.

Table of Contents

List of Tables

List of Figures

Abstract

This dissertation empirically tests whether adaptations resulting from the Neolithic Revolution, or the widespread adoption of sedentary agriculture for sustenance, have led to economic differences. The development of sedentary agriculture constituted an environmental shift from the previous sustenance method of hunting and gathering. This environmental shift resulted in the natural selection of certain traits. I seek to exploit differences in these traits to measure differing economic outcomes. Two main adaptations are considered in this work: the ability to consume milk, or lactose tolerance, and resistance to infectious Eurasian diseases, which is the result of genetic variation.

The first essay establishes a link between lactose tolerance and economic conditions in the pre-colonial era. The ability to digest milk, or to be lactose tolerant, is conferred by a gene variant, which is unequally distributed across the Old World. Digesting milk conferred qualitative and quantitative advantages to early farmers's diets, which ultimately, led to differences in the carrying capacities of respective countries. The second essay investigates the role of genetic differences in resistance to infectious diseases on contemporary health outcomes. The Neolithic Revolution led to the initiation and sustainability of new infectious diseases. The differential timing of the Neolithic Revolution created differences in exposure to these infectious pathogens. Ultimately, this led to differential selection of genetic resistance, in which diversity within a key component of the immune system,

the major histocompatibility complex, was favorable. We evaluate this advantage through the construction of a common measure of genetic diversity that is constructed solely from gene variants within the major histocompatibility complex, known as the human leukocyte antigen (HLA) system in humans. The final essay explores the complementarity between potatoes and milk in explaining the large population growth experienced throughout the Old World in the 18th and 19th centuries.

Chapter 1

Introduction

A recent and growing body of research has been focused on the fundamental, or ultimate, determinants of economic growth.[1] From which, history and geography play an intimate role in the differential patterns of economic growth across states. These differential rates are a contributing factor in the great disparities in income seen across the world today. Therefore, a greater understanding of historical differences in the wealth of states provides a window into understanding contemporary differences.

The development of sedentary agriculture, commonly referred to as the Neolithic Revolution, is key to the acquisition and accumulation of wealth (Diamond 1998; Hibbs and Olsson 20004; Putterman 2008). Before agriculture, humans lived in relatively small, nomadic bands as hunter-gatherers. The switch to agricultural subsistence led to a large change in the environment in which humans lived. Agriculture allowed for the development of large, dense populations, as well as the capability of specialized labor. If one farmer can produce food for many, not everyone in the population needs to produce food. This opens the possibility of a non-food producing class—kings, bureaucrats, priests, soldiers, artisans, etc.—which is not possible in the hunter-gatherer environment of the past.

The current work is concerned with the environmental shift associated with the Neolithic Revolution. For thousands of years humans lived within small, nomadic groups, but starting

[1] See Nunn (2009) for review. See also Diamond (1998), Galor & Moav (2007), Hibbs & Olsson (2004), Putterman (2007), Ashraf & Galor (2012), and Spolaore & Wacziarg (2009) in regards to the current work.

roughly 10, 000 years ago, humans began to live in large, dense societies with limited diets and close contact with a number of domesticate animals. The Neolithic Revolution changed the environment, and this change occurred at differing times for differing groups. This new agricultural environment prompted selection for traits favorable to an agricultural lifestyle. I seek to exploit the natural selection of these traits, and the advantages conferred by them, in order to measure differential levels of development and health in historic and contemporary times, respectively.

Recent research has looked at the effect of general genetic variation in explaining economic phenomena (see e.g., Ashraf & Galor 2012; Spolaore and Wacziarg 2009). These works do not consider the effects of natural selection; rather, they exploit neutral genetic variation, either within or between countries, to explain historic levels of development, contemporary variations in income per capita, and the diffusion of technology across states.[2] My work contributes to the literature by looking at the contribution of traits that were naturally selected, and therefore provided some advantage to the agricultural environment. I find that natural selection has played a role in the economic outcomes, past and present, in the wealth and health of nations.

A prime example of selection since the Neolithic Revolution is the ability to digest lactose. In the next chapter, I explore the role of country level differences in the frequency of lactose tolerance to explain variations in pre-colonial population density, a proxy for economic development due to the Malthusian economy. Lactose tolerance, also referred to as lactase persistence, results from a gene variant that allows for the continual production of lactase in the small intestine. Lactase is the enzyme responsible for digesting lactose, a sugar within milk. Of particular interest is how differences in lactose tolerance have developed. The production of lactase is conferred by a gene variant; this implies the benefits of milk consumption are measured by differences in the genome.[3] The gene variant responsible for

[2]The term "neutral genetic variation" is used to imply genetic differences arising for reasons other than natural selection (i.e., genetic drift, non-random mating, population bottlenecks, etc.).

[3]Measurement of lactase persistence is calculated through phenotypic, or observed ability to consume,

the production of lactase has arisen since the domestication of animals, or the beginning of agriculture, and is commonly used as a textbook example of natural selection within humans since the Neolithic Revolution. My hypothesis is that countries with high densities of lactose tolerant individuals were able to use an additional resource, milk. The use of this resource provided qualitative and quantitative benefits to the diet, which in turn, led to a greater carrying capacity for a given piece of land and, therefore, denser populations. I find that for a one standard deviation increase in the frequency of lactose tolerance, or lactase persistence, is associated with a 60% increase in population density in 1500 CE.

For the third chapter I explore how infectious disease has shaped the genome since agriculture. My specific question relates to the disparity in the efficaciousness of disease after the Columbian Exchange. Particularly, why did Old World diseases debilitate New World populations, while the opposite did not occur? Disease disparities arose from differences in population density and domesticate animals (Wolfe et al. 2009). Pathogens of domesticate animals mutated to affect readily available human hosts. Additionally, the presence of the resulting diseases was dependent on the number of hosts available, in which diseases would move rapidly through less dense populations, exhaust all potential hosts, and die out. [4] New World populations did not have a large number of domesticable animals from which to initiate infectious disease. Additionally, the population density of the New World was relatively low when compared to the Old World. This implies historical exposure to infectious disease varied between continents. In short, peoples of the Old World have had a greater exposure to specific diseases from domesticate animals. This greater exposure has led to the natural adaptation of the genome to these diseases. In other words, peoples of the Old World have an inherent genetic resistance that is not present in New World populations.

---

not through differences in the genome.

[4]Following epidemiological models, potential hosts are either susceptible to the disease, infected by the disease, or recovered from the disease. In the diseases considered, a recovered individual cannot be infected again.

With this understanding, I explore the natural selection of a key component of the human immune system, the human leukocyte antigen (HLA) system. Previous work has shown that exposure to the large number of diseases present within Eurasia led to balancing selection within the HLA system, resulting in a high level of genetic variation (Black 2004; Hughes and Yeager 1998; Jeffrey and Bangham 2000; Prugnolle et al. 2005). I therefore create a measure of genetic variation within HLA system for use in measuring inherent genetic resistance to the disease that developed within Eurasia. This measure is given for differing ethnicities that are then aggregated to the country level, allowing for a cross-country comparison. After controlling for other relevant determinates of health, I find that this measure of variation has a positive and statistically significant effect on pre-medicinal life expectancy.

The fourth chapter builds upon the topic of Chapter 2. In a recent work, Nunn and Qian (2011) show that the introduction of the potato led to larger population in the Old World. My work argues that the efficaciousness of the potato is dependent on milk consumption. Milk and potatoes are complements in the diet: milk provides fats, proteins, and vitamins, while potatoes provide a high number of calories. Therefore, the potato should have a greater effect on populations that are able to consume milk. This thought is echoed by the following statement in Nunn and Qian (2011, p. 601): "[A] single acre of land cultivated with potatoes and one milk cow was nutritionally sufficient for feeding a large family of six to eight." I find that potato consumption was complemented by milk consumption, where the marginal effect of potatoes is doubled if a country's entire population is able to consume milk.

In summary, the Neolithic Revolution constituted a major environmental shift for humans. This shift occurred at differing times for differing peoples, with Eurasia having an advantage in both initiating and spreading agriculture. This implies adaptation to this new environment also differs. I seek to exploit the economic outcomes of this differential adaptation, from which I show lactose tolerance provided historical benefits in the accumulation

of wealth and populations. In addition to the benefits of lactose tolerance, I also show that historic disease environments have an effect on pre-medicinal health outcomes through differences in a key component of the genome, where these differences are the product of the continual selection of individuals best able to subsist the wide array of infectious Eurasian diseases.

Chapter 2

The Role of Lactose Tolerance in Pre-Colonial Development

2.1 Introduction

The great disparities in productivity that are seen throughout the world today are not new. As of 500 years ago great variations in technology, state development, and industry were obvious across states and continents; most notable is the distinction between Europe and Sub-Saharan Africa. Europe was in the middle of the Renaissance, had complex systems of state organization, numerous divisions of labor, and was making great strides in seafaring, while Africa was vastly under populated and relatively under developed. What are the causes of variations in historic development? It is known that Eurasia contained advantages in initiating and spreading agriculture, but are there other factors which led to larger pre-colonial populations? Why did Europe in particular have an advantage over other Eurasian states? This paper argues the variation in an important food source, milk, is significantly related to differences in pre-colonial development, or pre-colonial populations.

The Neolithic Revolution radically changed the environment for humans.[1] Furthermore, this change occurred at different times for different peoples; implying, certain groups have had a longer time to evolve, or adapt, to the new environment. In the words of Clark (2008,

[1]The Neolithic Revolution is the name given to the transition from hunting and gathering to agriculture

P. 6; Galor and Moav 2002), "The Darwinian struggle that shaped human nature did not end with the Neolithic Revolution but continued right up until the Industrial Revolution." A major adaptation to the sedentary agricultural lifestyle is the ability to consume milk, or to be lactose tolerant. Milk was an additional resource that some could consume, while others could not. In the Malthusian economy of the pre-colonial era, this variation in the consumption of milk should be associated with variations in the productive capacity of land. Specifically, we seek to explain the differences in population density for 1500 CE using the fraction of lactose tolerant individuals within a country.

Lactose tolerance data is available by ethnicity for the second half of the twentieth century. A central assumption in our paper is that this has not changed much over the past 500 years (Section 2 includes a detailed discussion of the validity of this claim). Since our hypothesis concerns pre-colonial development, we also need a measure of ethnic composition for 1500 CE, which is not directly available. We follow two strategies. The first, and primary strategy, involves post-multiplying the matrix of current ethnic compositions countries with the inverse of a matrix that captures human migration from 1500 to 2000 CE (Putterman and Weil 2010). This, in theory, gives county level ethnic compositions for the year 1500 CE. In order to confirm our results, we also use a cruder strategy of assigning majority ethnic groups to represent countries in the 1500 CE.[2]

We show that our constructed measure of lactose tolerance has a positive and significant effect on population. Specifically, our baseline estimate states that a standard deviation increase in the fraction of lactose tolerant individuals within a country is associated with a 60% increase in pre-colonial population density. The results are robust to a large number of geographical and environmental variables. In particular, we show that the effect of lactose tolerance does not pick up the overarching advantages of earlier transitions into agricultural societies that have been documented extensively. The results are also robust to other measures of genetic distances that have been used to explain technological diffusion across

[2]This strategy is pursued in similar research, i.e., Spolaore and Wacziarg (2009).

countries, as well as variables that capture other environmental or cultural determinants of pre-colonial development. In addition to least squares estimation, we also consider an instrumental variables approach. Lower levels of sunlight result in a defficiency of vitamin D. A diet that is rich in milk can offset the harmful benefits of vitamin D deficiency through the addition of absorbable calcium (Flatz and Rotthauwe 1973). Therefore, we use a measure of solar radiation to instrument country level differences in the frequency of lactose tolerance. Due to concerns, however, our use of IV estimations are not meant to replace estimates through OLS; the use of instrumental variables is intended to supplement and confirm the relationship between dairying and pre-colonial population density.

An interest in the role of history in explaining economic disparities has recently been renewed. The idea that current development levels are path dependent has established the search for a more ultimate understanding of the long run causes of growth; knowing the causes of small differences in past growth rates gives valuable insights into the cross-country disparities in current economic conditions. According to Nunn (2009, P. 88): "The main fact . . . is that history matters." Specifically, a number of papers have established an empirical link between past and current economic events, where it is shown that variations in the past have economic repercussions that are felt today (see, e.g., Acemoglu et al. 2001; Bockstette et al. 2002; Chanda and Putterman 2004; Comin et al. 2007; Engerman and Sokoloff 1997, 2002; La Porta et al. 1997, 1998; Nunn 2008). The current work seeks to build upon this research.

One of the most comprehensive works in explaining pre-colonial populations and, therefore, pre-colonial development is Jared Diamond's *Guns, Germs, and Steel* (1997). Diamond's main argument is that societies on the Eurasian continent contained a geographical advantage in both initiating and spreading agriculture. In particular, the geographical advantages of Eurasia are the number of domesticable species (plants and animals) and the East-West orientation of the continent, where the former is associated with an ease of initiating agriculture and the latter an ease of agricultural diffusion. These advantages allowed

for an earlier transition to, and a more widespread use of, agricultural practices; which in turn, allowed for mass populations, the development of cities and states, the specialization of labor, and, ultimately, a head start in the acquisition of prosperity. Diamond's hypothesis is tested by Putterman (2007) and Hibbs and Olsson (2004), who find a positive correlation between agricultural transition dates and wealth levels in 1500 CE. The most tangible difference between the two papers is in the way agricultural transition dates are calculated: Putterman uses archeological facts in calculating the dates for particular countries, while Hibbs and Olsson use biogeographic and geographic conditions in order to estimate the transition dates for regions. Diamond's argument, however, does not give reason as to why variations within Eurasia may develop. This paper seeks to supplement Diamond's by providing a possible explanation to within levels of development; particularly, we use the varied use of milk as an explanation of varied levels of development throughout the Old World.

Instead of archeological evidence or environmental estimates, we use an observed genetic difference between societies as a predictor of past economic development. This genetic difference is primarily driven by differences in culture; and through the process of natural selection, this information has been passed through generations of humans until today. Diamond states: "History followed different courses for different peoples because of differences among peoples' environments, not because of biological differences among peoples themselves." A difference in environments, however, is the main cause in divergent evolutionary paths, according to Darwin: "In the struggle for survival, the fittest win out at the expense of their rivals because they succeed in adapting themselves best to their *environment*."[3] Therefore, a difference in environments, including both cultural and geographical differences, allows for differences in genetic adaptations. Conversely, the use of genetic variation may be used as an indicator of the usage or availability of a cultural or environmental advantage conferred to some societies and not others.

The effects environmental changes have on evolution are numerous and well documented.

[3]Emphasis my own.

The most common example involves the peppered moths of England before and after the industrial revolution (Kettlewell 1956). Before the revolution light colored moths were the vast majority due to camouflauge provided by light colored trees; however, the industrial revolution caused dark soot to form on the trees causing lighter colored moths to stand out. The darkening of the trees allowed for the darker variety of the peppered moth to have a greater relative probability of survival, thereby increasing the frequency of dark moths compared to light. Just as the dark colored moths had an advantage after the environmental shift, those peoples who were able to capitalize the additional resource of milk were able to increase their numbers relative to those who were unable to digest lactose.

A number of recent papers explore the effect that genetics may have on aggregate economic outcomes (see, e.g., Ashraf and Galor 2008; Spolaore and Wacziarg 2009; Michalopolous 2008). In general, these papers use broad genetic variation measures between, and within, particular countries to explore differing economic outcomes, historic and current. This paper differs by the use of a particular gene variant, not differences in the general genetic make-up of a population. In particular, the current work uses variation in an expressed genetic trait which has been naturally selected for since the Neolithic Revolution. To our knowledge, this is the first paper to explore the effect of a particular gene variant expression has on aggregate economic conditions.

A similar work by Nunn and Qian (2011) explores how the introduction of the potato to the Old World has affected populations in the 18th and 19th centuries. Specifically, they show that exogenously determined soil conditions, which are favorable for potato production, account for 25%-26% of the population increase from 1700 to 1900 and 27%-34% of the increased urbanization rate in the same time period. Both the current work and that of Nunn and Qian explore how the addition, or varied use, of a particular food source affects historic populations. A slight difference, however, is found in quantifying the spread of the respective food sources; Nunn and Quian use soil conditions, whereas we use the observed differences of an underlying genetic variation.

Natural selection since the Neolithic Revolution has been studied in recent research. Theoretically, Galor and Moav (2002) establish a unified model that captures the evolution of both man and economic outcomes, while Galor and Michalopolous (2011) show the selection for particular traits since the beginning of agriculture. Empirically, Galor and Moav (2008) show adaptation since the initiation of agriculture has a statistically significant relationship with contemporary variations in aggregate health measures. The work of Galor and Moav (2008) implies that differences have developed since the Neolithic Revolution and that these differences may be correlated with differing economic outcomes. This is this attitude that we seek to capture. Particularly, the variation in the timing of the Neolithic led to a variation in the genetic ability to consume milk.

2.1.1 Population Advantages of Milk Consumption

The consumption of milk today ranges from cows in Europe, America, Australia, and Africa to camels and goats in the Middle East, reindeer in the Arctic, mares and asses in the Eurasian steppe, and water buffalo in Southeast Asia (WHO 2009). [4] There is considerable evidence that milk stimulates growth, increases bone density, and provides essential vitamins and minerals (Hoppe et al. 2006). Milk is an incredibly complex liquid that contains fats, proteins, vitamins, and minerals; as the popular slogan states (McCann-Erikson 1990): "Milk: It Does a Body Good!"Along with these qualitative advantages, milking also allowed early farmers and pastoralists to obtain a greater number of calories from a fixed number of cattle. Through the qualitative and quantitative attributes of milk, greater populations could be supported for a fixed quantity of land.

A sugar found in milk, lactose, is responsible for the exclusivity in consumption. The enzyme required to break down lactose, lactase, is found within the small intestine.[5] If this enzyme is not present, the lactose will pass to the colon causing diarrhea or cramping to

[4]For simplicity we reference milk to be from cattle.

[5]Lactose is found in all milk

occur (Simoons 1969). Like all mammals, humans produce lactase from birth until the end of weaning in order to digest the numerous nutrients that are passed from mother to offspring.[6] Certain populations of humans, however, have developed an allele, or gene variant, that allows for the production of lactase throughout their adult lives; this is known as lactase persistence.[7] Considering that the vast majority of humans, and all other mammals, are unable to produce lactase beyond the weaning period, it must be the case that the inability to drink milk into adulthood is the original state (Simoons 1969). Accordingly, the ability to digest milk, or to be lactase persistent, is one of the most famous cases for continued evolution in humans (Ingram et al. 2009).

The quantitative advantages in the ability to digest lactose are apparent. Consider two farmers (or pastoralists) with identical numbers of cattle (or some other milk producer). One of the farmers is able to digest milk, while the other is not. The farmer who is able to digest milk immediately gains an additional resource from his set herd of cattle. Moreover, the farmer who is able to digest milk can now support a larger family, which in turn has the effect of increasing the population and increasing the percent of lactase persistence within the population.

It isn't necessarily the case that strict specialization in milk production is required to increase population densities. This paper argues that the supplementation of the additional resource is enough to improve pre-colonial population levels. Horticulture can supply vastly more calories per acre than any husbandry technique (Cooper and Spillman 1917). A homogenous diet of a few grains, however, led to adverse health effects in early farmers (Cohen and Armelagos 1984). The addition of fats, proteins, vitamins, and minerals found in milk provided a healthy balance to the early farmer's diet, which, in turn, allowed for longer lives and greater populations. According to the World Health Organization (2009, p. 3): "The profile of amino acids in milk complement those in grains and cereals, which is of

[6]Weaning is the process of an infant taking nourishment other than by suckling.

[7]As is consistent with the literature, we will use lactase persistence instead of lactose tolerance. Although, the two terms have equivalent definitions.

considerable benefit in communities where grains and cereals predominate." Additionally, Nunn and Quian state (2011, p. 7): ". . . a single acre of land cultivated with potatoes and one milk cow was nutritionally sufficient for feeding a large family of six to eight."[8] Considering two societies with equal resources, the society that is able to digest milk gains a qualitative dietary advantage that increases health and, therefore, population.

Milk provided both quantitative and qualitative advantages to the early farmer's diet, which, respectively, can be seen as a substitute or a complement to a farmer's diet. Again consider two identical farmers: one can digest milk while the other cannot. The farmer who is able to digest milk is able to sustain solely on the caloric output that milk provides—i.e., milk is a substitute for other food sources. The farmer is also able to supplement needed vitamins, minerals, and other essential nutrients, which a staple crop provides an insufficient amount—i.e., complementing the farmer's current diet. Both effects would increase pre-colonial populations.

In addition to the direct effects of consumption, the availability of milk may have increased the fecundity of early sedentary women. Postpartum amenorrhea, or infertility, is positively related to the length of time an infant weans (Jain et al. 1970). The use of milk as a substitute for mother's milk would have reduced weaning time and, therefore, the postpartum infertility period.[9] Implying, a mother who had access to milk would have been able to give birth to a larger number of children over her life span, which corresponds to the positive relationship between dairying and populations.

2.1.2 Selection for Lactase Persistence

The Neolithic Revolution radically changed the environment for humans, and this change has occurred at different times for different peoples. This implies that certain groups have

[8]Potatoes are nutritionally advantageous to other Old World staple crops, which implies the inclusion of milk is complimetary no matter the nutritional value of the staple crop.

[9]All infants produce lactase in order to digest mother's milk.

had a longer time to evolve, or adapt, to the new environment, and one adaptation is the continued production of lactase. Burger et al. (2007) have shown that the allele, or gene variant, that allows for lactase persistence in Europeans is absent, or rare, in early Neolithic Europeans. Considering that Europeans have the highest levels of lactase persistence in the world, this implies that the ability to digest lactose into adulthood is a new phenomenon that gives a significant advantage to its possessors. Toward this end, Bersaglieri et al. (2004) find that the differences in lactase persistence frequencies are due to a strong positive selection of an allele that allows for milk consumption occurring in the past 5,000-10,000 years, a time range that is consistent with the domestication of cattle and other milk producing domesticates. Furthermore, the gene variant that confers lactase persistence is the "textbook" example of a selective sweep (Nielsen et al. 2005; Ingram et al. 2009).[10]

Most gene mutations that occur do not confer any type of advantage. If, however, a gene mutation gives an advantage, then its possessor is more likely to survive and, in turn, produce more children. This process continues over time, with a larger and larger portion of the population containing this mutation, i.e. the allele is being naturally selected. Or in the words of Darwin:

> Owing to this struggle for life, variations, however slight and from whatever cause proceeding, if they be in any degree profitable to the individuals of a species, in their infinitely complex relations to other organic beings and to their physical conditions of life, will tend to the preservation of such individuals, and will generally be inherited by the offspring. The offspring, also, will thus have a better chance of surviving, for, of the many individuals of any species which are periodically born, but a small number can survive. I have called this principle, by which each slight variation, if useful is preserved, by the term Natural Selection.

Given the fast increase in the frequency of lactase persistence, then it must be the case

[10] A selective sweep is defined as, "The process in which a favorable mutation becomes fixed in a population (Hartl and Clark, P. 184)."

that digesting lactose did provide an advantage for the owners of a lactase producing gene variant. Bersaglieri et al. (2004) find that the ability to continually produce lactase has a selective advantage between .014 and .15: this implies that a population of 1,000 individuals that are able to produce lactase throughout their lives will have between 14 and 150 more offspring per generation compared to individuals without the ability to digest lactose.[11]

If no cattle were available, and therefore no milk, then there would be no advantage to producing lactase. This implies further that the availability of milk is a necessary condition for the rise in frequencies of lactase persistence. This co-evolution of dairying and lactase persistence is formally known as the "Cultural Historical Hypothesis" and is attributed to Simoons (1969). According to Simoons:

> Such an advantage most likely would occur in groups, not necessarily pastoral, that not only enjoyed a plentiful milk supply, but that had other foods inadequate in amount and quality, and that did not process milk into products low in lactose. Under these conditions, the lactase aberrant adults would better multiply, and would more successfully defend their families against others. And in their numerous descendants, high levels of adult lactase activity would come to prevail.

The "Cultural Historical Hypothesis" has received considerable attention lately with the discovery that the origination of lactase persistent alleles have coincided with the proposed dates of the domestication of cattle (Coelho et al. 2005, Mulcare 2006, Bersaglieri et al. 2004, and Tishkoff et al. 2007).

This indicates that the frequency of lactase persistence may just be a proxy for the origination of animal husbandry; whereby the frequency of lactase persistence is an increasing function of the years since the domestication of a particular mammal. While it is true that the availability of milk, or cattle, is a necessary condition for the evolution of lactase persistence, it is not, however, a sufficient condition. Southern Europe, Eastern Europe, the Near East, and the Middle East have had access to milk for as long, or longer, than Western

[11] This is dependent on the availability of milk. If no milk is available; no advantage exists.

Europeans, yet these areas have significantly lower levels of lactase persistence (Simoons 1978). This indicates that differences in dairying also have a cultural significance. For this reason, the use of lactase persistence frequencies does not measure the initial advantages of obtaining cattle; it measures the initial advantages of milking.

In summary, the gene variant that allowed its possesors to consume milk did provide an advantage. One question this work seeks to answer is whether or not this advantage led to differential economic outcomes. The next section provides a detailed explanation of the cross-country measure of lactase persistence.

2.2 Data

2.2.1 The Frequency of Lactase Persistence

Milk consumption has independent origins across the Old World, which has resulted in a number of gene variants, or alleles, responsible for the production of lactase (Ingram et al. 2009; Tishkoff et al. 2006). Further, the frequency of a particular variant is ethnic specific. In other words, the gene variant that allows for milk consumption in Northern Europeans is not identical to the allele that allows for milk consumption in Western Africans. It is for this reason that the observed, or phenotypic, ability to consume milk is the primary determinant of our measure of lactase persistence.[12][13]

The data for the frequencies of lactase persistence come from Ingram et al. (2009), in which the authors aggregate data from past studies of lactase persistence frequencies. The data are given at the ethnic level. The lactase persistence frequencies are obtained

[12] A phenotype is the physical expression of a genotype (Hartl and Clark 2007).

[13] A measure of the frequency of lactase persistence has been calculated by using the frequency of the gene that allows for the continued production of lactase in European populations. Substituting this measure into the estimating equation specified above leads to a positive and signficant coefficient, but the use of the European gene frequency is sensitive to the inclusion of a number of controls. This is to be expected, due to the genes positive relationship with milk consumption in Europeans and nonexistent relationship with milk consumption in all other ethnic populations, which results in a large measurement error on the explanatory variable of interest and an attenuation of the coefficient.

by conducting lactose tolerance tests on samples from an indigenous population. The data are collected from 1965 to 2007. While the tests do span a relatively large time scale, the testing methods used remain constant, and the gene frequencies themselves should have also remained constant over this relatively short period. There are two ways to test for lactase persistence: blood glucose and breath hydrogen. In both tests individuals are given lactose after an overnight fast in order to accurately conduct the tests. A description of the two tests from Ingram et al. (2009):

> A baseline measurement of blood glucose or breath hydrogen is taken before ingestion of the lactose, and then at various time intervals thereafter. An increase in blood glucose indicates lactose digestion (glucose produced from the lactose hydrolysis is absorbed into the bloodstream), and no increase, or a 'flat line' is indicative of a lactose maldigester...
>
> An increase in breath hydrogen indicates maldigestion and reflects colonic fermentation of the lactose...

The arbitrary cutoff levels in defining digesters and maldigesters, or, respectively, lactase persistence and non-lactase persistence, imply that measurement errors will be present.

2.2.1.1 Estimating the Ethnic Composition of Countries in 1500 CE

In creating a country wide measure for lactase persistence frequencies, two problems need to be overcome. First, we need to aggregate ethnic groups into countries. And secondly, I will need to scale this measure back 500 years as to measure the effect of lactase persistence on pre-colonial development.

In order to aggregate ethnic groups into country level measures, data on the ethnic make-up of countries is used from Alesina et al. (2003). The data from Alesina et al. (2003) give the ethnic composition of 190 countries from roughly 1990 to 1995. Using ethno-linguistic classifications, ethnic groups, which have lactase persistence frequencies from Ingram et al. (2009), have been matched to ethnic groups in Alesina et al. (2003). For example, "Western Europeans" in Sweden from Alesina et al. (2003; hereafter Alesina)

are assigned the lactase persistence frequency of "Dane" from Ingram et al. (2009; hereafter Ingram), "Filipinos" in Alesina are assigned to the "Maori" ethnic group in Ingram, and the "Fon" people from Benin are assigned to "Yoruba" from Nigeria.[14] This matching yields data for 118 Old World countries (i.e., Europe, Asia, and Africa) , of which 51 countries have a direct match between the majority ethnic group given by Alesina and ethnic data from Ingram. An additional level of measurement error is to be expected from using ethnolinguistic classification in the matching of ethnic groups. As a result the 51 countries that have exact matches are considered to be more conservative estimates of the country level lactase persistence frequencies, and separate estimations are performed using the reduced sample.

The aggregation from ethnic groups to countries gives a cross-country measure of the lactase persistence frequency; however, this measure is for the present period and may not be relevant in the prediction of variables in the pre-colonial period. A cross-country measure for lactase persistence 500 years in the past is needed. One way around this problem is to ascribe the largest ethnic group within a country as the country's sole ethnic group in the year 1500 CE(Spaloare and Wacziarg 2009). A cross-country lactase persistence frequency is calculated in this manner with one exception: if an ethnic group does not constitute over 60% of a countries present day composition and another ethnic group constitutes over 30% of the countries composition, the country's ethnic composition in 1500 CE is ascribed as 50% to each group. For example, Belgium's present ethnic composition from matching ethnic groups in Ingram to Alesina is given to be 58% German and 30% French, so in calculating ethnic composition in the year 1500, 50% is ascribed to German and 50% is ascribed to French. Lactase persistence frequencies for 126 countries are found in this manner with 54 countries having exact ethnic matches.

Our primary way of calculating country level ethnic compositions in 1500 CE involves

[14]Swedes and Danes belong to the East Scandinavian branch of the Indo-European language group, Filipinos and Maori belong to Malayo-Polynesian branch of the Austronesian language group, and the Fon and Yoruba belong to the Volta-Niger branch of the Niger-Congo language group.

using data on migration frequencies over the period 1500 to 2000 (Putterman and Weil 2010). If it is known where a county's current population has migrated from over the past 500 years, it is possible to effectively remove this fraction of immigrants from the current population, leaving a rough representation of the population in the year 1500 CE. Consider an $m \times n$ matrix, $E^{1500}_{m \times n}$, which contains the ethnic composition of countries in the year 1500 with m ethnic groups and n countries. If we take the product of $E^{1500}_{m \times n}$ and the $n \times n$ Putterman and Weil matrix of migration (denoted as $M^{1500-2000}_{n \times n}$), this should give a rough estimate of the ethnic composition today. For example, consider China and Malaysia, which were respectively composed of the Han and Maori groups in 1500:

$$E^{1500}_{m \times n} = \begin{array}{c|cc} & \text{Malaysia} & \text{China} \\ \text{Han} & 0 & 1 \\ \text{Maori} & 1 & 0 \end{array}$$

The matrix $E^{1500}_{m \times n}$ states that in 1500 CE the entire population of China is ascribed to the Han ethnic group and the entire 1500 population of Malaysia is ascribed to the Maori ethnic group. Migration over the last 500 years is given by:

$$M^{1500-2000}_{n \times n} = \begin{array}{c|cc} & \text{Malaysia} & \text{China} \\ \text{Malaysia} & 0.75 & 0 \\ \text{China} & 0.25 & 1 \end{array}$$

which says that 75% of Malaysia's population is derived from Malaysia and 25% of Malaysia's population has immigrated from China. And given that in 1500 China was entirely composed of the Han ethnic group and Malaysia was entirely composed of the Maori ethnic group, this implies that Malaysia's current ethnic composition is 75% Maori and 25% Han.

This is shown by:

$$A^{2000}_{m\times n} = E^{1500}_{m\times n} \times M^{1500-2000}_{n\times n} = \begin{array}{lcc} & \text{Malaysia} & \text{China} \\ \text{Han} & 0.25 & 1 \\ \text{Maori} & 0.75 & 0 \end{array}$$

However, we are interested in finding $E^{1500}_{m\times n}$ given $A^{2000}_{m\times n}$, which is found through methods described above using data from Alesina et al. (2003), and $M^{1500-2000}_{n\times n}$, which is given in Putterman and Weil (2010). In particular, post multiplying $A^{2000}_{m\times n}$ by the inverse of $M^{1500-2000}_{n\times n}$ gives $E^{1500}_{m\times n}$. In our example with Malaysia and China:

$$\begin{aligned} E^{1500}_{m\times n} &= A^{2000}_{m\times n}(M^{1500-2000}_{n\times n})^{-1} \\ &= \begin{pmatrix} & \text{Malaysia} & \text{China} \\ \text{Han} & 0.25 & 1 \\ \text{Maori} & 0.75 & 0 \end{pmatrix} \times \begin{pmatrix} & \text{Malaysia} & \text{China} \\ \text{Malaysia} & 1.33 & 0 \\ \text{China} & -0.33 & 1 \end{pmatrix} \\ &= \begin{pmatrix} & \text{Malaysia} & \text{China} \\ \text{Han} & 0 & 1 \\ \text{Maori} & 1 & 0 \end{pmatrix} \end{aligned}$$

In theory, post multiplying current country level ethnic compositions by the inverse of the Putterman and Weil migration matrix should remove all migration that has occurred over the last 500 years. This process, however, assumes an equality of migration across ethnic groups. It is improbable that migrations were ethnically equal. This problem is partly mitigated due to the high correlation between ethnicity and state in 1500 CE; e.g., France was entirely composed of French, Zimbabwe was entirely composed of Bantu, etc. Comparing lactase persistence frequencies calculated through inverting the migration matrix

to frequencies calculated through majority ethnic groups yields a correlation of roughly 98%. Assuming equality in migration appears to be a minor issue.

2.2.1.2 Monotonicity of Lactase Persistence

Although the ethnic composition of countries is somewhat mitigated due to the inversion of the migration matrix, it still remains that the frequencies of lactase persistence themselves are found roughly 500 years after the dependent variable to be explained. The main issue concerns the monotonicity, or relative relationships, of lactase persistence frequencies over the last 500 years. In order to create a false, positive relationship, either countries that were lightly populated in 1500 CE should have had a comparative decline in lactase persistence over the past 500 years, or relatively rich countries should have had a comparative increase in the frequency of lactase persistence.

To understand any potential biases that may occur, it is important to understand how gene frequencies come about. According to population geneticists three main variables affect how the frequency of a gene evolves: the selective, or survival, advantage conferred by the gene variant, the initial population containing the gene variant, and time (Hartl and Clark 2007). Considering all countries in the sample have the same time constraints, any differences in the frequency of lactase persistence must be attributed to either differences in the initial population containing the gene variant or the selective advantage conferred by the allele.

There is no valid reason to suspect variation in the lactase persistent allele prior to the domestication of milk producing animals. The availability of milk determines whether or not lactase persistence provides an advantage; if there is no milk, then there is no advantage, and according to the laws of natural selection: if there is no advantage, then a gene will not rise in frequency (Hartl and Clark 2007). This principle is shown in the absence of the lactase persistent allele in Europeans prior to the Neolithic Revolution (Burger et al. 2007). The possibility does remain, however, that migrations over the past 500 years have distorted

the respective genotype of a country. This potential source of bias is partially corrected for by in the methods described above; although, introgression, or the exchange of genes from interactions in the migrant and native populations, may have altered the respective native genotype for a particular country. For this to create a bias in my estimation, the lactase persistent allele would have to be passed only to densely populated countries, which seems unlikely. Further dimming the possibility of bias estimation is given by the inverse relationship between the size of a population and the speed at which a gene frequency rises (Hartl and Clark 2007).

Everything else constant, differences in the selective advantage of lactase persistence will cause differences in the speed in which the frequency of the population obtains the gene (Hartl and Clark 2007). Consider again the peppered moths of England. The advantage of the darker moths was dependent on the level of soot within a particular area: The greater the soot, the greater the advantage of having a dark complexion. Dark moths had a greater reproductive advantage relative to light moths in the darker areas, which in turn caused their numbers, or frequency, to increase at a faster rate in these areas. This same idea can be applied to the advantage conferred by the ability to digest milk, where differing areas could confer differing advantages which could cause a non-monotonic relationship to develop between lactase persistence frequencies today and lactase persistence frequencies in the year 1500 CE.

One potential source of a differing selective advantage arises from the environment in which the gene evolved. Flatz and Rotthauwe (1973) theorize that differences in the frequency of lactase persistence are caused by differences in exposure to ultra violet light. Countries with low levels of sun exposure lack the necessary ultraviolet light to adequately synthesize Vitamin D. Deficiency in vitamin D is associated with rickets, or a weakening of bones. The inclusion of milk, which is high in calcium, offset the negative effects of vitamin D deficiency. This implies a greater advantage for milk in areas with lower sunlight; therefore, lactase persistence should rise to a greater frequency in these areas. A number

of controls are used to account for this potential source of bias. These include a Western European dummy and the distance from the equator. In addition to the control variables, a sample truncation is conducted, in which all Western European states and all countries above and below the sample median distance from the equator are excluded. As an extension of this hypothesis, we consider sunlight to be an exogenous determinant of differences in the frequency of lactase persistence. With adequate controls to partial out the effect of sunlight on population density, we use a measure of solar radiation as an instrument for differences in the distribution of lactase persistence. This is further discussed in Section 3.2.2.

Conversely, it could be the case that moderately populated countries, which contained high frequencies of lactase persistence in 1500 CE, faced a situation in which the selective advantage to consuming milk became negative or nonexistent. There is currently no backing for any hypothesis suggesting a negative selective advantage associated with lactase persistence.[15] It is possible, however, that a particular country has lost its milk producing mammals in the past five hundred years, effectively giving no advantage to the ability of drinking milk. According to the Hardy-Weinberg principle, if a gene possesses no selective advantage its relative frequency should remain constant, not decline.[16] Indicating that if a country did lose its cattle stock in the last 500 years, the frequency of lactase persistent individuals within the country should have remained constant; further implying the improbability of a false relationship between lactase persistence frequencies and 1500 CE population density.

Table 2.1: Summary Statistics

Variable:	N	Mean	Std. Dev.	Min	Max
Lactase Persistence Frequency (Inverse of Migration Matrix)	103	0.4188	0.2447	0.0233	0.96
Conservative Lactase Persistence Freq. (Inverse of Migration Matrix)	48	0.4415	0.2608	0.0484	0.96
Lactase Persistence Frequency (Majority Ethnic Group)	103	0.4136	0.2588	0	0.96
Conservative Lactase Persistence Freq. (Majority Ethnic Group)	48	0.4385	0.2741	0	0.96
Population Density in 1500 CE	103	7.7954	9.5798	0.1429	45.9462
Millennia of Agriculture	103	5.3631	2.4103	1	10.5
Mean Suitability of Agriculture	103	0.4226	0.2482	0.003	0.9557
Distance from the Equator	103	29.1651	17.4126	0	60
Sub-Saharan Africa (Dummy)	103	0.3107	0.465	0	1
Western Europe (Dummy)	103	0.1359	0.3444	0	1
Genetic Distance from the U.K. in 1500 CE	103	0.8604	0.7529	0	2.288
Mean Crop Suitability for Potatoes	103	1.8149	4.8215	0.001	35.9686
Mean Crop Suitability for Old World Crops	103	6.4004	8.8663	.001	48.5282
Mean Crop Suitability for New World Crops	103	7.0632	14.9413	.001	116.2154
Solar Radiation ($kWh/m^2/day$)	103	4.623	1.1346	2.4225	6.6708
Elevation (Country Average in km)	103	0.6552	0.6113	0.0301	3.1859
Ruggedness (Country Average)	103	1.3064	1.2688	0.037	6.202
Mean Distance from Coast or River (in km)	103	0.3658	0.4535	0.0227	2.2917
% of Land within 100 KM of Coast or River	103	0.4242	0.3743	0	1
% of Land within the Tropics	103	25.4273	38.604	0	100
% of Land within a Desert	103	4.863	12.5898	0	77.28
Mean of Malarial Ecology Index	103	3.9254	6.9399	0	31.639

Notes: Lactase Persistence Measures calculated from Ingram et al. (2009), Alesina et al. (2003), and Putterman and Weil (2010). Population Density data are given by persons per km^2 and are from McEvedy and Jones (1978). Mean suitability of agriculture is from Michalopolous (2010) and Ramankutty et al. (2002). Distance from the Equator comes from Rodrik et al. (2002). Genetic distance is from Spolaore and Wacziarg (2009). Mean crop suitability for potatoes, New World staples, and Old World staples come from Nunn and Qian (2011). Solar Radiation data come from the Atmospheric Science Data Center at NASA. Elevation, mean distance to a coast or river, % of land within 100 km of a coast or river, and other water access controls are from Gallup et al. (1999). Ruggedness is from Nunn and Puga (2011). The malarial ecology index is from Kiszewski et al. (2004). Genetic distance, crop suitability data from Nunn and Qian, distance to a coast or river, and elevation have all been scaled by 1/1000.

2.2.2 Data: Summary and Sources

Using the ethnic compositions given by the inversion of the migration matrix, I am able to create a lactase persistence measure for the year 1500 CE; this is the primary measure of lactase persistence to be used. This method yields 118 countries, of which 51 have exact ethnic matches. Table 2.1 presents the descriptive statistics for the frequency of lactase persistence as well as all control and dependent variables. The mean frequency of lactase persistence in the base sample is 41.3%, which is similar to the world mean of 35% given by Ingram et al. (2009).[17] Figure 2.1 gives a shaded map of Old World lactase persistence frequencies. As expected lower frequencies of lactase persistence occur in Sub-Saharan Africa while higher frequencies are reported in Western Europe, Scandinavia in particular, with a max sample frequency of 96% in Sweden and a min of 2.33% in Zambia. Figure 2.2 gives historical areas of milking and non-milking. Comparing Fig. 2.1 and Fig. 2.2, there appears to be a relatively tight fit between historically non-milking areas and low levels of lactase persistence.

The main variable to be explained is population density in 1500 CE. This variable is from McEvedy and Jones (1978). Thomas Malthus's seminal work on the relationship between population and wealth has shown that any wealth increase prior to the Industrial Revolution was offset by an equivalent increase in population, thereby keeping income per capita constant. For this reason population densities are a viable proxy for wealth levels in 1500 CE; additionally, 1500 CE population densities are used regularly in similar research; e.g., Acemoglu et al. 2002, Ashraf and Galor 2008, Chanda and Putterman 2007, Putterman 2008. The hypothesis posed by this paper is that milking provided an extra resource to

[15]There is a hypothesis that states riboflavin rich milk allows for an increased risk to the contraction of malaria (Anderson and Vullo 1994), but this hypothesis is unproven (Meloni et al. 1998).

[16]The Hardy-Weinberg equilibrium states that allele frequencies in a population remain constant, that is, they are in equilibrium from generation to generation unless specific disturbing influences are introduced. Those disturbing influences include non-random mating, mutations, selection, limited population size, "overlapping generations", random genetic drift and gene flow (Hartl and Clark 2007).

[17]The world lactase persistence frequency calculated by Ingram et al. (2009), however, is based on a flawed population weighted average.

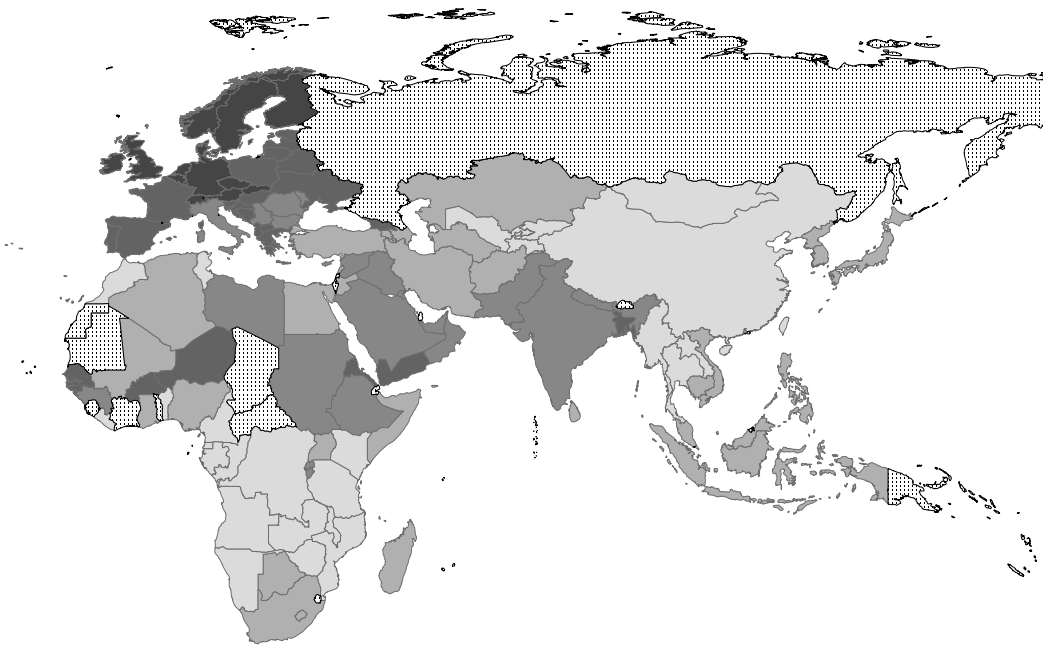

Figure 2.1: Distribution of Lactase Persistence

Note: Darker areas represent a greater frequency of lactase persistence. Dotted areas represent countries not in the data set. Western European countries are shown to have high levels of lactase persistence, while Sub-Saharan Africa and Southeast Asia have low levels of lactase persistence. This corresponds to the historical levels of milking from Simoons (shown in Fig. 2.2)

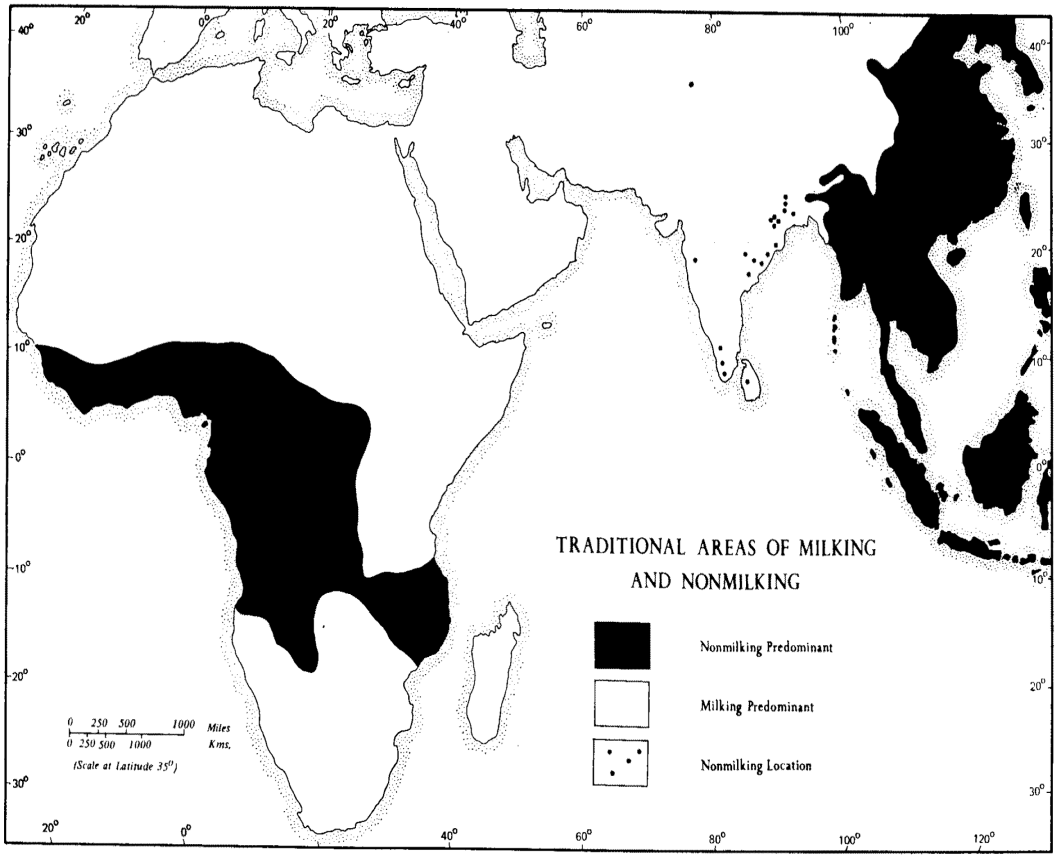

Figure 2.2: Historical Milk Consumption (Simoons 1969)

Note: Darker areas represent historically non-milking areas. There appears to be a high level of overlap of the historically non-milking areas and areas with low frequencies of lactase persistence shown in Fig. 2.1.

certain peoples that expanded the carrying capacity of their environment, thereby increasing population densities, or wealth. Figure 2.3 gives a simple plot with the natural log of population density on the y-axis and the country level frequency of lactase persistence on the x-axis.

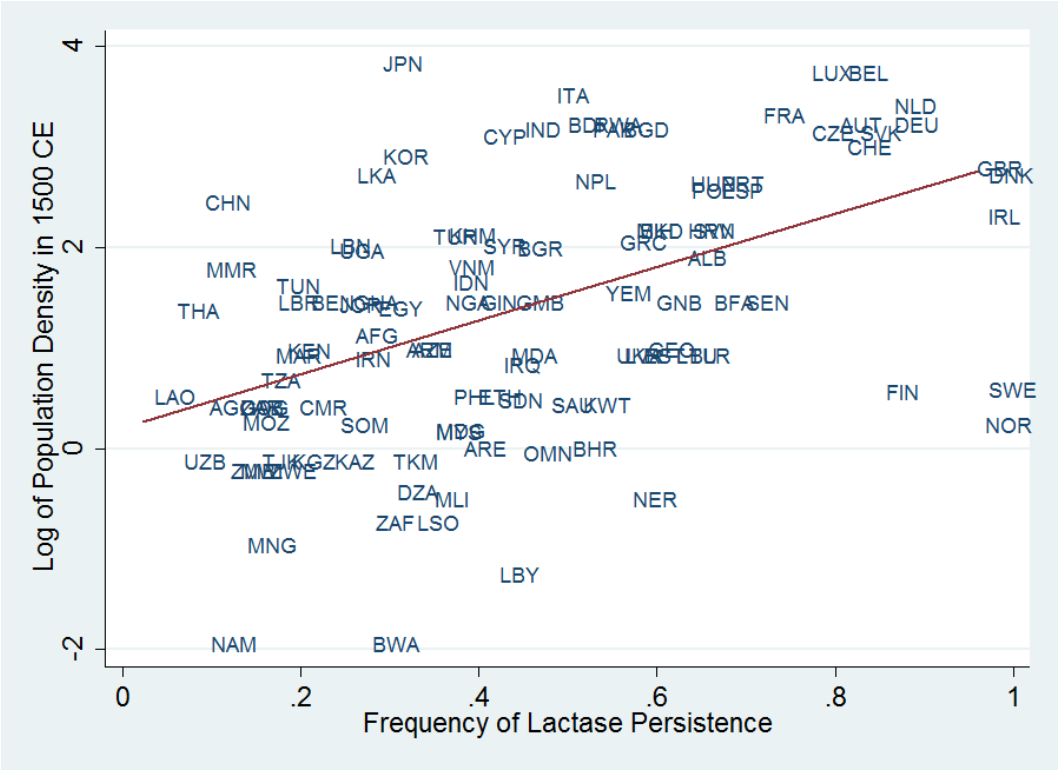

Figure 2.3: The Freq. of Lactase Persistence and the ln of Pop. Density in 1500 CE

As previously mentioned, the presence of mammals is a necessary, but not sufficient, condition for milking. This denotes that the frequency of lactase persistence may be picking up some of the effects of extended agricultural use. In order to show that milking itself increased population densities, agricultural transition dates need to be controlled for. As stated earlier, two different measures for agricultural transition dates have been used previously: the region specific measures from Hibbs and Olsson and the country specific measures from Putterman. Although Putterman's method is measured with greater certainty,

the measure by Hibbs and Olsson may have the effect of capturing unseen technological similarities between countries assigned to the same region. To conserve space we use the millennia of agriculture measure given by Putterman (2007). The results have been checked using millennia of agriculture from Hibbs and Olsson (2004) with little difference in estimation. Ideally, the distribution of livestock within the Old World would also be controlled for; however, such a measure is unavailable.

In addition to the initiation of agriculture, the yield from agriculture is also extremely important to food production and, therefore, variations in pre-colonial populations. Controlling for land quality is necessary to the estimation of pre-colonial populations. The land quality measure used in this paper is the mean suitability of agriculture (Ramankutty et al. 2002, Michalopoulos 2008). The mean suitability of agriculture is constructed by the country average of 0.5 degree latitude by longitude grids that give a probability of cultivation. Additionally, the soil suitability of potatoes, Old World staple crops, and New World staple crops from Nunn and Qian (2011) are used in the sensitivity analysis.

An additional genetic control comes from Spolaore and Wacziarg (2009) in which the authors measure the genetic distance, or variation, from the world's technological frontier. Using the genetic distance from the U.K. in the year 1500 CE gives a viable control for other alleles that may be highly correlated with lactase persistence. In other words, the frequency of lactase persistence may be accounting for a broad, underlying genetic capital possessed by Western Europeans; therefore, it is useful to see the effect of lactase persistence while controlling for other possible genetic variations.

When conducting sensitivity analyses for omitted variables, additional terrain, water access, environmental, cultural, and genetic controls are used.[18] The distance from the equator is intended to control for geographical variation that lactase persistence may be picking up; this variable is from Rodrik et al. (2002). Terrain and water access controls come from the Center of International Development. These include average elevation, av-

[18]Table 2.1 gives the source of all variables.

erage distance to the coast or navigable river, and the percent of land that is within 100 kilometers of the coast or navigable river. Terrain ruggedness and land within the tropics or deserts are from Nunn and Puga (forthcoming); to account for disease environments the stability of malaria transmission is used from Kiszewski et al. (2004); and whether or not a particular country belonged to the Roman Empire is also used from Acemoglu, Johnson, and Robinson (2005).

2.3 Results

The main hypothesis presented in this paper, a higher frequency of lactase persistence is associated with greater population densities in the pre-colonial era, is tested with the following estimating equation:

$$ln(\text{Population Density})_i^{1500} = \alpha + \beta(\text{Frequency of Lactase Persistence})_i + \mathbf{\Phi'X}_i + \epsilon_i \quad (2.1)$$

where i is a country index, β is the coefficient of interest throughout the paper, and $\mathbf{X}_i$ is a vector of country specific relevant controls. Equation (1) is estimated by OLS with robust standard errors. Robustness exercises use varied samples and variations in $\mathbf{X}_i$.

2.3.1 Baseline Estimation

The baseline estimations of Equation (1) are given in Table 2.2. Table 2.2 establishes the empirical relationship between the frequency of lactase persistence calculated by inversing the Putterman and Weil migration matrix within a country and the log of the 1500 population density for that particular country while controlling for relevant variables.

Column (1) displays the simple bivariate regression of 1500 population density on the frequency of lactase persistence within a particular country. The explanatory variable has a positive coefficient that is significant at the 1% level and explains roughly 20% of the

Table 2.2: Baseline Estimation

Dependent Variable: ln Population Density in 1500 CE								
		Extended Sample						Conservative Sample
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Freq. of Lactase Persistence (Inverse of Migration Matrix)	2.5992*** (0.4779)			2.2754*** (0.4471)	2.5573*** (0.5197)		2.4404*** (0.5102)	3.3765*** (0.7801)
Millennia of Agriculture		0.2062*** (0.0470)		0.1582*** (0.0425)		0.1608*** (0.0532)	0.1393** (0.0554)	0.0640 (0.0846)
Avg. Suitability of Agriculture			2.7205*** (0.3882)		2.5783*** (0.3677)	2.7151*** (0.3757)	2.5801*** (0.3608)	2.5032*** (0.5377)
Dist. from Equator			-0.0230** (0.0093)		-0.0364*** (0.0099)	-0.0213** (0.0086)	-0.0342*** (0.0091)	-0.0496*** (0.0125)
Sub-Saharan Africa			-1.2242*** (0.3162)		-1.3162*** (0.2991)	-0.5404 (0.3576)	-0.7197* (0.3884)	-1.8698*** (0.5973)
Western Europe			1.4545*** (0.2623)		0.5095* (0.2668)	1.5559*** (0.2722)	0.6405** (0.2715)	0.0286 (0.4133)
N	103	103	103	103	103	103	103	48
$R2$	0.2353	0.1437	0.4861	0.3162	0.5871	0.5241	0.6154	0.6999
$Adj.R^2$	0.2277	0.1352	0.4652	0.3025	0.5658	0.4996	0.5914	0.6560
F	29.5820	19.2234	32.3458	26.3302	37.4354	28.4010	38.4315	27.9297

Notes: OLS coefficients are reported in each column. *, **, and *** represent significance at the 10, 5, and 1% significance level, respectively. Robust standard errors are in parentheses. Lactase persistence frequency is the percent of a country's population that is able to digest milk. The extended sample includes the classification of ethnic groups based on language groups. The conservative sample represents direct matches in ethnic groups between Ingram et al. (2009) and Alesina et al. (2003). Country level lactase persistence frequencies are created by using the weighted average of a representative country's ethnic make-up in 1500 CE. Ethnic compositions in 1500 CE are found by the inverse of the Putterman and Weil migration matrix (2010). Millennia of agriculture come from Putterman (2007); separate estimations have been conducted using millennia of agriculture from Hibbs and Olsson (2004). Results are unchanged.

variance in the log of 1500 population density. To be more precise, column (1) reveals that a one standard deviation increase in the frequency of lactase persistence is associated with roughly a 63% increase in the number of people per kilometer. For the median country in the sample, the Sudan, this corresponds to a rough increase of two people per square kilometer. In column (2) a bivariate regression is run to show the impact of the millennia of agriculture within a country (Putterman 2008) on population densities; this is a direct test of the hypothesis proposed by Diamond. The coefficient is positive and significant at the 1% level with the explanatory variable accounting for roughly 14% of the variation in the dependent variable. Column (3) shows the effects of environmental variables, measured by the mean suitability of agriculture, distance from the equator, and dummies for Western Europe and Sub-Saharan Africa, on pre-colonial levels of development. The coefficient of our measure for the suitability of agriculture is positive and significant at the 1% level, which indicates improved land quality led to greater agricultural yields and larger populations. Column (3) also shows that a larger distance from the equator is associated with less dense populations in 1500 CE.[19] As expected, Western European countries had greater population densities, or wealth, relative to other countries, while Sub-Saharan African countries were relatively worse off.

Column (4) exhibits that when controlling for the millennia of agriculture, the coefficient of the frequency of lactase persistence remains significant at the 1% level, which further indicates that the frequency of lactase persistence is accounting for an additional advantage to a longer presence of agriculture. Column (5) shows the results of including the frequency of lactase persistence while controlling for environmental variables. The coefficient of the frequency of lactase persistence remains significant at the 1% level while also leading to a 10% increase in the explained variation of population densities in 1500. Col. (6) introduces millennia of agriculture while controlling for environmental variables; all signs are as expected, although the significance of the Sub-Saharan African dummy dissipates.

[19]The coefficient of the distance to the equator is influenced by the use of only Old World countries.

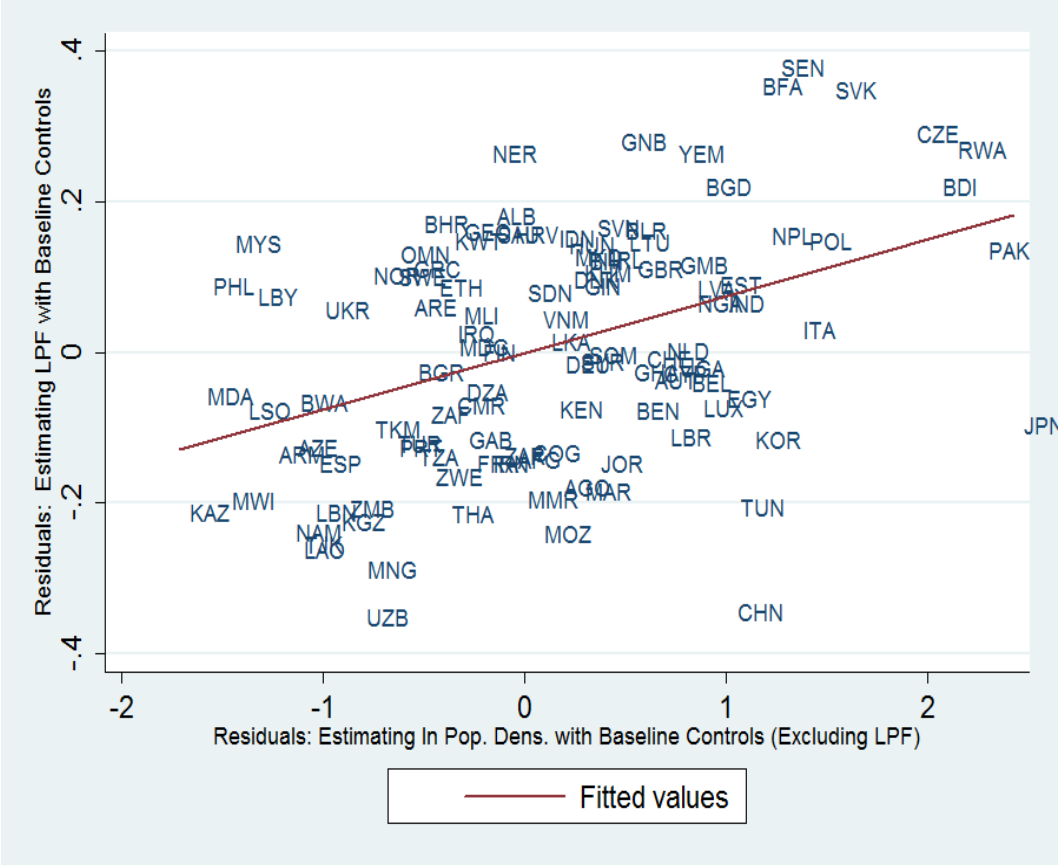

Figure 2.4: Orthogonal Plot of Estimated Effect of Lactase Persistence (Col. 7, Table 2.2)

The baseline result is given by columns (7) and (8). As stated previously, the availability of cattle is a necessary condition for the development of a gene that allows for digesting lactose; this implies the frequency of lactase persistence may be only capturing the effects of the millennia of agriculture within a particular country. Column (7) shows that when controlling for the country specific measures of millennia of agriculture, as well as environmental controls, the coefficient on lactase persistence remains both positive and significant at the 1% level. Also, comparing columns (6) and (7), the addition of the lactase persistence frequency increases the explained variation of population density in 1500 by roughly 8%. The coefficient of interest in column (7) is consistent with the bivariate estimation of column (1): an increase of one standard deviation in the ability of a population to digest lactose is associated with roughly a 60% increase in the population density in 1500. This suggests that the consumption of milk did indeed have a positive effect on the population density, or pre-colonial living standards, within a particular country.

Column (8) repeats the regression given by column (7); however, the sample is reduced to the countries in which the majority ethnic group is directly matched between Ingram et al. (2009) and Alesina et al. (2003). The coefficient of lactase persistence in Col. (8) is significantly larger than that in Col. (7); this is to be expected given the reduction in measurement error from using the more conservative sample. Also as expected the use of the smaller sample results in a larger standard error. In particular, a one standard deviation increase when using the coefficient in Col. (8) is associated with roughly a 87% increase in 1500 population density.

Table 2.3 performs the same estimations as Table 2.2, but instead uses the frequency of lactase persistence calculated by taking the majority ethnic groups within a country. The results, both magnitude and significance, are similar to those found in Table 2.2. For the baseline estimate of Col. (7), a one standard deviation increase in the frequency of lactase persistence corresponds to a 54% increase in 1500 population density; if we consider the conservative sample given in Col. (8), a one standard deviation increase in the frequency of

lactase persistence corresponds to an increase in population density of 82%. Given the high correlation and the similarity of coefficients between the two lactase persistence measures, hereafter we will use the measure calculated with the inverse of the migration matrix.

Tables 2.2 and 2.3 corroborate our main hypothesis. Those societies who consumed milk had the advantage of an additional resource; this additional resource, in turn, allowed for the development of greater pre-colonial populations. This relationship remains stable and significant while controlling for agricultural transition dates, agricultural suitability, and other relevant geographic determinants of pre-colonial wealth.

Whether or not the relationship between the frequency of lactase persistence and pre-colonial population density is causitive, depends upon the source of the cross-country differences in lactase persistence. In some sense, lactase persistence is analogous to the land suitability of potatoes found in Nunn and Qian (2011); in which, the frequency of lactase persistence can be seen as an exogenous suitability of consumption (rather than production) for a common good. Lactase persistence, however, has arisen in part due to cultural variation. The cultural cause of differences in lactase persistence creates an ambiguity in the exogeneity of our measure. In other words, did those cultures that adopted dairying have other unseen population advantages? The next section will attempt to alleviate the ambiguity in causation through sample adjustments, the inclusion of possible omitted variables, and instrumental variables estimation.

2.3.2 Sensitivity Analysis and Identification

The relationship between the frequency of lactase persistence and pre-colonial populations is established in Table 2.2; however, the nature of this relationship is unclear. The endogeneity of lactase persistence seems plausible: cultures which adopted dairying may have contained additional advantages that allowed for greater levels of pre-colonial development, geographic conditions that permitted dairying may have also permitted larger populations, etc. This suggests that OLS is unlikely to confirm a causative relationship between dairying and

Table 2.3: Baseline Estimation
Freq. of Lactase Persistence Calculated through Majority Ethnic Group

Dependent Variable: ln Population Density in 1500 CE								
		Extended Sample						Conservative Sample
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Freq. of Lactase Persistence (Majority Ethnic Group)	2.5465*** (0.4530)			2.2158*** (0.4216)	2.3470*** (0.4773)		2.2343*** (0.4678)	2.9711*** (0.6682)
Millennia of Agriculture		0.2062*** (0.0470)		0.1498*** (0.0419)		0.1608*** (0.0532)	0.1387** (0.0545)	0.0700 (0.0840)
Avg. Suitability of Agriculture			2.7205*** (0.3882)		2.5133*** (0.3659)	2.7151*** (0.3757)	2.5186*** (0.3620)	2.4697*** (0.5471)
Dist. from Equator			-0.0230** (0.0093)		-0.0359*** (0.0100)	-0.0213** (0.0086)	-0.0338*** (0.0091)	-0.0481*** (0.0125)
Sub-Saharan Africa			-1.2242*** (0.3162)		-1.2616*** (0.2981)	-0.5404 (0.3576)	-0.6700* (0.3865)	-1.7153*** (0.5861)
Western Europe			1.4545*** (0.2623)		0.5973** (0.2605)	1.5559*** (0.2722)	0.7260*** (0.2637)	0.2073 (0.3846)
N	103	103	103	103	103	103	103	48
R^2	0.2528	0.1437	0.4861	0.3243	0.5846	0.5241	0.6126	0.6882
$Adj.R^2$	0.2454	0.1352	0.4652	0.3108	0.5632	0.4996	0.5884	0.6426
F	31.6054	19.2234	32.3458	26.2841	38.3242	28.4010	41.4398	30.2402

Notes: OLS coefficients are reported in each column. *, **, and *** represent significance at the 10, 5, and 1% significance level, respectively. Robust standard errors are in parentheses. Lactase persistence frequency is the percent of a country's population that is able to digest milk. Country-level lactase persistence frequencies are created by using the weighted average of a representative countries ethnic make-up in 1500 CE. Ethnic compositions in 1500 CE are simply ascribed to the largest current ethnic group. The extended sample includes the classification of ethnic groups based on language groups. The conservative sample represents direct matches in ethnic groups between Ingram et al. (2009) and Alesina et al. (2003). Millennia of agriculture come from Putterman (2007); separate estimations have been conducted using millennia of agriculture from Hibbs and Olsson (2004). Results are unchanged.

population densities. This section attempts to strengthen perceptions of the relationship between the frequency of lactase persistence and 1500 CE population densities. Firstly, we perform truncations and include possible omitted variables to control for a potential spurious relationship. Secondly, we use the average solar radiation a country receives as an exogenous determinant of cross-country differences in the frequency of lactase persistence in order to determine causation. In all specifications the coefficient on the frequency of lactase persistence remains positive, significant, and, for the most part, is consistent in magnitude to the baseline estimate.

2.3.2.1 Sensitivity Analysis

Table 2.4 restricts the baseline estimation to each of the three continents that makeup the Old World.[20] The purpose of this is to show that Europe is not responsible for the significance of the coefficient in the baseline estimate, and that the positive relationship between a greater frequency of lactase persistence and pre-colonial population densities is seen within other continents. Column (1) performs the baseline estimation for countries contained only within Europe. The coefficient of lactase persistence in column (1) is positive, significant at the 1% level, and roughly double the magnitude of the baseline estimate given by column (7) of Table 2.2. This result implies the effects of milk consumption on population density are more pronounced within Europe; this is to be expected, since Europe has a greater history of milk consumption and, therefore, a greater exposure to the population advantages of milk (Simoons 1971). Column (2) constricts the sample to countries within Africa alone. The coefficient of interest is significant at the 10% level and the magnitude of the coefficient is lower than that given by the baseline estimate. The estimates of column (2), however, do show that milk consumption did have a positive effect on population density. The results are similar to those of column (3), which restricts the sample to only Asian countries. Within Asia, a greater frequency of lactase persistence is associated with

[20]The Western European and Sub-Saharan African dummies are excluded.

a greater population density; this effect is significant at the 10% level and differs slightly in magnitude from the baseline estimate. Column's (2) and (3) provide support that it is lactase persistence itself that led to larger populations and not an externality associated with Europe. This result is further confirmed in column (4), in which only Asian and African countries are considered. In column (4), the coefficient of the frequency of lactase persistence is once again significant at the 1% level and the magnitude only differs slightly from that given in the baseline estimate. Table 2.4 provides substantial evidence that the effect of lactase persistence is not being driven by a European externality, narrowing the possibility of a spurious correlation and providing a better understanding of the role of lactase persistence in explaining variations in pre-colonial population density.

Table 2.5 conducts column specified sample truncations. Columns (1) and (2) of Table 2.5 give the results of the baseline regression (Col. (7) of Table 2.2) while omitting Western European countries from the sample.[21] The purpose of the omission of Western European countries is in the fact that Western European countries have both the highest population densities and the highest levels of lactase persistence. Additionally, Columns (3) and (4) drop Sub-Saharan African countries from the sample. The reasoning for the omission of Sub-Saharan African countries is due to the fact that these countries contain on average lower frequencies of lactase persistence and lower population densities; the opposite of Western Europe. Column (5) omits both Western Europe and Sub-Saharan Africa, in effect dropping the highest and lowest frequencies of lactase persistence and the highest and lowest regional averages of population density in 1500 CE. In all cases the significance of the coefficient on the frequency of lactase persistence remains at the 1% level, and all point estimates are similar to the baseline case.

Columns (6) and (7) estimate the baseline regression while considering countries that are respectively above and below the median distance from the equator. The median ab-

[21]The baseline regression does include a Western European dummy, but the omission of Western European countries should further show that Western Europe is not the driving factor of the results given in the baseline case.

Table 2.4: Baseline Estimation: Within Continent Estimation

Dependent Variable: ln Population Density in 1500 CE				
	(1) Europe	(2) Africa	(3) Asia	(4) Asia + Africa
Freq. of Lactase Persistence (Inverse of Migration Matrix)	4.7439*** (1.2601)	1.5296* (0.8134)	2.3879** (1.1115)	1.9205*** (0.6017)
Millennia of Agriculture	0.0107 (0.2151)	0.4836*** (0.1132)	0.0929 (0.1115)	0.2457*** (0.0515)
Avg. Suitability of Agriculture	2.3344 (1.4321)	3.2697*** (0.7910)	2.5841*** (0.5218)	2.8111*** (0.4989)
Dist. from Equator	-0.0755** (0.0354)	-0.0359** (0.0156)	-0.0136 (0.0190)	-0.0298** (0.0117)
N	30	37	36	73
R^2	0.5732	0.6102	0.4129	0.5117
$Adj.R^2$	0.5049	0.5615	0.3372	0.4830
F	13.4362	18.5609	9.1871	20.0893

Notes: OLS coefficients are reported in each column. *, **, and *** represent significance at the 10, 5, and 1% significance level, respectively. Robust standard errors are in parentheses. The baseline estimation is given by column (7) in Table 2.2. The sample in column (1) includes only European countries; column (2) includes only Asian countries; column (3) includes only African countries; and column (4) excludes all European countries.

Table 2.5: Baseline Estimation: Sample Truncations

Dependent Variable: ln Population Density in 1500 CE							
Truncation:	Western Europe		Sub-Saharan Africa		WE and SSA	Dist. from Equator < Median	Dist. from Equator > Median
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Freq. of Lactase Persistence	2.3698***	2.4995***	2.5856***	1.8927***	2.0216***	2.1679***	2.9521**
(Inverse of Migration Matrix)	(0.5329)	(0.5273)	(0.5518)	(0.6937)	(0.7323)	(0.5943)	(1.2157)
Millennia of Agriculture	0.2085***	0.1324**	0.1038*	0.1138**	0.1003*	0.3795***	-0.0616
	(0.0403)	(0.0574)	(0.0566)	(0.0558)	(0.0574)	(0.0785)	(0.1337)
Avg. Suitability of Agriculture	2.4355***	2.4044***	2.4070***	2.4921***	2.2060***	3.4520***	2.4344***
	(0.4066)	(0.3919)	(0.4186)	(0.4150)	(0.4585)	(0.5153)	(0.6171)
Dist. from Equator	-0.0236***	-0.0331***	-0.0196**	-0.0209**	-0.0187*	-0.0494***	-0.0810**
	(0.0070)	(0.0091)	(0.0097)	(0.0096)	(0.0096)	(0.0148)	(0.0386)
Sub-Saharan Africa		-0.7242*				0.1990	
		(0.3952)				(0.4367)	
Western Europe				0.6453**			0.5384
				(0.3018)			(0.4288)
N	89	89	71	71	57	52	46
R^2	0.5221	0.5448	0.5073	0.5273	0.4273	0.6920	0.6105
$Adj.R^2$	0.4993	0.5174	0.4775	0.4910	0.3833	0.6585	0.5619
F	27.0615	25.5653	21.3149	21.6238	12.9582	29.6344	16.7108

Notes: OLS coefficients are reported in each column. *, **, and *** represent significance at the 10, 5, and 1% significance level, respectively. Robust standard errors are in parentheses. Columns (1) and (2) drop Western European countries from the sample. Col.'s (3) and (4) drop Sub-Saharan African countries from the sample. Col. (5) drops both Sub-Saharan African and Western European countries from the sample. Col.'s (6) and (7) respectively drop data points below and above the median sample distance from the equator. The median distance from the equator corresponds to a latitude of 33 degrees, which corresponds to an area slightly above North African states, roughly equal to the Levant, and above India and Southeast Asia.

solute latitude of our sample is 33 degrees. This corresponds to an area just above the tropics or roughly equal to the Levant and slightly above North African states, India, and Southeast Asia. The truncation is done to control for any biases that may occur due to the relationships between milk consumption, vitamin D, and the availability of sun light.[22] After the respective truncations, the point estimates of the coefficient on lactase persistence remains significant at the 1% level and is similar in magnitude to the base line estimation. From the truncations, a selection bias seems improbable.

Our method for approximating ethnic compositions in 1500 CE is prone to measurement error. This is due to disparities in the current ethnic composition and country compositions in the migration matrix (Putterman and Weil 2010). Further, this error is larger in countries that have experienced large immigrations between 1500 and 2000 CE. To account for this potential error Table 2.6 truncates the base sample by the fraction of the current population that is derived from the 1500 CE population. Column (1), for example, excludes all countries which have less than 50% of the current population originating from the within country 1500 CE population. This results in the exclusion of only two countries from our baseline sample; as a result, the significance and magnitude of the coefficient of interest are analogous to those in column (7) of Table 2.2. Column (2) excludes countries in which less than 75% of the contemporary population is derived from the 1500 CE population. This results in the exclusion of 10 countries that are included in the baseline sample. The coefficient of the frequency of lactase persistence remains consistent in magnitude and significance. Column (3) performs the same truncation as columns (1) and (2) but sets the threshold of within country population to 85%; again, the estimates are similar to the baseline case. Column (4) excludes countries in which 95% of the current population is derived from 1500 CE populations. This results in excluding 50 countries from the baseline sample. The estimate of the coefficient of interest, however, remains roughly equivalent to the baseline estimate. As a further check, column (5) replaces the frequency of lactase persistence derived by

[22]This idea is further explored with the inclusion of a solar radiation variable into our baseline estimation.

post-multiplying by the inverse of the migration matrix with the measure calculated by assuming the majority ethnic group. Again, the coefficient of lactase persistence is positive, significant at the 1% level, and similar in magnitude to estimations with the full sample. The measurement error that results in our approximation of 1500 CE ethnic compositions does not appear to affect our results. This gives further credence to the relationship between milk consumption, measured by the frequency of lactase persistence, and population density posed in this paper.

Tables 2.7, 2.8, 2.9, 2.10, 2.11, and 2.12 explore whether additional controls can make the effects of lactase persistence frequencies disappear. Table 2.7 includes an additional genetic measure. Table 2.8 replaces the mean suitability of agriculture in the baseline estimation (Michalopolous 2008; Ramankutty 2002) with soil suitability measures from Nunn and Qian (2011); these include the suitability for potatoes, New World staples, and Old World staples. Table 2.9 includes additional environmental controls: elevation, ruggedness, whether a country is within the tropics or desert, a measure of malarial intensity, and whether or not a country belonged to the Roman Empire . Water access variables are included in Table 2.10. Table 2.11 includes biogeographic variables from Hibbs and Olsson (2004), while Table 2.12 includes all additional variables specified in the previous tables.

As noted earlier lactase persistence is a function of the genotype of a respective individual. It may be the case that a genotype that allows for lactase persistence may also allow for other growth promoting attributes, or, in other words, there may be some underlying genetic capital which is beneficial to development. Table 2.7 introduces the genetic distance from the technological frontier, Britain, in the year 1500 CE to the baseline model (Spolaore and Wacziarg 2009). Spolaore and Wacziarg argue that a smaller genetic distance (i.e. similar genotypes) allowed for an easier diffusion of technology. This is seen in the bivariate regression of Col. (2) in Table 2.7, where a greater genetic distance from Britain in 1500 CE is associated with lower population densities. The significance of genetic distance remains while controlling for the frequency of lactase persistence (Col. (3)); however, the

Table 2.6: Baseline Estimation: Truncations Due to Migration

Dependent Variable: ln Population Density in 1500 CE					
% of Population Derived from 1500 CE Population:	(1) >50%	(2) >75%	(3) >85%	(4) >95%	(5)† >95%
Freq. of Lactase Persistence	2.4181***	2.4414***	2.3937***	2.1290***	1.8251***
	(0.5222)	(0.5392)	(0.5461)	(0.7343)	(0.6640)
Millennia of Agriculture	0.1251**	0.1403**	0.1337**	0.1097	0.1117
	(0.0584)	(0.0631)	(0.0665)	(0.1234)	(0.1210)
Avg. Suit of Agr.	2.4751***	2.6191***	2.6878***	2.7767***	2.7257***
	(0.3836)	(0.4158)	(0.4220)	(0.5598)	(0.5745)
Dist. from Equator	-0.0322***	-0.0375***	-0.0332***	-0.0292*	-0.0293*
	(0.0092)	(0.0090)	(0.0096)	(0.0158)	(0.0162)
Sub-Saharan Africa	-0.6740*	-0.7891**	-0.7420*	-0.7658	-0.7203
	(0.3919)	(0.3794)	(0.3834)	(0.7360)	(0.7281)
Western Europe	0.6036**	0.5821*	0.4794	0.3976	0.5457
	(0.2695)	(0.2953)	(0.3035)	(0.4478)	(0.4581)
N	99	86	82	50	50
R^2	0.5777	0.6118	0.6049	0.5575	0.5453
$Adj.R^2$	0.5502	0.5823	0.5732	0.4958	0.4819
F	34.4443	31.9720	28.9073	13.6354	14.4287

Notes: OLS coefficients are reported in each column. *, **, and *** represent significance at the 10, 5, and 1% significance level, respectively. Robust standard errors are in parentheses. The baseline estimation is given by column (7) in Table 2.2. The sample truncations are based on the percent of a countries 2000 CE population that is derived from the 1500 CE population (Putterman and Weil 2010).

† The frequency of lactase persistence in column (5) is calculated by the majority ethnic group.

inclusion of relevant agricultural and geographic controls makes the coefficient of Spolaore and Wacziarg's genetic distance statistically insignificant (Col. (4)). The additional genetic control does not alter the significance or magnitude of lactase persistence. Lactase persistence is of importance, not because it is part of some larger genetic package, but because lactase persistence allowed for the consumption of an additional resource. This singular genetic adaptation gave an advantage, which in turn, allowed for the development of larger historic populations.

Column (1) of Table 2.8 includes the average country-level soil suitability for potatoes while excluding the baseline soil suitability measure. The introduction of the potato in between the 18th and 19th centuries is associated with a large increase in population over this time period (Nunn and Qian 2011). The inclusion of this suitability measure is intended to capture any additional effects that this measure may be accounting for in regards to population variation. As seen in Col. (1) the potato suitability measure is positive and significant, indicating an additional relationship between the soil suitability and population density. The inclusion of this variable, however, does not affect the significance or magnitude of the coefficient of the frequency of lactase persistence. Columns (2), (3), and (4) respectively introduce the suitability for Old World staple crops, New World staple crops, and jointly controls for both measures of soil suitability. Again, the significance and magnitude of the coefficient of lactase persistence remain similar to the baseline estimation. Col. (5) replaces the measure for New World staple crops with that for potatoes; the role of lactase persistence is unaffected, while the suitability of both potatoes and Old World staples are positive and significantly related to pre-colonial population densities. Table 2.8 again confirms that dairying did have a strong association with historic population densities. This relationship is not the by product of soil suitability; rather, dairying was an important determinant to pre-colonial populations.

Columns (1), (2), and (3) of Table 2.9 introduce elevation (in km), ruggedness, and ruggedness squared into the estimation. Ruggedness is roughly the variation in elevation of

Table 2.7: Additional Genetic Control

Dependent Variable: ln Population Density in 1500 CE					
	(1)	(2)	(3)	(4)	(5)
Freq. of Lactase Persistence	2.5992***		1.9239***		2.4473***
(Inverse of Migration Matrix)	(0.4779)		(0.5606)		(0.5484)
Genetic Dist. from U.K. in 1500 CE		-0.7464***	-0.4187**	-0.3749	0.0093
(Spolaore and Wacziarg 2009)		(0.1632)	(0.1885)	(0.2752)	(0.2823)
Millennia of Agriculture				0.1173**	0.1403**
				(0.0576)	(0.0614)
Avg. Suitability of Agriculture				2.7483***	2.5789***
				(0.3778)	(0.3620)
Dist. from Equator				-0.0223***	-0.0343***
				(0.0082)	(0.0093)
Sub-Saharan Africa				-0.2702	-0.7269
				(0.4221)	(0.4804)
Western Europe				1.3711***	0.6425**
				(0.2993)	(0.2832)
N	103	103	103	103	103
R^2	0.2353	0.1837	0.2772	0.5344	0.6154
$Adj.R^2$	0.2277	0.1756	0.2627	0.5053	0.5871
F	29.5820	20.9094	19.2919	24.9998	32.6228

Notes: OLS coefficients are reported in each column. *, **, and *** represent significance at the 10, 5, and 1% significance level, respectively. Robust standard errors are in parentheses. Genetic distance from the United Kingdom is measured by differences in neutral genetic distances across peoples. The inclusion of this variable is intended to capture genetic capital differences that may affect population density.

Table 2.8: Additional Soil Suitability Measures

Dependent Variable: ln Population Density in 1500 CE					
	(1)	(2)	(3)	(4)	(5)
Freq. of Lactase Persistence	2.9272***	3.1352***	2.9334***	3.1416***	3.1238***
	(0.5577)	(0.5396)	(0.5579)	(0.5467)	(0.5376)
Millennia of Agriculture	0.1212*	0.1181*	0.1210*	0.1189*	0.1170*
	(0.0635)	(0.0616)	(0.0632)	(0.0619)	(0.0625)
Dist. from Equator	-0.0396***	-0.0353***	-0.0331***	-0.0354***	-0.0364***
	(0.0107)	(0.0098)	(0.0100)	(0.0098)	(0.0107)
Sub-Saharan Africa	-0.9105*	-0.9890**	-0.9304**	-0.9846**	-0.9916**
	(0.4636)	(0.4472)	(0.4534)	(0.4488)	(0.4521)
Western Europe	0.3808	0.2480	0.3737	0.2400	0.2617
	(0.3459)	(0.3534)	(0.3444)	(0.3676)	(0.3495)
Suitability of Potatoes	0.0564***				0.0116
	(0.0159)				(0.0221)
Suitability of Old World Crops		0.0368***		0.0396**	0.0324**
		(0.0093)		(0.0160)	(0.0124)
Suitability of New World Crops			0.0162***	-0.0022	
			(0.0050)	(0.0067)	
N	103	103	103	103	103
R^2	0.4191	0.4396	0.4136	0.4398	0.4403
$Adj.R^2$	0.3827	0.4046	0.3769	0.3985	0.3991
F	15.7640	17.4407	14.1398	14.8472	14.9377

Notes: OLS coefficients are reported in each column. *, **, and *** represent significance at the 10, 5, and 1% significance level, respectively. Robust standard errors are in parentheses. The baseline soil suitability measure (Michalopoulos 2009) is not included. Inclusion of the baseline soil suitability measure in the above table does not change the magnitude or significance of the coefficient on the frequency of lactase persistence.

particular cells within a country, which are then averaged to the country level (Nunn and Puga forthcoming). For our concerns, ruggedness and elevation may account for land variations that make farming difficult; and, therefore, may promote the use of animal husbandry, which increases the likelihood of milk consumption. The addition of these additional geographic controls should alleviate any potential biases that may occur due to land conditions that lead to an increased use of pastoralism. Col. (1) includes elevation into the estimation; results remain significant and similar to the baseline estimates. The inclusion of ruggedness and its square in column (2) produce trivial differences in the estimates of the coefficient of interest.

An argument has been put forward that extreme environments may contribute to variations in lactase persistence (Cook and al-Torki 1975). The idea being that extreme environments have fewer resources in which to support populations; therefore, the ability to drink milk becomes essential to surviving and will rise to a greater frequency within the population. Columns (3) and (4) of Table 2.9 control for environmental differences by including, respectively, the percent of land within the tropics and the percent of land which is desert. The percent of land within the tropics, for our purposes, represents an environment in which resources are rich; consequently, there should be little need for dairying. At the other extreme, deserts are poor in resources, implying a greater need for dairying. This is verified by the coefficients on the respective environments. Deserts have a negative and significant effect on pre-colonial population density, while the tropics have a positive but insignificant effect. Neither variable alters the effect of the frequency of lactase persistence. The coefficient of lactase persistence remains positive, significant, and similar in magnitude to the baseline estimate; this is true while including the environmental variables separately (Col.'s (3) and (4)) or jointly (Col. (5)).

An additional environmental effect that may act on the number of cattle (and, in turn, the number of milk drinkers) and population density is the disease environment. Cattle and other milk producers are extremely sensitive to the tsetse fly, while people are subject to

Table 2.9: Additional Environment, Disease, and Cultural Controls

Dependent Variable: ln Population Density in 1500 CE								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Freq. of Lactase Persistence	2.4447***	2.4831***	2.3936***	2.4214***	2.3979***	2.3836***	2.5196***	2.5083***
	(0.5842)	(0.5252)	(0.5019)	(0.5025)	(0.4963)	(0.5658)	(0.5176)	(0.6339)
Millennia of Agriculture	0.1391**	0.1271**	0.1583**	0.1511***	0.1604**	0.1384**	0.1304**	0.1407**
	(0.0555)	(0.0573)	(0.0645)	(0.0566)	(0.0643)	(0.0554)	(0.0554)	(0.0625)
Avg. Suitability of Agriculture	2.5802***	2.5260***	2.4396***	2.1578***	2.1082***	2.5980***	2.4389***	1.8301***
	(0.3627)	(0.3710)	(0.4117)	(0.4100)	(0.4441)	(0.3621)	(0.3695)	(0.4678)
Dist. from Equator	-0.0343***	-0.0320***	-0.0264**	-0.0347***	-0.0306***	-0.0331***	-0.0333***	-0.0218*
	(0.0094)	(0.0092)	(0.0112)	(0.0093)	(0.0113)	(0.0100)	(0.0091)	(0.0117)
Sub-Saharan Africa	-0.7211*	-0.6432*	-0.6156	-0.7655*	-0.7081*	-0.7965*	-0.7231*	-0.6745
	(0.3851)	(0.3867)	(0.4203)	(0.3944)	(0.4228)	(0.4079)	(0.3894)	(0.4326)
Western Europe	0.6398**	0.6124**	0.5993**	0.5474**	0.5311**	0.6548**	0.2722	0.0074
	(0.2749)	(0.2855)	(0.2637)	(0.2645)	(0.2612)	(0.2798)	(0.3508)	(0.3752)
Mean Elevation	0.0034							0.0357
	(0.1564)							(0.3073)
Mean Ruggedness		0.2698						0.2991
		(0.2288)						(0.2640)
Mean Ruggedness^2		-0.0568						-0.0627
		(0.0427)						(0.0416)
% in Tropics			0.0037		0.0019			0.0038
			(0.0037)		(0.0037)			(0.0046)
% in Deserts				-0.0160***	-0.0151**			-0.0143**
				(0.0059)	(0.0059)			(0.0066)
Malaria Ecology Index						0.0092		0.0105
						(0.0174)		(0.0157)
Member of the Roman Empire							0.6365**	0.8368**
							(0.3095)	(0.3245)
N	103	103	103	103	103	103	103	103
R^2	0.6154	0.6231	0.6189	0.6313	0.6322	0.6165	0.6248	0.6554
$Adj.R^2$	0.5871	0.5911	0.5908	0.6042	0.6010	0.5882	0.5971	0.6051
F	32.5697	26.8209	36.5071	34.6062	31.8748	33.2684	33.7513	19.5925

Standard errors in parentheses
* $p < .1$, ** $p < .05$, *** $p < .01$

Notes: OLS coefficients are reported in each column. *, **, and *** represent significance at the 10, 5, and 1% significance level, respectively. Robust standard errors are in parentheses. Tropics represent both a lush environment and a potential control for diseases, which may have affected cattle distributions. Deserts represent an extreme environment, in which reliance on milk may be necessary for survival. Roman occupation may have instilled both favorable institutions and an intogression of lactase persistent alleles.

malaria and other tropical disease from similar environments. Looking at Figures 2.1 and 2.2, areas with a low frequency of lactase persistence are similar to areas with historic levels of malaria. This indicates that the relationship between lactase persistence and historic populations may be driven by the disease environment. Column (6) controls for the disease environment by including the stability of malarial transmission within a particular country, which can also be seen as a proxy for the tsetse fly (Kiszewski et al. 2004). While this is a contemporary measure, we have little evidence to believe it is an ineffective control variable. The inclusion of the disease proxy does not affect the coefficient of lactase persistence. The estimated coefficient of the frequency of lactase persistence is unaffected by the inclusion of the malaria ecology index. Particularly, the coefficient remains significant at the 1% level and is of a consistent magnitude to the baseline estimate.

Column (7) of Table 2.9 includes a dummy for whether or not a country was part of the Roman Empire. Using historical evidence Acemoglu, Johnson, and Robinson (2005) argue that being included in the Roman Empire may have contributed to the advanced growth of Western Europe. This is a cultural variable that may be included with the diffusion of technology, development levels in the pre-colonial era, and, ultimately, the practice of dairying. Col. (7) shows that being a part of the Roman Empire did have a significant effect on population densities in 1500; however, this effect is not coming at the expense of lactase persistence. The inclusion of the Roman Empire dummy causes no meaningful difference in the magnitude or statistical significance in the coefficient of lactase persistence.

Column (8) in Table 2.9 introduces all environmental, disease, and cultural controls. Again, the significance and magnitude of the coefficient on lactase persistence are unaltered. The relationship between dairying and historic populations is not the result of a simultaneous correlation with an environmental or cultural variable.

Table 2.10 includes a number of water access controls. These include the distance from an ice free coast, the distance from a navigable river, the distance to either an ice free coast or a navigable river, the percent of land within 100 kilometers of an ice free coast,

and the percent of land within 100 kilometers of an ice free coast or a navigable river. Neither individually nor jointly introducing water access controls affects the significance or magnitude of the coefficient on lactase persistence. Specifically, column (6) gives the baseline estimation while including both the distance from a coast or a river and the percent of land within 100 kilometers of a coast or river; the coefficient of lactase persistence is significant at the 1% level and resembles the baseline estimate.

Domesticable animals were a necessary condition for the development of lactase persistence. But domesticable animals also provide population benefits, e.g., meat, labor, etc. Table 2.11 uses the number of potential domesticate animals as a proxy for the additional benefits conferred by domesticate animals, as well as other biogeographic controls from Hibbs and Olsson (2004). Column (1) gives the baseline estimates with the sample reduction; results are similar to the larger sample in column (7) of Table 2.2. Column (2) includes the number of domesticable animals into the baseline estimation. The inclusion of this variable has a negligent effect on the coefficient of lactase persistence. This supports our main hypothesis that a greater level of milk consumption led to denser populations in the pre-colonial era. Columns (3) and (4) include a measure for the number of domesticable crops and a measure for the East-West orientation of a country respectively. The coefficient of interest remains roughly equivalent to the baseline estimate. Column (5) includes both the number of domesticable plants and animals, while column (6) includes all variables from Hibbs and Olsson into the baseline estimation. The inclusion of biogeographic controls does not influence the estimated relationship between the frequency of lactase persistence and population density in 1500 CE.

Table 2.12 simultaneously introduces the potential omitted variables discussed in Tables 2.7, 2.9, 2.10, and 2.11. [23] The inclusion of all additional variables does not affect the coefficient of lactase persistence; this is shown in column (5). Column (6) reproduces the

[23]The only soil suitability measure considered in Table 2.12 is the baseline measure from Michalopolous (2011). Inclusion of differing suitability measures has an insubstantial effect on the coefficient of the frequency of lactase persistence.

Table 2.10: Inclusion of Water Access Controls

Dependent Variable: ln Population Density in 1500 CE						
	(1)	(2)	(3)	(4)	(5)	(6)
Freq. of Lactase Persistence	2.3735***	2.2633***	2.2143***	2.4442***	2.2844***	2.2138***
	(0.5848)	(0.5451)	(0.6295)	(0.5479)	(0.6538)	(0.6869)
Millennia of Agriculture	0.1410**	0.1506**	0.1475***	0.1390***	0.1462***	0.1476***
	(0.0546)	(0.0594)	(0.0545)	(0.0526)	(0.0543)	(0.0537)
Mean Suitability of Agriculture	2.5410***	2.3393***	2.4605***	2.5828***	2.4682***	2.4601***
	(0.3806)	(0.4486)	(0.3848)	(0.3617)	(0.3961)	(0.4039)
Dist. from Equator	-0.0330***	-0.0365***	-0.0318***	-0.0343***	-0.0335***	-0.0318***
	(0.0103)	(0.0090)	(0.0101)	(0.0093)	(0.0094)	(0.0102)
Sub-Saharan Africa	-0.6874*	-0.7338*	-0.6246	-0.7248**	-0.6404*	-0.6243*
	(0.3823)	(0.3850)	(0.3843)	(0.3521)	(0.3678)	(0.3651)
Western Europe	0.6253**	0.6311**	0.6441**	0.6419**	0.6590**	0.6442**
	(0.2767)	(0.2670)	(0.2725)	(0.2764)	(0.2753)	(0.2769)
Mean Dist. to Coast	-0.0876					
	(0.2349)					
Mean Dist. to River		-0.1943				
		(0.1626)				
Mean Dist. to Coast or River			-0.2389			-0.2384
			(0.2221)			(0.2360)
% within 100 Km of Coast				-0.0114		
				(0.3607)		
% within 100 Km of Coast or River					0.1934	0.0012
					(0.3586)	(0.3934)
N	103	103	103	103	103	103
R^2	0.6161	0.6234	0.6202	0.6154	0.6169	0.6202
$Adj.R^2$	0.5878	0.5956	0.5922	0.5871	0.5886	0.5879
F	33.2601	34.2635	35.2240	32.6504	33.1905	31.2543

Notes: OLS coefficients are reported in each column. *, **, and *** represent significance at the 10, 5, and 1% significance level, respectively. Robust standard errors are in parentheses. Additional water access controls are included into the baseline estimation. Easier access to water is intended to represent an ease to trade, which may affect 1500 CE population densities.

Table 2.11: Additional Biogeographic Controls

Dependent Variable: ln Population Density in 1500 CE						
	(1)	(2)	(3)	(4)	(5)	(6)
Freq. of Lactase Persistence	2.3929***	2.4414***	2.3673***	2.4257***	2.5728***	2.6226***
	(0.6166)	(0.6529)	(0.5985)	(0.6064)	(0.6284)	(0.6002)
Millennia of Agriculture	0.2327***	0.1792**	0.2254***	0.2557***	0.1896**	0.2178**
	(0.0747)	(0.0896)	(0.0764)	(0.0690)	(0.0858)	(0.0917)
Avg. Suitability of Agriculture	2.3611***	2.3424***	2.3578***	2.4031***	2.3492***	2.3923***
	(0.5056)	(0.4909)	(0.5075)	(0.5068)	(0.4997)	(0.5073)
Dist. from Equator	-0.0279**	-0.0405**	-0.0297*	-0.0266**	-0.0376*	-0.0350*
	(0.0128)	(0.0183)	(0.0161)	(0.0127)	(0.0190)	(0.0197)
Sub-Saharan Africa	-0.6883	0.4149	-0.6793	-0.8716	0.8204	0.6231
	(0.5162)	(0.8054)	(0.5279)	(0.6088)	(0.8371)	(1.1978)
Western Europe	0.3088	0.2363	0.2978	0.3003	0.2556	0.2563
	(0.3168)	(0.3312)	(0.3251)	(0.3257)	(0.3298)	(0.3335)
Number of Potential Domesticate Animals		0.2008			0.2818*	0.2793
		(0.1363)			(0.1677)	(0.2031)
Number of Potential Domesticate Plants			0.0044		-0.0190	-0.0222
			(0.0196)		(0.0238)	(0.0250)
East-West Orientation				-0.0003		-0.0003
				(0.0003)		(0.0003)
N	70	70	70	70	70	70
R^2	0.6718	0.6878	0.6722	0.6790	0.6926	0.6990
$Adj.R^2$	0.6406	0.6526	0.6352	0.6427	0.6523	0.6539
F	31.3909	28.7499	27.2766	27.6883	26.4689	25.0621

Notes: OLS coefficients are reported in each column. *, **, and *** represent significance at the 10, 5, and 1% significance level, respectively. Robust standard errors are in parentheses. Additional biogeographic controls—Number of Domesticable Animals, Number of Domesticable Plants, and East-West Orientation—are from Hibbs and Olsson (2004).

estimate of column (4), while only considering our conservative sample.[24] Again, neither the magnitude or significance of the coefficient are meaningfully affected. The effect of lactase persistence is robust to the inclusion of a large and theoretically important set of additional controls. Omitted variable bias seems to be insubstantial.

In summary, the coefficient on lactase persistence remains relatively constant throughout the numerous empirical specifications performed. Throughout the sensitivity analysis, the coefficient of the frequency of lactase persistence remains significant at the 1% level and is rarely different in magnitude from the bivariate or baseline estimations (Columns (1) and (7) of Table 2.2). This robustness is shown through differing samples and the inclusion of theoretically relevant variables, which should, in the least, mitigate a potential selection or simultaneity bias. A strong association exists between milk consumption and population densities in 1500 CE. This implies that the intensity of milk consumption did play some role in the development of larger pre-colonial societies. Those who were able, and did, consume milk gained both qualitative and quantitative advantages which led to larger populations; larger populations in turn led to greater armies, technological gains, and eventually a head start to prosperity differences seen today.

This works primary goal is to explore the role milk consumption, measured through the ability to digest lactose, had in the accumulation of pre-colonial populations. The coevolution of the ability to consume milk with the cultural adaptation of dairying, however, prevents the genetically given lactase persistence measure to be truly exogenous. The omitted reason as to why some cultures initiated dairying while others did not may also be correlated with the accumulation of pre-colonial populations, implying a potential simultaneity bias. Without the use of an exogenous instrument, causality cannot be established. The next section will attempt to alleviate the lack of causation with the use of an exogenous instrument.

[24] Biogeographic controls are omitted due to sample considerations.

Table 2.12: All Controls

Dependent Variable: ln Population Density in 1500 CE						
		Extended Sample				Conservative Sample
	(1)	(2)	(3)	(4)	(5)	(6)
Freq. of Lactase Persistence	2.8383***	2.2420***	2.6170***	2.7210***	3.0125***	3.9665***
	(0.6294)	(0.6895)	(0.7050)	(0.7036)	(0.7041)	(1.0285)
Controls:						
Baseline	Y	Y	Y	Y	Y	Y
Genetic	Y	Y	N	Y	Y	Y
Environment	Y	N	Y	Y	Y	Y
Water Access	N	Y	Y	Y	Y	Y
Biogeographic	N	N	N	N	Y	N
N	103	103	103	103	68	48
R^2	0.6765	0.6204	0.6789	0.6810	0.7387	0.7806
$Adj.R^2$	0.6250	0.5836	0.6236	0.6217	0.6353	0.6674
F	16.2606	27.9260	15.5175	14.9061	10.1204	11.9294

Notes: OLS coefficients are reported in each column. *, **, and *** represent significance at the 10, 5, and 1% significance level, respectively. Robust standard errors are in parentheses. Baseline controls include the millennia of agriculture, the average suitability of agriculture, the distance from the equator, and dummies for Sub-Saharan Africa and Western Europe. The additional genetic control is the genetic distance to the United Kingdom. Environmental controls include average elevation, average ruggedness and its square, the percent of a country within a desert or the tropics, a measure for the suitability of malaria, and whether or not a country belonged to the Roman Empire. Water access controls include the distance from the coast or a navigable river and the percent of a country within 100 kilometers from a coast or river. Biogeographic controls include the number of potential domesticate plants and animals and a measure for the country's East-West orientation. Inclusion of differing soil suitability measures does not change the magnitude or significance of the coefficient on the frequency of lactase persistence.

2.3.2.2 Identification

In order to establish causation we consider the proposed relationship between lactase persistence and low sunlight areas (Flatz and Rotthauwe 1973). In adequate sunlight, the body is able to synthesize vitamin D; however, if sunlight is low, individuals may be deficient in vitamin D. A major disease associated with deficieny in vitamin D is rickets, which results in the softening of bones. A diet heavy in milk would increase calcium absorption, thereby partially offsetting the harmful effects of Vitamin D deficiency (Flatz and Rotthauwe 1973; Gueguen and Pointillart 2000).[25] Therefore, those societies in low sunlight countries, i.e. Western Europe, gained an additional benefit from the consumption of milk. With this understanding, we use a 22 year average of solar radiation as an exogenous determinant of the frequency of lactase persistence.

The measure of solar radiation comes from the Atmospheric Science Data Center of NASA (NASA Surface Meteorology and Solar Energy 2011). With the use of country latitude and longitude from the CIA World Factbook, we calculate the 22 year average of solar radiation of a horizontal surface, given in the kilowatts per hour of a squared meter, for all countries in our sample. Figure 2.5 plots the relationship between this measure of solar radiation and our measure of the frequency of lactase persistence. The relationship appears to be nonlinear. At low levels of sunlight, lactase persistence is widespread; however, as sunlight increases beyond an adequate amount, the frequency of lactase persistence becomes more varied. We therefore use solar radiation and its square in order to instrument the frequency of lactase persistence.

While the relationship between solar radiation and the frequency of lactase persistence is strong in our sample, the use of solar radiation as an instrument is problematic. First, solar radiation may correlate with factors that influence population density. This is partially alleviated by the inclusion of relevant controls, i.e. the mean suitability of agriculture, dis-

[25] Milk also contains small amounts of vitamin D.

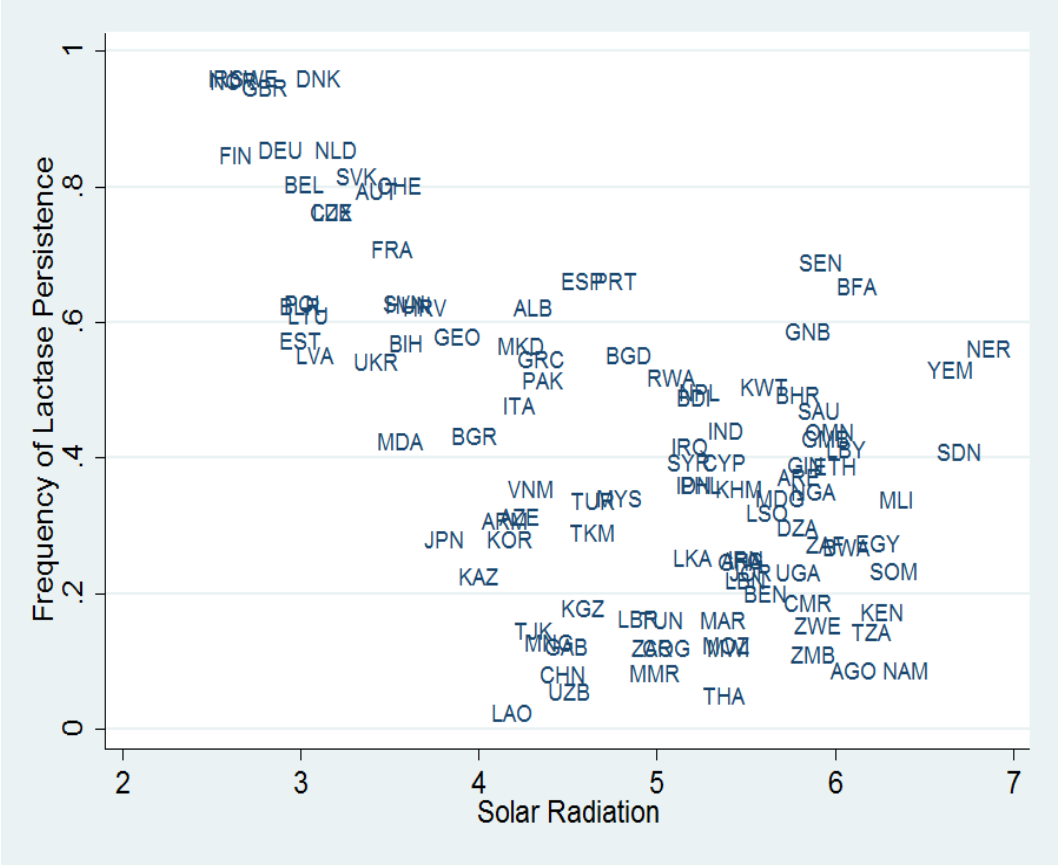

Figure 2.5: Solar Radiation and the Freq. of Lactase Persistence

tance from the equator, and a Western European dummy, but the relationship between solar radiation and population density may not be fully accounted for.[26] Second, the proposed relationship between sunlight and lactase persistence has come under recent criticism. Itan et al. (2009) simulate the evolution and spread of the gene associated with lactase persistence in Europeans. When controlling for relevant factors, they find that the low sunlight areas of Northern Europe do not correlate with a higher frequency of lactase persistence. Gerbault et al. (2009), however, find evidence supporting the relationship between solar radiation and lactase persistence. In short, the relationship between the frequency of lactase persistence and sunlight is still in question. Given the problems of our proposed instrument, we use IV estimation as a supplement to the estimates given by least squares. [27]

The baseline IV estimates are given in Table 2.13. Column (1) displays the bivariate regression of 1500 population densities on the frequency of lactase persistence. Solar radiation and its square have a strongly correlated with the frequency of lactase persistence; this is shown by the first stage F statistic of 113.39. The IV estimated coefficient of the frequency of lactase persistence is positive and significant at the 1% level. Additionally, the magnitude of the coefficient is similar to the bivariate, OLS estimate of Table 2.2.

In the bivariate case, however, the IV estimates may be bias. This is due to the agricultural benefits of sunlight. Therefore, columns (2) and (3) respectively add in millennia of agriculture and the suitability of agriculture, as well as other geographic variables. The inclusion of the additional controls does weaken the strength of our proposed instruments, but the instruments remain strong. The IV estimated coefficient of the frequency of lactase persistence is positive and significant while including the millennia of agriculture in column (2); however, the coefficient becomes insignificant in column (3), which includes the

[26] After including relevant controls, neither solar radiation or its square are insignificant from zero at the 10% level. However, they are jointly significant.

[27] Given the shortcomings of solar radiation, we have also used the number of potential domesticate animals from Hibbs and Olsson (2004). While correlated with the frequency of lactase persistence, the number of potential domesticate animals is a weak instrument. This is especially true when including additional controls.

Table 2.13: Baseline Estimations: Instrumental Variables

Dependent Variable: ln Population Density in 1500 CE					
		Extended Sample			Conservative Sample
	(1)	(2)	(3)	(4)	(5)
			First Stage Estimates:		
Solar Radiation	-0.9179***	-1.0053***	-0.7767***	-0.8513***	-0.8655***
	(0.1071)	(0.1005)	(0.12)	(0.1149)	(0.159)
Solar Radiation^2	0.0889***	0.1001***	0.0823***	0.0884***	0.0813***
	(0.0124)	(0.0116)	(0.0138)	(0.0132)	(0.0182)
First Stage F Statistic	106.602	118.999	23.284	30.606	21.133
			Second Stage Estimates:		
Freq. of Lactase Persistence	1.8155***	1.4604**	1.2686	2.2389**	4.3741***
	(0.6538)	(0.6041)	(1.0044)	(0.8543)	(0.9256)
Millennia of Agriculture		0.1754***		0.1411**	0.0481
		(0.0444)		(0.0555)	(0.0883)
Avg. Suitability of Agriculture			2.6500***	2.5912***	2.4444***
			(0.3631)	(0.3557)	(0.5606)
Dist. from Equator			-0.0296***	-0.0332***	-0.0537***
			(0.0109)	(0.0092)	(0.0127)
Sub-Saharan Africa			-1.2698***	-0.7049*	-2.0774***
			(0.3085)	(0.3890)	(0.6325)
Western Europe			0.9857**	0.7161*	-0.4226
			(0.4223)	(0.3960)	(0.4888)
N	103	103	103	103	48
R^2	0.2139	0.2941	0.5615	0.6148	0.6872
$Adj.R^2$	0.2061	0.2800	0.5388	0.5907	0.6414
F	7.7111	14.7675	31.9264	33.3851	29.0264

Notes: IV coefficients are reported in each column. *, **, and *** represent significance at the 10, 5, and 1% significance level, respectively. Robust standard errors are in parentheses. Table 2.13 is a re-estimates Table 2.2 with solar radiation and its square as an instrument for the frequency of lactase persistence.

suitability of agriculture and other relevant geographic controls.

Column (4) gives the baseline IV estimate. Solar radiation and its square are highly related to the frequency of lactase persistence. The first stage F statistic of column (4) is 31.13, which satisfies the maximum Stock-Yogo criteria. The IV estimated coefficient of the frequency of lactase persistence is positive, significant at the 1% level, and roughy identical to the OLS estimate.[28] Column (5) reduces the sample to the conservative estimates, leading to a slight reduction in instrument strength and a larger estimated coefficient of interest. Again, the estimated coefficient is similar to that given by least squares estimation. As stated before, the use of IV estimates is meant to supplement the estimations by least squares. The consistency of the IV coefficient in magnitude and significance to the OLS estimates provided further evidence that the relationship between dairying and population density is substantial.

Table 2.14 performs IV estimations while including the additional controls of Tables 2.7, 2.9, 2.10, and 2.11. Column (1) includes genetic distance from the U.K. into the baseline IV estimation of column (4) of Table 2.13. The instruments remain strong, and the estimated coefficient is similar to the baseline IV estimate, as well as the baseline least squares estimate. All environmental variables of Table 2.9 are included in column (2). Again the coefficient remains similar to the baseline estimates. Column (3) includes water access controls given by column (6) of Table 2.10. This results in a reduction in magnitude in the coefficient of lactase persistence, which leads to the coefficient being insignificant at the 10% level. Column (4) includes biogeographic controls of Hibbs and Olsson (2004) into the baseline estimation. The coefficient of interest is significant at the 5% level and is similar to previous estimates in magnitude. Column (5) includes all additional controls. The coefficient remains similar in magnitude to previous estimates with statistical significance dropping to the 10% level. [29] Aside from the lack of significance in column (3), IV estimates of the coefficient of

[28]The OLS estimated coefficient is 2.34, while the IV estimated coefficient is 2.24.

[29]Due to the sample adjustment of the Hibbs and Olsson data, we exclude biogeographic controls from

Table 2.14: Additional Controls: IV Estimates

Dependent Variable: ln Population Density in 1500 CE					
	(1)	(2)	(3)	(4)	(5)
			First Stage Estimates:		
Solar Radiation	-0.7717***	-0.7781***	-0.6521***	-0.7463***	-0.6736***
	(0.1206)	(0.1354)	(0.1258)	(0.1601)	(0.1492)
Solar Radiation^2	0.079***	0.0814***	0.0674***	0.0742***	0.0696***
	(0.0136)	(0.0151)	(0.0141)	(0.0194)	(0.0162)
First Stage F Statistic	22.83	16.62	14.08	18.62	10.25
			Second Stage Estimates:		
Freq. of Lactase Persistence	2.4261**	2.4606**	1.8378	2.3365**	2.4090*
	(0.9445)	(1.1148)	(1.1590)	(1.0749)	(1.3881)
Controls:					
Baseline	Y	Y	Y	Y	Y
Genetic	Y	N	N	N	Y
Environmental	N	Y	N	N	Y
Water Access	N	N	Y	N	Y
Biogeographic	N	N	N	Y	N
N	103	103	103	68	103
R^2	0.6154	0.6554	0.6185	0.7046	0.6600
$Adj.R^2$	0.5870	0.6051	0.5860	0.6588	0.5968
F	28.9754	17.6517	27.0862	19.1196	14.5441

Notes: IV coefficients are reported in each column. *, **, and *** represent significance at the 10, 5, and 1% significance level, respectively. Robust standard errors are in parentheses. Baseline controls include the millennia of agriculture, the average suitability of agriculture, the distance from the equator, and dummies for Sub-Saharan Africa and Western Europe. The additional genetic control is the genetic distance to the United Kingdom. Environmental controls include average elevation, average ruggedness and its square, the percent of a country within a desert or the tropics, a measure for the suitability of malaria, and whether or not a country belonged to the Roman Empire. Water access controls include the distance from the coast or a navigable river and the percent of a country within 100 kilometers from a coast or river. Biogeographic controls include the number of potential domesticate plants and animals and a measure for the country's East-West orientation.

the frequency of lactase persistence remain similar in magnitude and statistical significance to the baseline IV estimates, as well as the baseline OLS estimates.

The use of solar radiation is potentially problematic. However, IV estimations provide further evidence for the relationship between dairying and population densities posed in this paper. Furthermore, given the uniformity in magnitude and significance of the IV and OLS estimates, we have no reason to suspect a potential spurious relationship. Milk consumption aided the diet of early farmers; these benefits appear to have resulted in denser populations.

2.4 Conclusion

Diamond has stated that, "History followed different courses for different peoples because of differences among peoples' environments, not because of biological differences among peoples themselves." This paper does not intend to dispel this argument; rather this paper merely alters this view. Diamond is correct that the environment is the ultimate causal factor in the differing fates of humanity, but to assume the environment has not caused differences in people undermines one of the basic laws of evolution. The use of genetic frequencies above is merely an indicator for differing environments.

Toward this end, our work establishes an empirical relationship between milk consumption and pre-colonial development. Milk had the ability to improve both the quality and quantity of calories for Neolithic farmers and pastoralists. Both effects had the outcome of increasing populations. This relationship holds through a number of specifications and estimations, and gives important insights into the numerous advantages contained within Eurasian continent and Europe in particular.

column (5); however, when these variables are included, no significant change is seen.

Chapter 3

Genetic Determinants of Health Differentials: The Role of Disease in Natural Selection since the Neolithic Revolution

3.1 Introduction

The role of infectious diseases in low income economies is devastating in both a humanitarian and an economic sense. In the developed world the majority of diseases are associated with aging, e.g. cancer and heart disease, but in developing states preventable infectious diseases remain the primary cause of mortality (Cutler et al. 2006). Why are these diseases so prevalent and destructive in developing states? The most obvious answer is tied to the socioeconomic conditions that continually plague inhabitants of less developed countries. The widespread incidence and mortality associated with infectious disease is seen as a byproduct of a lack of hygiene, insufficient preventative measures, and inadequate treatment options, which ultimately stem from low levels of economic development.

This work questions whether there is a more fundamental determinant to the destructiveness of infectious disease. Particularly, are differences in disease resistance influenced by

historical exposure? A notorious illustration is given by the numerous contacts between European explorers and native populations of the Americas and Oceania, in which previously unexposed native populations exhibited greater susceptibility and loss of life from many common European diseases.[1] With this idea in mind, the current work explores the historical origins of infectious pathogens in order to exploit the natural selection of resistance to these pathogens. Furthermore, we argue that these genetic differences have remained to the present and are associated with the efficaciousness of infectious disease before the widespread distribution of effective medicines and vaccines.

To test this idea we construct a measure of inherent genetic resistance for a widespread group of infectious pathogens that developed as a result of the Neolithic Revolution. Differences in environments have led to differences in the initiation and sustainability of infectious diseases. In turn, this variation in disease environments has led to differing natural selections of genes associated with disease resistance. The large number of diseases developing from agriculture led to selection favoring variation *within* populations (Prugnolle et al. 2005; Jeffrey and Bangham 2000). In other words, variation within the immune system was naturally selected; this is referred to as balancing selection.[2] With this idea in mind, we create a measure of genetic diversity that is based solely on gene variants found within a key component of the immune system, the major histocompatibility complex (MHC). In humans the MHC is referred to as the human leukocyte antigen (HLA) region and this region also contains the greatest amount of diversity within the human genome (Hughes and Yeager 1998; Jeffrey and Bangham 2000).[3] Using this measure, we then estimate differences in the pre-medicinal health outcomes of states.

Our measure of genetic diversity, in theory, results from natural selection since the Neolithic Revolution. The effects of the agricultural transition on historic disease environments

[1] Another example is given by the sickle cell trait, which provides resistance to malaria in Sub-Saharan Africa. The sickle cell trait is the result of historical exposure to malaria.

[2] Section 2.2 gives a full explanation of balancing selection.

[3] MHC and HLA can be used interchangeably when discussing humans.

are twofold. First, agriculture allowed for the development of large, dense populations. Large, dense populations, in turn, allowed for an ease in the transmission of infectious diseases, as well as a large number of potential hosts. It is for this reason that diseases resulting from the Neolithic Revolution are known as crowd diseases (Wolfe et al. 2007). Second, the domestication of animals in the Neolithic provided closer contact between animals and humans. This close contact allowed animal pathogens to infect human hosts (Wolfe et al. 2007). The Neolithic Revolution provided the conditions for the initiation and sustainability of infectious crowd diseases. Societies which domesticated animals earlier and developed large, dense populations were the ones most likely to encounter infectious crowd diseases. This earlier exposure to disease led to the selection of genes that provide resistance to these diseases.

Starting with the Columbian Exchange, and accelerated by the mass development of roads, rail lines, and air ports, infectious pathogens initiated by the Neolithic Revolution have spread throughout the world (Crosby 1973; Arroyo et al. 2006; Brownstein et al. 2006; Wilson 1995). This rapid pace of globalization has created a global disease pool, resulting in the introduction of diseases into previously unexposed populations (McNeil 1976; Harrison 2004). Given the slow change of the genome, genetic resistance, or susceptibility, is assumed to be relatively constant during this period of globalization, while exposure to infectious pathogens has changed. This implies that differences in selection due to historical exposure provide an inherent protection to this global disease environment, and should therefore be associated with differing health outcomes.

In a recent paper, Galor and Moav (2007; hereafter GM) highlight the role of historical environmental differences on contemporary health outcomes. They show that an earlier transition to agriculture is associated with higher levels of life expectancy today. Those societies who adopted agriculture at an earlier date have had an advantage in adapting to the new agricultural environment. In particular, GM (P. 1) state, "...the Neolithic transition altered the evolutionary optimal allocation of resources towards somatic investment, repairs,

and maintenance (e.g., enhanced immune system, DNA repairs, accurate gene regulation, tumor suppression, and antioxidants)." To test this theory GM use a migration weighted average of the millennia a particular country has practiced agriculture. The authors find a strong positive relationship between a prolonged history of agriculture and variations in health outcomes in the year 2000, which gives credence to theory of adaptation proposed by GM. However, GM aren't able to measure adaptation directly; the weighted millennia of agriculture is a proxy for adaptation. The current work seeks to build on that of GM by measuring a specific adaptation to the agricultural environment: genetic variation within the HLA region.

The relationship between aggregate genetic variations and economic outcomes is explored in a number of new papers. Ashraf and Galor (forthcoming) show that genetic variation within a country leads to differences in historical and contemporary levels of development. The authors suggest this is due to benefits resulting from moderate levels of genetic variation, where lower levels of genetic variation are associated with lessened creativity and higher levels of genetic variation are associated with frequent in-fighting. The method used to measure of genetic variation used in Ashraf and Galor (forthcoming) is the same as in the current work; however, we consider only genes within the HLA system that have been influenced by presence of infectious pathogens. A complete discussion of this variable is given in Section 3.1.

Spolaore and Wacziarg (2008; hereafter SW) use genetic difference as an explanation for the diffusion of technology between states. In short, SW theorize that a greater genetic distance to the technological frontier is associated with a lag in the adoption of new technologies and, therefore, a lower level of income.[4] SW consider only random, or neutral, mutations that have occurred due to the separation between populations; we, on the other

[4]The technological frontier is the U.S. in the year 2000 and Great Britain in the year 1500. The measure of genetic distance which is used in SW is Sewall Wright's fixation index, or a measure of genetic diversity *between* populations.

hand, consider genes which have encountered selection.[5] In addition to the work of SW, Guiso et al. (2004) and Giuliano et al. (2006) use genetic distance in the explanation of bilateral trade flows between countries in Europe.

3.2 The Neolithic Revolution and the Natural Selection of Disease Resistance

This section will i) discuss the role of the Neolithic Revolution in the initiation and sustainability of infectious crowd diseases and ii) describe the natural selection for variation within the HLA region of the genome.

3.2.1 Crowd Disease

The rise of infectious disease in man is dependent upon agriculture. The domestication of animals created close contact between farmers and their animals, which allowed animal pathogens to infect new human hosts (Wolfe et al. 2007). Diphtheria, influenza A, measles, mumps, pertussis (whooping cough), rotavirus A, smallpox, and tuberculosis all have similar diseases within the domesticate animals of Eurasia and "probably or possibly reached humans from domesticate animals (Wolfe et al. 2007, P. 281)." Those peoples who first domesticated a particular animal had a greater probability of contracting the domesticate's diseases; implying that the peoples of Eurasia, with the highest number of potential domesticate animals, were the initial hosts of many crowd diseases (Diamond 1998; Hibbs and Olsson 2004).

In addition to initiation, the sustainability of a pathogen within a population is necessary for the development of genetic resistance. Endemicity, or the sustained presence, of infectious crowd disease is dependent on population density (Dobson 1996; Anderson and May 1991; Wolfe et al. 2007). In order for a disease to persist within a population, the

[5]To test for directional selection, or the selection to unity of a gene that generates a favorable trait, we construct a measure of genetic distance comprised of HLA genes. This is discussed in Section 5.3.

population must be large enough so that newly susceptible individuals, or hosts, are present. The diseases of interest in this paper are those that either kill the host or provide the host with antibodies so that he or she develops immunity to the disease. This implies that in small populations all susceptible individuals will either die or become immune, causing the disease itself to die out. As an example, it has been shown that measles becomes endemic in island communities with populations roughly greater than 500,000 individuals (Black 1966). If populations aren't sufficient in size, epidemics occur in which the disease sweeps through a population; leaving its members either dead or immune.

Hunter-gather societies could not support large enough populations for endemic diseases. Only the relatively large societies resulting from agriculture can supply hosts in such large numbers in which the disease could be continually maintained. Eurasian countries contained an advantage in the initiation of agriculture, implying these states had the necessary population size to replenish hosts necessary for the endemicity of the pathogens under consideration (Diamond 1998). Larger populations also led to greater cities that facilitated the spread of disease through closer contacts and lower hygiene (McNeil 1976). Additionally, the sedentary lifestyle of the agricultural environment allowed for the contamination of water supplies and the collection of rodents and other pests that carry vectors for disease.[6]

While large populations are necessary for the endemicity of infectious diseases, they are not sufficient in explaining the differences in disease resistant alleles. This point is most apparent when considering the ruinous results disease played on New World populations. The Mayans, Aztecs, Incas, and certain North Amerindian communities all developed agriculture and had populations sufficient in size to support the endemicity of disease, yet obviously had not developed an inherent genetic resistance to sustain the diseases from European conquerors and settlers during the colonial period (Crosby 1986). The reason is that these societies simply had not been exposed to diseases; if the selective force (i.e.,

[6]Plague and typhus are primarily distributed through lice, which are native to rodents that can only be supported in large sedentary human settlements.

disease) is not present, then selection for disease resistance will not take place.

A hypothesis posed by Barnes et al. (2010) corroborates the role dense populations in selection for resistance. Barnes et al. (2010) find that a history of living within a city has a close association with genetic resistance to tuberculosis. In summary, the highly dense populations associated with cities have allowed for a greater spread and sustainability of tuberculosis, which in turn, has led to a greater selection for an allele that provides resistance to tuberculosis.[7] Their findings suggest differential disease environments have led to differences in adaptation. This idea is naturally extended by exploring the effect of differential adaptation on contemporary health outcomes; this is our main hypothesis.

The development of infectious crowd disease is dependent upon both the wide domestication of animals and large, dense societies. Eurasia contained the advantage of contracting diseases earlier and also having large enough populations in which to sustain the particular diseases. In the words of Wolfe et al. (2007, P. 281):

> Thus, the rise of agriculture starting 11,000 years ago played multiple roles in the evolution of animal pathogens into human pathogens. Those roles included both generation of the large human populations necessary for the evolution and persistence of human crowd diseases, and generation of large populations of domestic animals. Moreover, as illustrated by influenza A, these domestic animal herds served as efficient conduits for pathogen transfers from wild animals to humans, and in the process may have evolved specialized crowd diseases of their own.

This process led to a continual selection for individuals containing a greater inherent resistance. The selection process is explored in the next section.

[7]Selection for resistance to tuberculosis, relative to other crowd diseases, should occur at a slower rate. Tuberculosis has a golden age of about 15-25 in which the mortality of the disease becomes quite low; this corresponds to the ability to produce offspring. Measles and mumps, on the other hand, usually affect infants who are unable to produce offspring.

3.2.2 Pathogen Driven Selection

If a disease enters into a primitive society in which no medicine exists, some individuals may die from the disease while others may not. It is this variation amongst individuals, corresponding to variations within the genome, which causes disease resistance to be selected. This is natural selection in which the strong survive, where in this case, strength is determined by some unseen phenotypic difference that allows some to be more resistant to disease (e.g., better recognition of potential infections, better disposal of harmful pathogens,etc.).[8] Furthermore, disease environments have differed, leading to differences in selection. This concept is expressed by Inhorn and Brown (1990, P. 89):

> ...infectious diseases including both great epidemics, such as plague and small pox, which have devastated human populations from ancient to modern times, and less dramatic, unnamed viral and bacterial infections causing high infant mortality have likely claimed more lives than all wars, noninfectious diseases, and natural disasters taken together. In the face of such attack by microscopic invaders, human populations have been forced to adapt to infectious agents on the levels of both genes and culture.[9] As agents of natural selection, infectious diseases have played a major role in the evolution of the human species.

Therefore, holding socioeconomic conditions constant across peoples, those societies that have been in contact with infectious crowd diseases for longer periods, or have had more time to adapt to the infectious diseases, have developed a greater genetic resistance to these particular diseases.

The high level of variation within the HLA system is based on a theory of balanced selection (de Bakker et al. 2006; Jeffrey and Bangham 2000; Traherne et al. 2006; Klein

[8] A phenotype is the expression of the genotype (Hartl and Clark 2007). The recognition and response of the immune system is a phenotypic expression of the underlying genes that constitute the HLA system.

[9] An example of cultural adaptation would be the washing hands, thoroughly cooking meat, or the wearing of a surgical mask while in public. Footnote our own.

1987). Balancing selection is selection for genetic diversity and results from two distinct reasons: overdominance and frequency-dependence (Slade and McCallum 1992). Overdominance implies heterozygotes, or individuals with differing alleles at a particular locus, have an advantage compared to homozygotes, or individuals with identical alleles at a particular locus.[10] A prime example of overdominance is the advantage conferred by the sickle-cell trait (Allison 1956). Heterozygous individuals contain a greater resistance to malaria, while homozygotes either contain no resistance to malaria or are afflicted by sickle-cell anemia. This leads to the natural selection of variation at the gene locus responsible for the the sickle-cell trait.

Frequency-dependent selection results from a comparative advantage of rare alleles. Infectious pathogens don't constitute a static selection pressure. Infectious pathogens—bacteria, viruses, protozoa, etc.—are living things also undergoing natural selection. If a particular allele were to provide complete resistance to a certain pathogen, variants of the pathogen, which avoid resistance, would thrive. In effect, this results to an "arms race" between the pathogen and the person. The relatively short time between generations of most pathogens, however, provides a time advantage in this "arms race." This implies that any resistance developed in the human genome should be overcome by genetic mutations within the pathogen that avoid the resistance provided by the genome. In other words, infectious pathogens have greater defenses to more common HLA gene variants; therefore, rarer, or lesser frequency, HLA alleles are better able to recognize and dispose of disease, implying a constant selection for rarer HLA alleles. As a result, the optimal strategy for disease resistance is variation, or allowing alleles associated with recognition to be played in equal frequency. Prugnolle et al. (2005) confirm this idea in finding that pathogen richness, or a high number of infectious pathogens, is associated with diversity within the HLA system. Furthermore, adaptation of infectious pathogens is routinely seen in the development of

[10]Individuals contain alleles at a gene locus from both the mother and father, implying two alleles at a given locus.

antibiotic resistance.

Natural selection has taken place within the HLA system since the initiation of agriculture (Sabetti et al. 2006). This natural selection has led to high levels of diversity within the HLA system, implying balancing selection (Prugnolle et al. 2005). Differences in HLA diversity remain, and these differences affect the immune response to the large number of diseases developed during the Neolithic Revolution (Bhatia et al. 1995; Black 1994; Black et al. 1974) Given differences in immune response and the widespread distribution of Neolithic crowd diseases, we postulate that differences in HLA diversity have an affect on contemporary health outcomes before the distribution of effective medicines and vaccines.

3.3 Disease Based Genetic Diversity

This section will outline the creation of the genetic diversity measure. First, a commonly used measure of genetic diversity is described. Second, we will discuss specific gene variants that will comprise our measure of genetic diversity. Finally, we will discuss aggregation to the country level.

3.3.1 A Measure of Genetic Diversity

In a recent work Ashraf and Galor (forthcoming; hereafter AG) explore the role of genetic variation in explaining historical and contemporary levels of development. In order to measure genetic diversity AG use a common measure within population genetics: expected heterozygosity. Expected heterozygosity is roughly defined as "the probability that two randomly selected individuals differ with respect to the gene in question (AG, P. 3)." Expected heterozygosity is a function of gene variants, or alleles, at a particular site on the genome, or locus. Mathematically, expected heterozygosity is defined by:

$$H_{exp} = 1 - \frac{1}{m} \sum_{l=1}^{m} \sum_{i=1}^{k_l} p_i^2 \tag{3.1}$$

where p_i represents the frequency of allele i, and expected heterozygosity is found by the average across m loci.

Our measure of heterozygosity differs slightly from that found in AG. AG attempt to measure variation across the genome, not a certain region of the genome; this is done to measure the effects of fractionalization and creativity associated with high and low levels of diversity, respectively. Our work differs in that we seek to measure heterozygosity in order to show balancing selection from the numerous infectious pathogens that became endemic after the Neolithic Revolution. Therefore, we only consider genes within a key component of the immune system, the major histocompatiblity complex.

3.3.2 Alleles Associated with Infectious Disease: The Major Histocompatibility Complex

The specific genes to be considered in the construction of our genetic diversity measure are based on the major histocompatibility complex . The major histocompatibility complex is a group of genes associated with the recognition of foreign substances within the body and is very important in disease resistance and susceptibility (Klein 1987; Traherne et al. 2006). In short, the MHC is responsible for locating foreign proteins in order to direct cells of the immune system to initiate an immune response (Piertney and Oliver 2006).

In humans the MHC is known as human leukocyte antigen system (Encyclopedia Britannica 2011).[11] The HLA system is a cluster of 239 genes located on the sixth chromosome (Shiina et al. 2004). The MHC is broken into two major classes, Class I and Class II, with both classes being associated with the recognition of certain pathogens.[12] This work, however, targets the entire system, and not sole genes. The use of all gene variants within the HLA system allows for a more complete measurement of diversity resulting from exposure the numerous Neolithic crowd diseases.

[11] White blood cells are also known as leukocytes.

[12] In the recognition of cells, Class I molecules are expressed on nucleated cells and are associated with defense against viruses, while Class II molecules are expressed on antigen-presenting cells and are associated with extracellular parasites (Piertney and Oliver 2006).

In the construction of our expected heterozygosity measure we consider gene variants within the HLA system; therefore, our main measure is HLA heterozygosity. HLA heterozygosity is constructed with data on SNP's from the Allele Frequency Database at Yale University, referred to as ALFRED. A SNP (pronounced "snip") is a single change along a strand of DNA. Each SNP has two variants, or alleles. From the website, "ALFRED is a free, web-accessible, curated compilation of allele frequency data on DNA sequence polymorphisms in anthropologically defined human populations." The use of the ALFRED gives allele frequency data for 19 HLA genes of 51 ethnic groups. In calculating heterozygosity, 156 SNPs are used from the 19 HLA genes.

The HLA system is associated with resistance and susceptibility to infectious disease (Traherne et al. 2006). Theoretical differences should exist within this system due to differences in the disease environments (Jeffrey and Bangham 2000). Our primary measure quantifies these differences by measuring diversity within the HLA system. The next subsection explains the aggregation of ethnic groups to the countries. And the following subsection defines the infectious crowd disease origin, from which HLA genetic distance is taken.

3.3.3 Aggregation from Ethnic Groups to Country

Allele frequency data is given by distinct ethnic groups; however, many (or most) relevant economic data are given only on the country level. This implies that an aggregation is needed in which countries are constructed of ethnic groups. Following Spolaore and Wacziarg (2009), we aggregate ethnic groups to the country level with the use of ethnic compositions found in Alesina et al. (2003). [13]

The matching of ethnic groups from ALFRED to Alesina et al. (2003) is not perfect.

[13]The ethnic compositions found in Alesina et al. (2003) are from the 1990's. This creates a measurement problem in measuring the effect of HLA genetic distance on 1960 life expectancy and the years of life lost to communicable disease in 2002, our two primary dependent variables. However, there is no reason to suspect a nonrandom error. Therefore, this measurement should lead to an attenuation bias, understating the true relationship of HLA genetic distance.

ALFRED contains allele frequency data for 51 differing ethnic groups, while Alesina et al. (2003) contain hundreds of differing ethnic groups in the ethnic composition of countries. In order to get around this problem, language classifications are used to match distinct ethnic groups in Alesina et al. (2003) to a similar ethnic group in ALFRED. For example, Hutu from Alesina are classified as Bantu in Alfred, Amayara are classified as Amerindian, and Polish are classified as Russian.

In addition to the matching of ethnic groups, additional ethnic groups have been created through combinations found in ALFRED. [14] The primary example of this is given by the ethnicity Black in Alesina et al. (2003). The term Black refers only to the color of skin, not ethnicity. Ultimately, Black indicates a hereditary history from Sub-Saharan Africa, but Sub-Saharan Africa is not made up of a sole ethnic group. In order to get around this problem, I first assign Sub-Saharan African countries to one of three ethnic groups based on a map in Shillington (1989, P. 50; Reader 2002, P. 692) [15] Next, using data on the Trans-Atlantic slave trade from Nunn (2009), we create a representative Black ethnic group through the weighted average of the number of slaves from an African country that has been assigned to a specific ethnic group. This leads to the representative Black ethnic group comprised of 49% Bantu, 12% Mandenka, and 39% Yoruba. Other notable combinations include: White which is 50% Italian and 50% French, Mestizo which is 50% White and 50% Amerindian or Mayan (depending on whether the respective country is in North or South America), and Germanic which is 50% French and 50% Orcadian. Through this method I am able to find the genetic diversity 175 countries, which is based solely on genes associated with disease resistance.

[14] These combinations are not counted in the calculation of the heterozygosity score.

[15] West African countries are assigned to Mandenka, countries around the Gulf of Guinea are assigned to Yoruba, and South African countries are assigned to Bantu. Note that most Northeast African/Nilo-Saharan states are unused due to the lack of a close ethnic group in ALFRED.

3.4 Other Data

Table 3.1 gives summary statistics for all variables used in estimation. The origin of our measure of HLA heterozygosity is given in detail above, while sources and explanations of all other variables are given below.

Table 3.1: Summary Statistics of Baseline Variables

Variable:	N	Mean	Std. Dev.	Min	Max
HLA Heterozygosity	155	0.3167	0.0237	0.2110	0.3529
Continent:					
Europe	36	0.3337	0.0102	0.3153	0.3529
Asia	37	0.3154	0.0144	0.2711	0.3298
Africa	41	0.3169	0.0171	0.2844	0.3352
Americas	32	0.3108	0.0224	0.2588	0.3503
Oceania	9	0.2738	0.0503	0.211	0.3439
Life Expectancy 1960	155	54.4396	11.7350	31.1261	73.5498
GDP per capita 1960	155	4783.2839	6096.0276	425.0000	4.86e+04
Genetic Dist. from USA	155	975.1945	535.1245	0.0000	2088.0100
Ethnic Fractionalization	155	0.4231	0.2547	0.0000	0.9302
Absolute Latitude	155	25.9355	16.4027	0.0000	60.0000
Out of Africa Migratory Dist.	155	6882.7103	3375.8363	2883.0200	1.86e+04
Fraction of Population from Eurasia	155	0.6287	0.4337	0.0000	1.0000
Weighted Millennia of Agriculture	140	5.4356	2.1448	1.3570	10.4000

3.4.1 Dependent Variable

The primary health variable to be considered in this paper is life expectancy in 1960 (WDI). The use of 1960 life expectancy is meant to capture health variations before the widespread distribution of effective medicines and vaccines (Acemoglu and Johnson 2006).[16] In theory,

[16] Acemoglu and Johnson (2006) exploit the "epidemiological transition" which began in the 1940s. While many vaccines and medicines were invented in the 1940s and 1950s, the widespread distribution and use of these medicines was slowed. Earlier years have been considered, but due to a lack of measurement in relatively poor countries, very few data are available. Therefore, the use of 1960 is seen as a trade-off between data and the timing of medicinal distributions.

our measure of genetic diversity should affect the resistance of a country's population to infectious crowd disease in the absence of medicine; i.e., if a country's population has relatively low levels HLA heterozygosity, then in the absence of medicine, a greater fraction of the population will die from infectious crowd diseases. This higher mortality then is associated with a lower level of life expectancy. In other words, life expectancy indirectly measures the burden of disease. And this burden of disease is more accurately measured before the widespread use of effective medicines.

3.4.2 Control Variables

Given the importance of income to health outcomes, GDP per capita needs to be accounted for. GDP per capita data from 1960 come from estimates by Maddison found in Avakov (2010). Estimates are used due to the lack of data in Oceanic and Sub-Saharan African countries.

The primary aim of HLA heterozygosity is to measure country level susceptibility to crowd disease. The use of genetic diversity, however, may capture unintended genetic differences. Therefore, it is essential to control for neutral genetic differences. This is achieved through the use of genetic distance given by Spolaore and Wacziarg (2009; hereafter SW). SW's measure of genetic distance is calculated by F_{ST}, a measure of between variation *between* groups, and is based on 120 alleles from Cavalli-Sforza et al. (1994). SW argue that their resulting measure of genetic distance is not based on selection, but instead, based on random genetic drift between isolated populations; therefore, SW's measure of genetic distance can be seen as a baseline level of genetic differentiation. SW posit that this measure of genetic differences is associated with the diffusion of technology; therefore, in creating a cross-country measure we use a country's genetic distance to the United States, the contemporary technology frontier. Additionally, controlling for this baseline genetic differentiation can be used to capture biological and cultural differences that may be correlated with HLA heterozygosity. As a further control for genetic differentiation, we

create a measure of genetic distance based on our data of the HLA region. This is intended to control for directional selection, or a rise in the frequency of particular alleles, that may be associated with disease resistance. This measure is discussed further in Sec. 5.3.

In aggregation to the country level, data from Alesina et al. (2003) are used. This implies our measure of HLA heterozygosity may inherently account for ethnic fractionalization. To measure the direct effect of diversity within the HLA system, we control for ethnic fractionalization.

Additional controls to be used include the malaria ecology index (Kiszewski et al. 2004), the weighted millennia of agriculture (Putterman 2007; Putterman and Weil 2011), the rule of law (Dollar and Kraay 2001), and the fraction of a countries population derived from Eurasia.

3.4.3 Migratory Distance from East Africa

Homo sapiens originated within Africa roughly 200,000 years ago; around 100,000 years ago modern humans began to migrate out of East Africa, resulting in human colonization of the entire planet (Ashraf and Galor *forthcoming*; Prugnolle et al. 2005a; Ramachandran et al. 2005). This process of migrating out of Africa resulted in population bottlenecks and a decline in genetic diversity as a result of migratory distance. In other words, migrating populations carried only a fraction of genetic diversity, reducing the overall level of genetic diversity within the sub-population. This implies a clear linear relationship between heterozygosity and migratory distance from East Africa, the jumping off point for the "Out of Africa" migrations.

This relationship is exploited in Ashraf and Galor (forthcoming). Figure 3.1 plots the expected migratory paths out of East Africa, in which "Out of Africa" migratory distance is calculated as sum of the distance between country and the closest way point and the distance of this way point to East Africa (along the proposed migratory path). Using neutral (not solely HLA) genetic variation, the migratory distance from East Africa explains roughly

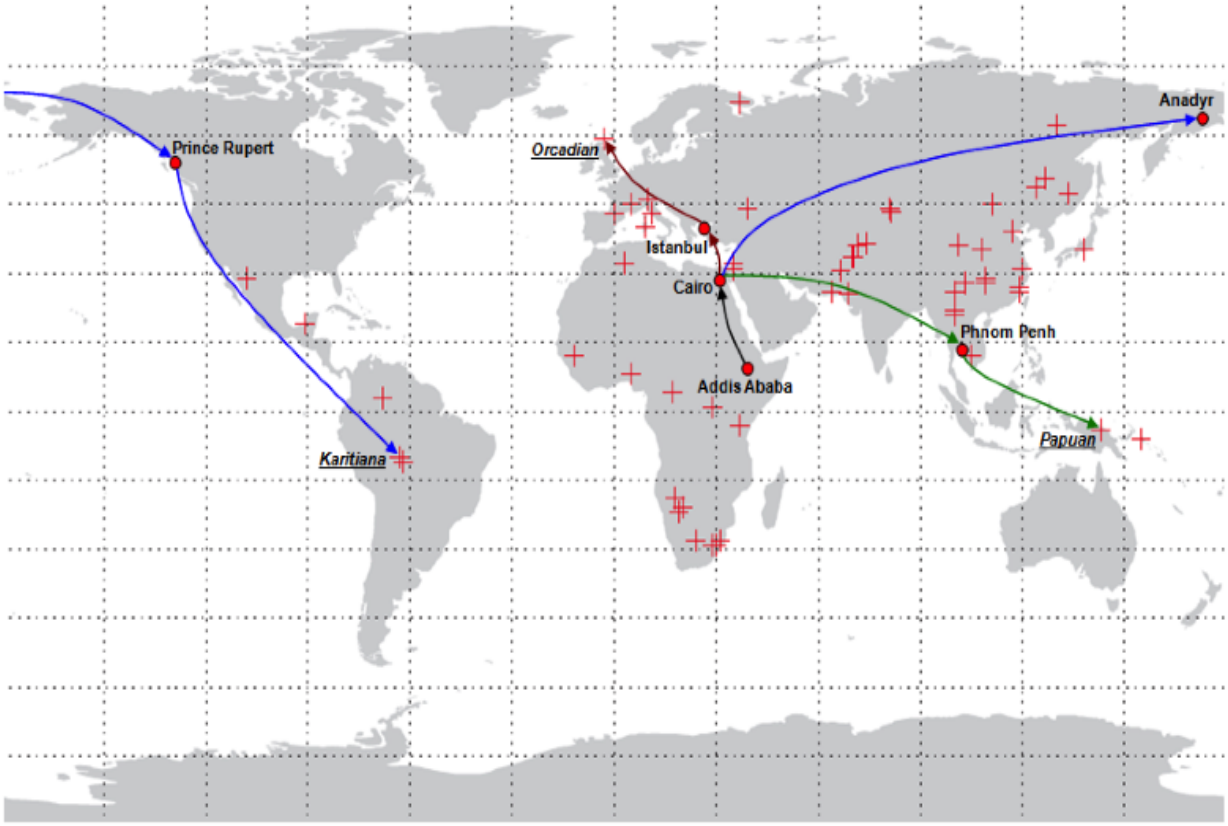

Figure 3.1: Migratory Paths from East Africa (from Ashraf and Galor *AER 2012*)

85% of the variation in expected heterozygosity. Given this strong relationship, AG use migratory distance to predict country level heterozygosity.

Our measure of HLA heterozygosity isn't based on neutral variation; it is based on a key component of the immune system that has under gone recent selection (Sabettie et al. 2006). This strong linear relationship between "Out of Africa" migratory distance and HLA heterozygosity need not be the case, since genetic bottlenecks due to the serial founder effect are not the only factor in determining HLA heterozygosity. Prugnolle et al. (2005b) find migratory distance explains only 17% to 39% of diversity within HLA genes, not 86% as found in Ashraf and Galor (2012). In explaining the residual variation in HLA heterozygosity, Prugnolle et al. (2005b) show that pathogen richness has a strong, positive association with variation within HLA genes. [17] This implies infectious pathogens are responsible for shaping the diversity within HLA genes.

Instead of pathogen richness, which may be endogenous, we too use migratory distance from East Africa as an instrument for HLA heterozygosity, but we exploit the nonlinearity in the relationship between migratory distance and HLA heterozygosity. This nonlinearity is the result of disease initiation from the Neolithic Revolution. Middle Eastern countries that initiated agriculture at an earlier date are shown to have a higher HLA heterozygosity. This is due to natural selection within the HLA system, resulting from exposure to the wide array of pathogens. Figure 3.2 gives the relationship between country level HLA heterozygosity and "Out of Africa" migratory distance.

Country level migratory distance is found from migratory distance of ethnic groups given in Ashraf and Galor (2012). Ethnic migratory distances are aggregated to the country level through data found in Alesina et al. (2003); the same weight is used in determining country level HLA heterozygosity.

[17] Pathogen richness is the total number of intracellular diseases within a country.

3.5 Results

3.5.1 Explaining HLA Heterozygosity

Two primary factors are responsible for explaining heterozygosity within the HLA system. First, heterozygosity is a declining function of the distance from East Africa. Due to the serial founder effect, variation within populations declined as people moved out of africa. Second, the early development of agriculture and the intense domestication of animals within Eurasia facilitated the development of a large number of infectious pathogens. In explaining HLA heterozygosity, the estimating equation is of the following form:

$$ln\ \text{HLA}_i = \alpha + \beta_1(ln\ \text{MD}_i) + \beta_2(ln\ \text{MD}_i)^2 + \boldsymbol{\gamma}'(\mathbf{NR}_i) + \epsilon_i$$

Where i is a country index, HLA is our measure of heterozygosity, MD is "Out of Africa" migratory distance, and **NR** is a vector of variables relating to the Neolithic Revolution. All estimations are by OLS with robust standard errors.

The relationship between "Out of Africa" migratory distance and heterozygosity within the HLA system is explored in column (1) of Table 3.2. The coefficients on HLA heterozygosity and its square indicate heterozygosity increased up to a certain distance outside of africa before decreasing due to the serial founder effect, with maximal heterozygosity found 4,600 km from East Africa. This distance corresponds to countries within the Middle East and Southern Europe, and further verifies the role of agriculture in shaping variation within the immune system. The relationship in column (1) is given by Figure 3.2.

Column (2) shows a direct affect between agriculture and HLA heterozygosity. The primary argument proposed by Galor and Moav (2008) is that an earlier adoption of agriculture led to a head start in adaptation to the Neolithic environment. In turn, this greater adaptation has led to a divergence in contemporary levels of life expectancy. One of the goals of the current work is to create an intermediate empirical measure of GM's adap-

Table 3.2: Explaining HLA Heterozygosity

Dependent Variable: ln HLA Heterozygosity					
	(1)	(2)	(3)	(4)	(5)
ln *Dist. Out of Africa*	1.8548***		1.6860***	1.0857***	1.0640***
	(0.2903)		(0.2651)	(0.2490)	(0.2462)
ln *Dist. Out of Africa*^{2}	-0.1099***		-0.1008***	-0.0670***	-0.0661***
	(0.0166)		(0.0152)	(0.0142)	(0.0140)
ln *Weighted Millennia of Agriculture*		0.0234**	0.0338***	-0.0262**	
		(0.0098)	(0.0084)	(0.0118)	
Fraction of Pop. from Eurasia				0.0847***	0.0541***
				(0.0174)	(0.0108)
ln *Weighted 1500 CE Pop. Density*					0.0103***
					(0.0032)
N	142	142	142	142	142
R^2	0.4539	0.0266	0.5064	0.6033	0.6148
F	48.7675	5.7228	42.2803	39.7284	43.8366

Notes: OLS coefficients are reported in each column. *, **, and *** represent significance at the 10, 5, and 1% significance level, respectively. Robust standard errors are in parentheses. Non-Eurasian sample excludes countries within Europe, Asia, and North Africa. Column (7) includes all available data, including countries within Eurasia. Weights for Millennia of Agriculture are from the Putterman and Weil Migration Matrix (2010). The dependent variable, HLA heterozygosity, is calculated using Equation (1).

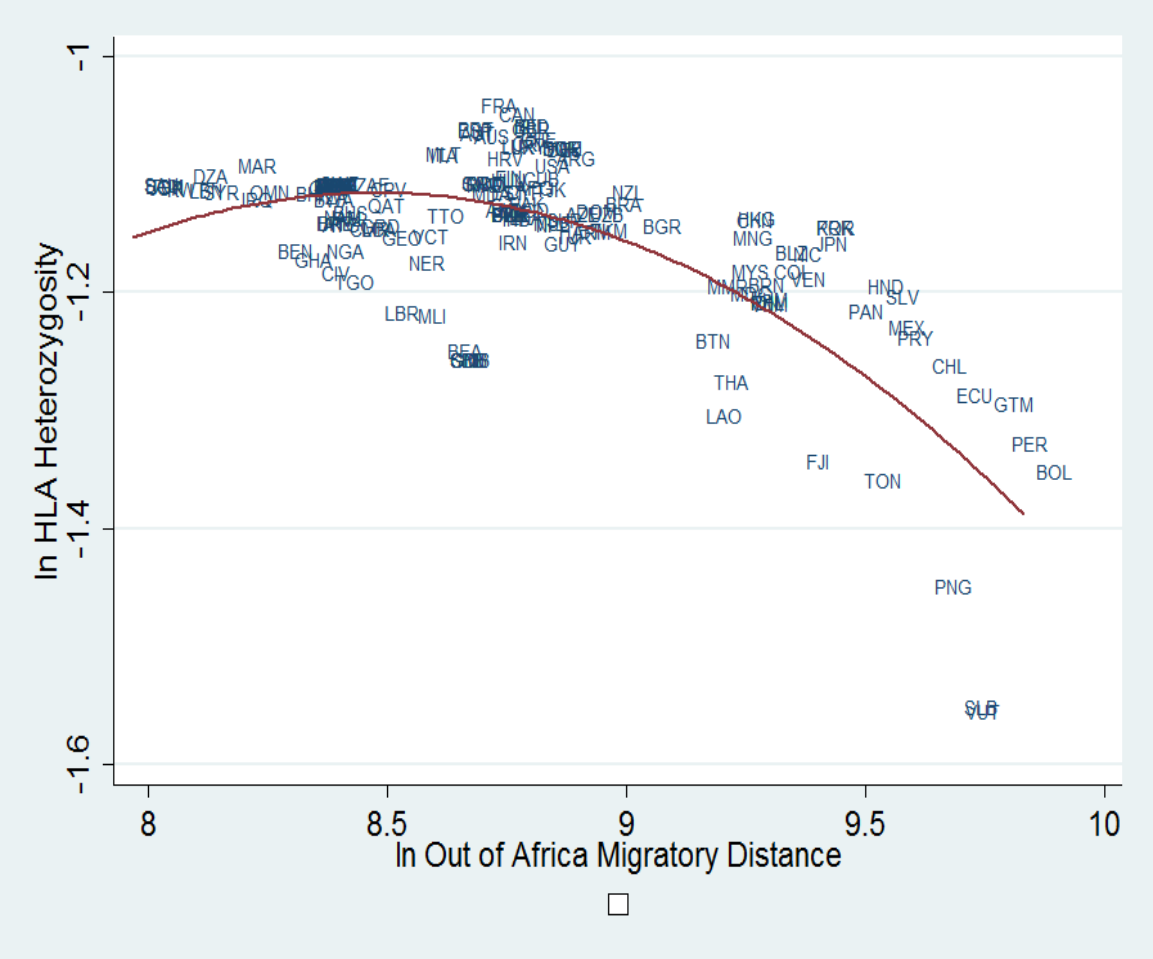

Figure 3.2: Relationship of "Out of Africa" Mig. Dist. and HLA Heterozygosity

tation. Column (2) shows the relationship between the natural log of weighted millennia of agriculture, where it is shown that a greater history of agriculture is associated with a greater genetic relationship to our disease origins. This relationship is significant at the 5% level with the weighted millennia of agriculture explaining roughly 3% of the variation in HLA heterozygosity. Specifically, column (2) shows that a 10% increase in the adoption of agriculture is associated with roughly a 0.2% decline in HLA genetic distance to the disease origin.

Column (3) satisfies the estimating equation above by including explanatory variables for "Out of Africa" migratory distance and the timing of the Neolithic Revolution. The direction of each coefficient is as expected, and all coefficients are significant at the 1% level. Migratory distance has a positive effect on HLA heterozygosity due to balancing selection from the development of Eurasian diseases, but outside of Eurasia, the serial founder effect is dominant, implying migratory distance results in a reduction in heterozygosity.

To further stress the role of Eurasia in the selection of variation within the HLA system, column (4) includes the fraction of a country's current population that is derived from Eurasia. This fraction has a positive and significant effect on HLA heterozygosity: a 10% increase in Eurasian derived population is associated with a 0.8% increase in HLA genetic diversity.

The inclusion of Eurasian derived population into the estimation of column (3) results in the sign of the coefficient of weighted millennia of agriculture to reverse. After controlling for the Eurasian population, and the high millennia of agriculture dates associated with Eurasia, the weighted millennia of agriculture accounts mainly for the prolonged agriculture found in the Americas. The Americas, however, were relatively disease free and geographically distant from East Africa, resulting in less genetic variability. This creates the negative coefficient of weighted millennia of agriculture seen in column (4).

The Neolithic Revolution provided a means for novel diseases to cross into human populations, as well as a large number of hosts, which allowed the pathogens to become endemic.

Therefore, substituting historical populations densities should alleviate the negative effect of agriculture in column (3). Column (5) shows estimation when weighted 1500 CE population densities are substituted for the weighted agricultural transition dates. Historical populations are shown to have a positive and significant effect on HLA heterozygosity, after controlling for the fraction of the population derived from Eurasia. All other coefficients are similar to previous estimates in magnitude and significant at the 1% level.

HLA diversity is a function of both distance from East Africa and the Neolithic Revolution. Distance out of Africa, is associated with a decline in genetic diversity (Ramachandran et al. 2005). This loss in diversity, however, has been overturned within Eurasia. This is due primarily to the development of a large array of diseases from the initiation of agriculture. The Neolithic Revolution provided close contact between numerous domesticate animal species and humans. This close contact facilitated the transmission of novel pathogens into human populations. This transmission was supported by the large, sedentary populations that resulted from the Neolithic Revolution. The prolonged exposure associated with agriculture led to balancing selection for genes responsible in the recognition of foreign pathogens. Therefore, we should see greater variations within the HLA system for Eurasians, despite a geographical distance from East Africa. The estimates of Table 3.2 confirm these two primary effects on HLA heterozygosity. The next subsection explores whether this inherent genetic variation can explain differences in contemporary health differences.

3.5.2 The Role of HLA Heterozygosity in Explaining Pre-Medicinal Life Expectancy

Genetic variation within the HLA system has naturally selected since the initiation of agriculture (Prugnolle et al. 2005; Sabeti et al. 2006). This variation provides an inherent resistance to the numerous, Eurasian crowd diseases that were spread across the globe following European colonization. The spread of disease was further accelerated by the

widespread development of ports, roads, rail lines, and air ports (Arroyo et al. 2006; Brownstein et al. 2006; Wilson 1995). Given the relatively slow pace of natural selection and the widespread distribution of infectious pathogens, we should expect differences in HLA heterozygosity to persist into contemporary times.

In order to show the contemporary effects of differences in HLA heterozygosity, we regress life expectancy of 1960 on our measure of HLA heterozygosity. The use of 1960 data is meant to capture the effects of inherent resistance before the widespread distribution of effective medicines and vaccines. Most effective medicines and vaccines were developed in the 1950's; however, data for 1950 is severely restricted, especially for lower income countries. We sacrifice the potential problems associated with medicinal distribution, in order to have a fuller sample.[18] The base specification is given by:

$$ln\ \mathrm{LE}_i^{1960} = \alpha + \beta(ln\ \mathrm{HLA}_i) + \boldsymbol{\gamma'G_i} + \boldsymbol{\delta'I_i^c} + \epsilon_i$$

Where i is a country indicator, LE^{1960} is life expectancy in 1960, and HLA is our measure of heterozygosity. $\boldsymbol{G}$ is a vector of relevant controls, including GDP per capita, genetic distance to the United States, ethnic fractionalization, and absolute latitude. I_i^c is an indicator variable as to whether or not country i is within continent c. The coefficient of interest is β, and all estimations are found through OLS with robust standard errors.

Column (1) of Table 3.3 gives the bivariate regression of 1960 life expectancy on HLA heterozygosity. The coefficient on HLA heterozygosity is positive and significant at the 1% level. In other words, greater genetic diversity within the HLA system is associated with higher pre-medicinal life expectancy. Specifically, a 10% increase in HLA heterozygosity is associated with a 10% increase in life expectancy. Column (2) includes continental dummies into the bivariate regression of column (1). The inclusion of continental dummies does not affect the magnitude or significance of the the coefficient of HLA heterozygosity. Further-

[18] Any potential effects of medicine should be captured within GDP per capita.

Table 3.3: Baseline Estimation of Life Expectancy in 1960: OLS

Dependent Variable: ln Life Expectancy in 1960							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ln *HLA Heterozygosity*	1.0155***	1.0467***	0.5633***	0.8891***	1.0057***	0.8667***	0.4575***
	(0.2350)	(0.1752)	(0.1619)	(0.1767)	(0.1488)	(0.1732)	(0.1422)
ln *GDP per capita 1960*			0.1187***				0.1173***
			(0.0201)				(0.0199)
ln *Genetic Dist. from USA*				-0.0477**			-0.0042
				(0.0204)			(0.0099)
ln *Ethnic Fractionalization*					-0.0425**		-0.0447***
					(0.0171)		(0.0129)
ln *Abs. Latitude*						0.0558***	0.0169
						(0.0153)	(0.0166)
Continental Dummies	N	Y	Y	Y	Y	Y	Y
N	155	155	155	155	155	155	155
R2	0.1367	0.5332	0.6901	0.5522	0.5581	0.5599	0.7242

Notes: OLS coefficients are reported in each column. *, **, and *** represent significance at the 10, 5, and 1% significance level, respectively. Robust standard errors are in parentheses. Continental dummies include indicator variables for the Americas and Africa.

more, given the limited number of ethnic groups used to determine HLA heterozygosity in Sub-Saharan Africa, the inclusion of continental dummies provides more precise estimation of the coefficient of interest.[19]

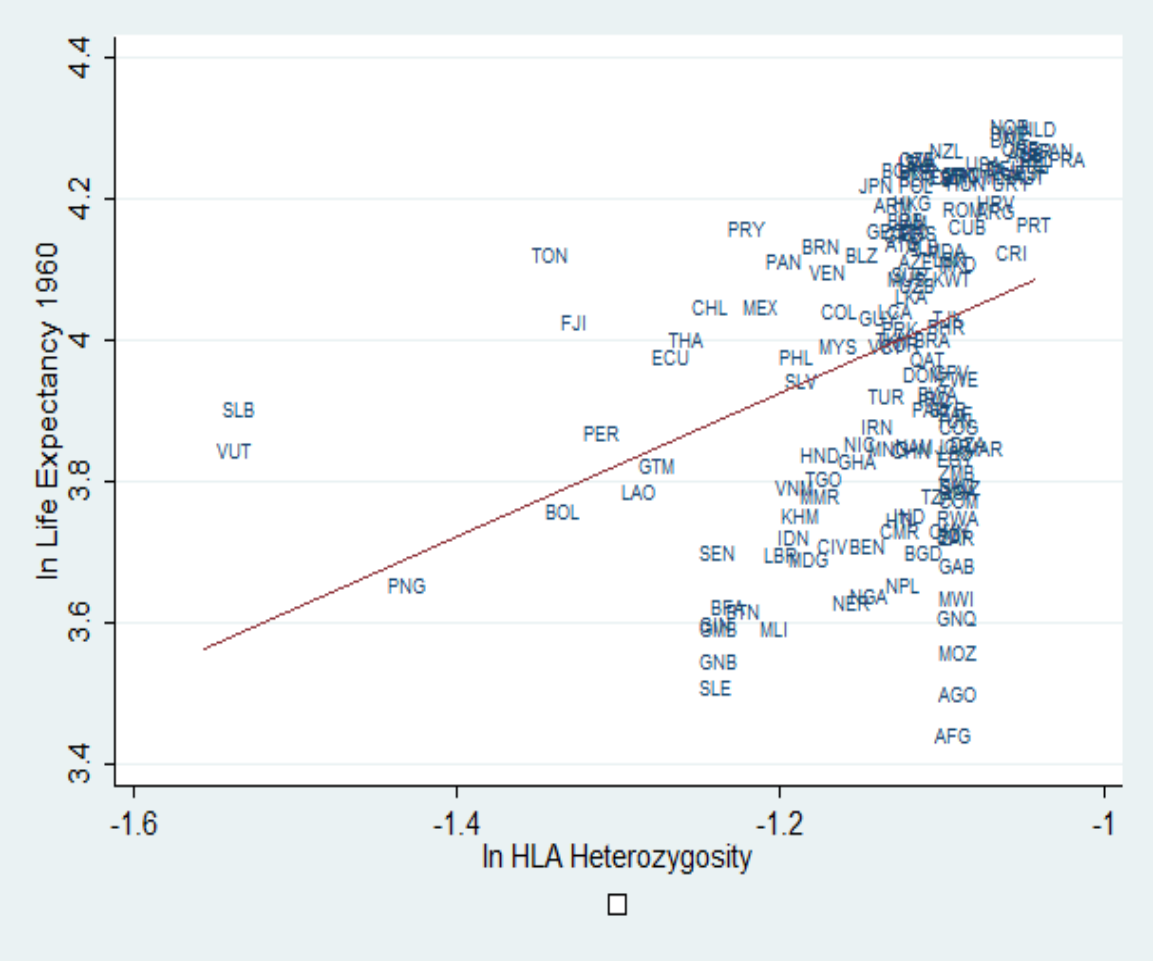

Figure 3.3: Relationship of HLA Heterozygosity and 1960 Life Expectancy

Columns (3)-(6) add in other relevant determinants of life expectancy in 1960 in a piecemeal fashion. The most obvious determinant of health differentials is income. More developed states are better able to provide greater nutrition, sanitation, and care to the

[19]Continental dummies include indicator variables for the Americas and Africa. Looking at Figure 3.2, there is very little difference in Sub-Saharan HLA heterozygosity, but great variation in 1960 life expectancy. Most countries within Sub-Saharan Africa are assigned to the Bantu ethnic group.

sick. Column (3) includes GDP per capita of 1960 into the estimation of column (2). The coefficient of HLA heterozygosity is roughly halved but remains positive and significant at the 1% level. As expected GDP per capita has a positive and significant effect on life expectancy.

Following the theory of Spolaore and Wacziarg (2009), a greater level of genetic distance from the technological frontier is associated with a slower diffusion of technology; this slow diffusion of technology, in turn, should be associated with lower production, lower medical technologies, and therefore, lower life expectancy. In order to accurately ensure HLA heterozygosity, which is based on genetic differences, isn't picking up an ommitted effect of technological diffusion, it is necessary to control for this measure of genetic distance. Column (4) includes the genetic distance from the United States into the bivariate estimation of column (2). The inclusion of genetic distance from the United States results in a slight attenuation in the coefficient of interest, but the effect of HLA heterozygosity remains significant. Genetic distance from the U.S. is shown to have a negative and significant effect on life expectancy in 1960, giving credence to the theory of Spolaore and Wacziarg (2009). This result indicates that our measure of HLA heterozygosity is not being driven by general genetic differentiation; instead, variation within the HLA system is picking up a specific effect of the genome on aggregate health outcomes.

The use of ethnic compositions from Alesina et al. (2003) implies a possible bias may arise from ethnic fractionalization. This is addressed in column (5), which includes ethnic fractionalization into the bivariate regression of column (2). The inclusion of ethnic fractionalization does not alter the magnitude or significance of our coefficient of interest. Column (6) controls for geographic differences in life expectancy by controlling for absolute latitude; this does not significantly alter the coefficient of interest.

All relevant determinants and controls are included in column (7), representing the base specification above. The effect of HLA heterozygosity is roughly half of the estimate given by the bivariate regression. The effect of HLA heterozygosity, however, remains positive and

significant at the 1% level. All other coefficients are as expected, although genetic distance from the U.S. and absolute latitude are statistically insignificant. The estimated coefficient of column (7) provides strong support for the main hypothesis of this paper: long running genetic differences within the immune system have an effect on health outcomes before the widespread distribution of medicine. This result is robust to the inclusion of an additional genetic control, GDP per capita, ethnic fractionalization, and absolute latitude.

Furthermore, the results of column (7) provide support for the prolonged effects of the timing of the Neolithic Revolution. The initiation of agriculture and the domestication of animals resulted in the development and sustainability of infectious pathogens. Prolonged exposure to these pathogens provided selection pressures favoring variation within the portion of the genome responsible for recognition of foreign bodies. Globalization resulted in the spread of these disease to populations that had no previous exposure, and therefore, lower diversity within the HLA system. Before efficacious medicines and vaccines, the difference in HLA diversity led to differences in how populations were able to cope with these infectious diseases, resulting in differences in life expectancy. This is confirmed by the results of Table 3.3.

As shown in Ashraf and Galor (forthcoming), "Out of Africa" migratory distance is a strong predictor of heterozygosity, or variation within the genome. Given the exogeneity of migratory distance, the distance from East Africa serves as an ideal instrument for our measure of HLA heterozygosity. The role of the Neolithic Revolution in providing and sustaining infectious pathogens, however, alters the linear relationship between "Out of Africa" migratory distance and HLA heterozygosity. As is shown in Table 3.2, Eurasian countries, which initiated agriculture earlier, have greater levels of diversity within the HLA system. This implies a nonlinear relationship between migratory distance from East Africa and HLA heterozygosity. This nonlinearity results from balancing selection provided by the numerous agricultural diseases. Therefore, as in column (1) of Table 3.2, we use migratory distance and its square as our primary instruments in Table 3.4.

Table 3.4: Baseline Estimation of Life Expectancy in 1960: IV

Dependent Variable: ln Life Expectancy in 1960							
	(1)	(2)	(3)	(4)	(5)	(6)	
ln *HLA Heterozygosity*	0.2858	0.9796***	0.5517***	0.7933***	1.0304***	0.7818***	0.5705***
	(0.2159)	(0.1897)	(0.1876)	(0.2047)	(0.1787)	(0.2049)	(0.1846)
ln *GDP per capita 1960*			0.1192***				0.1144***
			(0.0200)				(0.0200)
ln *Genetic Dist. from USA*				-0.0519**			-0.0017
				(0.0228)			(0.0100)
ln *Ethnic Fractionalization*					-0.0423**		-0.0442***
					(0.0170)		(0.0131)
ln *Abs. Latitude*						0.0594***	0.0143
						(0.0160)	(0.0170)
Continental Dummies	N	Y	Y	Y	Y	Y	Y
N	155	155	155	155	155	155	155
R^2	0.0661	0.5326	0.6901	0.5512	0.5581	0.5591	0.7230
First Stage F	28.648	38.993	36.986	35.412	37.331	37.769	34.062

Notes: IV coefficients are reported in each column. *, **, and *** represent significance at the 10, 5, and 1% significance level, respectively. Robust standard errors are in parentheses. Continental dummies include indicator variables for the Americas and Africa. The natural log of migratory distance from East Africa and its square are used as exogenous instruments.

Column (1) gives the bivariate IV estimates of regressing life expectancy in 1960 on HLA heterozygosity. The point estimate of the coefficient is positive but statistically insignificant. As stated earlier, Sub-Saharan African states have similar levels of HLA heterozygosity with a wide range of life expectancy in 1960. This is controlled for in column (2), which includes continental indicators. The IV estimated coefficient of column (2) is positive, significant at the 1% level, and similar in magnitude to the OLS estimate in column (2) of Table 3.3. In both columns (1) and (2) migratory distance from East Africa and its square are strong instruments with first stage F statistics of 29 and 39, respectively.

Columns (3)-(6) mirror the respective columns in Table 3.3. The IV estimates of Table 3.4 are very similar to the OLS estimates in both magnitude and significance of the coefficient of HLA heterozygosity. The IV estimates of the coefficient of HLA heterozygosity are positive and significant at the 1% level.

Column (7) of Table 3.4 satisfies the baseline estimating equation while instrumenting HLA heterozygosity with the migratory distance from East Africa and its square. The coefficient of HLA heterozygosity is positive and significant at the 1% level and slightly larger than the estimate given by OLS. This may be attributed to an attenuation bias due to measurement error in HLA heterozygosity. In particular, the estimated coefficient of column (7) states that a 10% increase in HLA heterozygosity leads to an 8% increase in life expectancy. For mean life expectancy in 1960, this corresponds to an increase in life expectancy of 4 years.

Through instrumental variables estimation and least squares estimation, greater diversity within the HLA system is shown to cause improvements in life expectancy before medicine. Without effective medicines and vaccines, individual resistance to infectious disease was dependent upon either cultural or genetic traits. After controlling for necessary cultural differences and instrumenting genetic differences, heterozygosity within the HLA system does provide an aggregate health advantage. In other words, holding constant cultural resistance, the genetic differences do lead to differences in the efficaciousness of

infectious pathogens. The next section explores the sensitivity of this relationship through sample truncations and the inclusion of additional variables.

3.5.3 Robustness

The relationship between HLA heterozygosity and life expectancy in 1960 may reflect some underlying role of Eurasia in promoting greater health outcomes. Aside from diversity in the HLA system, Eurasian populations may contain unseen cultural or additional genetic benefits. Therefore, it is worthwhile to explore the effect of HLA heterozygosity in countries of differing concentrations of Eurasian descent.

Column (1) of Table 3.5 recreates the baseline IV estimation of column (7) in Table 3.4, while excluding countries that contain any fraction of the population derived from Eurasia. The coefficient of interest is significant at the 5% level and nearly 3 times the magnitude of the baseline estimate; although, the small sample should be taken into account.

In columns (2)-(5) of Table 3.5, the coefficient of HLA heterozygosity remains positive and significant as larger and larger fractions of Eurasian derived populations are included within the sample. The coefficient of interest, however, declines in magnitude as the fraction of a country's population derived from Eurasia increases. This implies the effect of HLA heterozygosity is not due to unseen effects of Eurasian migration, but instead, diversity within the HLA system has a stronger effect on life expectancy when Eurasian populations are excluded. In other words, it's not the high diversity of Eurasian populations, but the low diversity of non-Eurasian populations that leads to the relationship with life expectancy. This idea is carried further in column (6), in which the sample only includes countries entirely derived from Eurasia. The effect of HLA heterozygosity is shown to be negative but highly insignificant. Given the high amount of HLA genetic diversity within Eurasian populations, additional variation within the system has an insignificant statistical relationship with life expectancy in 1960.

The estimations of Table 3.5 suggest that unseen benefits of Eurasian populations are

Table 3.5: Truncation Based on Fraction of Pop. Derived from Eurasia

Dependent Variable: ln Life Expectancy in 1960						
Fraction from Eurasia:	(1) = 0%	(2) < 25%	(3) < 50%	(4) < 75%	(5) < 100%	(6) = 100%
ln *HLA Heterozygosity*	1.5788**	1.1118***	1.0827***	0.9756***	0.8061***	-1.5399
	(0.6728)	(0.2451)	(0.3208)	(0.2335)	(0.1912)	(1.5699)
ln *GDP per capita 1960*	0.0081	0.0508*	0.0530**	0.0571***	0.0561**	0.1526***
	(0.0395)	(0.0257)	(0.0255)	(0.0212)	(0.0231)	(0.0364)
ln *Genetic Dist. from USA*	-0.1152	-0.0580	-0.0801	-0.0381	-0.0076	-0.1105
	(0.2097)	(0.1208)	(0.1081)	(0.0577)	(0.0076)	(0.0770)
ln *Ethnic Fractionalization*	-0.0024	-0.0161	-0.0130	-0.0177	-0.0141	-0.0705
	(0.0159)	(0.0174)	(0.0138)	(0.0155)	(0.0133)	(0.0199)
ln *Abs. Latitude*	0.0255	0.0216	0.0150	0.0158	0.0161	0.0760
	(0.0296)	(0.0247)	(0.0213)	(0.0191)	(0.0167)	(0.0641)
Continental Dummies	Y	Y	Y	Y	Y	Y
N	29	46	56	69	81	74
R^2	0.2530	0.6515	0.6684	0.7168	0.7804	0.6152
First Stage F	48.169	17.054	11.649	37.171	39.469	5.459

Notes: IV coefficients are reported in each column. *, **, and *** represent significance at the 10, 5, and 1% significance level, respectively. Robust standard errors are in parentheses. Continental dummies include indicator variables for the Americas and Africa. The natural log of migratory distance from East Africa and its square are used as exogenous instruments.

not driving the relationship between HLA heterozygosity and pre-medicinal life expectancy. In fact, it appears that the inclusion of Eurasian populations actually reduces the effect of HLA heterozygosity on life expectancy.

Table 3.6 includes additional variables to the baseline IV estimation of column (7) in Table 3.4. Galor and Moav (2007) show a prolonged history of agriculture has a positive association with contemporary health outcomes. We argue, however, that the relationship between the Neolithic Revolution and contemporary health is due to prolonged exposure to crowd disease and resulting genetic adaptation. In other words, GM's weighted millennia of agricultural is a general measure of adaptation, whereas our measure is a direct result of adaptation. This effects of each variable are given in column (1) of Table 3.6. The coefficient of HLA heterozygosity is positive and significant at the 10% level, while the coefficient of weighted millennia of agriculture is negative and insignificant. The more direct relationship between HLA diversity and life expectancy is shown in column (1).

Column (2) includes a measure for the ease of transmission of malaria into the baseline IV estimation (Kiszewski et al. 2004). As expected, this measure of malaria ecology has a statistically significant negative effect on life expectancy in 1960. The inclusion of this variable, however, does not weaken the effect of HLA heterozygosity. The coefficient of HLA heterozygosity is significant at the 1% level and slightly larger in magnitude in comparison to the baseline estimate.

The natural selection for diversity within the HLA system may also be associated with the selection of particular traits. That is to say, the effect of HLA heterozygosity may be accounting for the frequency of favorable alleles, not diversity. Column (3) attempts to control for differences across countries by using a measure of genetic distance to a proposed crowd disease origin. This measure of genetic distance is calculated with Sewall Wright's fixation index, or F_{ST}, and uses the 156 SNP's of the HLA system. In order to measure for the selection of particular traits, we take average HLA genetic distance from a set of disease origin countries. The disease origin countries are defined by having the highest number of

Table 3.6: Additional Explanatory Variables of 1960 Life Expectancy

Dependent Variable: ln Life Expectancy in 1960						
	(1)	(2)	(3)	(4)	(5)	(6)
ln *HLA Heterozygosity*	0.5146*	0.6234***	0.9821**	0.5334**	0.5053*	1.2765***
	(0.2778)	(0.2173)	(0.3935)	(0.2506)	(0.2717)	(0.3051)
ln *Weighted Millennia of Agriculture*	-0.0285					-0.0179
	(0.0478)					(0.0521)
ln *Malaria Ecology Index*		-0.0214*				-0.0330***
		(0.0112)				(0.0110)
ln *HLA Genetic Dist. from Disease Origin*			0.0678			0.1497**
			(0.0598)			(0.0676)
Rule of Law				0.0397**		0.0341*
				(0.0187)		(0.0194)
Fraction of Pop. Derived from Eurasia					0.0235	0.0504
					(0.0556)	(0.0475)
Baseline Controls:	Y	Y	Y	Y	Y	Y
Continental Dummies	Y	Y	Y	Y	Y	Y
N	128	129	129	129	129	128
R^2	0.7674	0.7758	0.7646	0.7789	0.7667	0.7846
First Stage F	36.800	49.090	51.973	38.878	72.158	27.059

Notes: IV coefficients are reported in each column. *, **, and *** represent significance at the 10, 5, and 1% significance level, respectively. Robust standard errors are in parentheses. Continental dummies include indicator variables for the Americas and Africa. The natural log of migratory distance from East Africa and its square are used as exogenous instruments.

potential domesticate animals and the longest estimated history of agriculture (Hibbs and Olsson 2004). While controlling for HLA distance to the disease origin, the IV estimated coefficient of HLA heterozygosity is positive and significant at the 5% level. This implies that variation within, not differences between, is responsible for differing health outcomes. HLA based genetic distance has an insignificant effect on life expectancy in 1960.

Column (5) includes a measure of the rule of law (Kaufmann, Kraay, and Zoido-Lobato 2001). The measure for institutional quality is problematic in that it measures institutional quality for the 1990's, while our dependent variable is life expectancy in 1960. It should serve, however, as a proxy for the rule of law and a useful control. The coefficient of institutional quality has a positive and significant effect on 1960 life expectancy, but the inclusion of this variable does not deter the effect of HLA heterozygosity. The coefficient of HLA heterozygosity is positive, significant at the 1% level, and similar in magnitude to the baseline estimate.

As a further check against a potential bias from Eurasian populations in Table 3.5, column (6) of Table 3.6 includes the fraction of a countries population derived from Eurasia. This variable has an insignificant effect on life expectancy in 1960, while the coefficient of HLA heterozygosity remains positive and similar in magnitude to the baseline estimate, although significance falls to the 10% level.

All additional controls are included in column (7) of Table 3.6. The inclusion of all controls results in the magnitude of the coefficient of HLA heterozygosity more than doubling while significant at the 1% level. The estimate of column (7) implies a 10% increase in HLA heterozygosity is associated with roughly a 13% increase in 1960 life expectancy, or a standard deviation increase in HLA heterozygosity (an eight percent increase) is associated with a 10% increase in life expectancy. Holding constant all other relevant factors, this implies that if Mexico had the HLA diversity of the United States in 1960 life expectancy would have been 10 years greater.

3.6 Conclusion

The Neolithic Revolution radically changed the environment of early humans. As a result of the new environment, and the large populations that resulted, pathogens previously constrained to animal populations came to infect the readily available human hosts. These pathogens selected for variation within the immune system. Those societies and peoples that came into contact with these pathogens at an earlier date, as well as allowing for a sustained presence through large populations, have had a greater selection for particular traits that provide resistance to the resulting diseases. This has resulted in a contemporary variation in disease resistance across countries, which corresponds to a variation in aggregate health measures.

Galor and Moav (2007) show that an earlier agricultural transition has given a head start in adaptation to the environmental shift. This head start, in turn, is associated with contemporary variations in aggregate health outcomes. We see the current work as an intermediary to that of GM, where we attempt to more narrowly define the adaptation resulting from the Neolithic Revolution. Towards this end, we explore genetic variation within the HLA system across countries, where we propose that HLA diversity corresponds to differences in historical disease environments and a measures adaptation to these disease environments.

Chapter 4

Potatoes, Milk, and the Old World Population Boom

4.1 Introduction

An interest in historic levels of development has been recently renewed. Of particular focus has been the rapid increase in populations that has occurred in the past few centuries. This work seeks to further understand this population boom by exploring the potential complementarities of two important food sources, potatoes and milk. [1]

The introduction of the potato into the Old World has been shown to have had a large effect on populations and urbanization (Nunn and Qian 2011). Larger populations resulted from the nutritional superiority of potatoes when compared to Pre-Columbian, Old World staple crops. After Columbus's seminal voyage, the introduction of South American potatoes to Old World farmers led to an increase in caloric output of a given acre of land, which in turn, led to an increased carrying capacity, everything else constant.

In a recent work, Nunn and Qian (2011; hereafter NQ) show an exogenous determinant of potato usage, land suitability, has a significant effect upon population growth between the 18th and 19th centuries. This work seeks to supplement that of Nunn and Qian by including a unique measure for varied milk consumption. Milk complemented the high caloric output

[1] We define the "Old World" as Europe, Asia, and Africa, or as excluding Oceania and the Americas.

of potatoes by providing necessary vitamins and proteins. Our main hypothesis is echoed by the following statement in Nunn and Qian (2011, P. 601; McNeil 1999): "[A] single acre of land cultivated with *potatoes* and one *milk* cow was nutritionally sufficient for feeding a large family of six to eight.[2]" This statement implies complementarities in the diet of those families that were able to consume both potatoes and milk, where isolated consumption of either food source would be insufficient in feeding such a large family. Estimations in NQ, however, don't account for varied milk consumption. Therefore, we augment the explanatory power of potatoes by including an exogenous measure for the suitability of milk consumption, the country level frequency of lactose tolerance. We treat milk and potatoes as inputs in the production of larger populations, and in so doing, allow for measurement of the complementarity between the two food sources in the production of population.

It is shown through a number of specifications that milk consumption did have a positive and significant effect on Old World populations after the introduction of the New World potato. [3] This result is consistent when including the agricultural suitability for potatoes, as well as other relevant controls. The isolated effect of dairying becomes insignificant, however, with the inclusion of the interaction between the frequency of lactose tolerance and the agricultural suitability of potatoes. This implies the effect of milk consumption on populations is contingent on the spread of potatoes. Milk and potatoes each contain differing vital components of diet. Milk contains fats and proteins, while potatoes are made up mostly of carbohydrates. In other words, milk is a complement of potatoes in the production of larger populations, where the estimated effect of potatoes found in NQ roughly doubles as lactase persistence approaches unity.

In a similar work we outline many health and population benefits of milk consumption in the precolonial era (Cook 2011). Particularly, it is shown that the frequency of lactase

[2]Emphasis our own.

[3]As proposed in NQ, the introduction date of potatoes is 1750 CE.

persistence is strongly associated with population density in 1500 CE.[4] This relationship holds through a large number of estimations, which include sample truncations, the inclusion of a large number of theoretically relevant controls, and identification with the use of instrumental variables.

History plays a role in contemporary economic development. A number of economic studies outline this fact (see e.g., Acemoglu et al. 2001; Bockstette et al. 2002; Comin et al. 2007; Engerman and Sokoloff 2002; La Porta et al. 1998; Nunn 2008). While many of these works highlight the role of historical variables in contemporary society, our focus is strictly in the past. In the succinct words of Nunn (2009, P. 88), "The main fact ... is that history matters."

Of particular importance in the study historical economies is population. This is due to the close association between technological advancements and larger populations. This relationship was first explored by Malthus, in which he posed that wages tend to move to a subsistence level in the long run. As productivity rises within an economy, a temporary increase in the standard of living occurs. The higher income per capita allows for a larger population. If population growth isn't restrained, the population will grow until incomes again approach the subsistence level. This idea is confirmed in a more recent study; Ashraf and Galor (2011) show with the use of an exogenous instrument that improvements in technology do lead to increases in population. Before the industrial revolution, population can be seen as a proxy for the level of technological and economic development.

Population is of particular focus in the current work. NQ explore the role of potatoes on both population growth and city population share. We exclude the effect of dairying on city populations. Historically, dairying is seen as a rural activity. Our hypothesis is that dairying increased populations by aiding a diet of potatoes. Milk may not necessarily relate to a greater fraction of a country's population living within a city. This is due to how milk is consumed. Milk needs to be fresh. This implies an efficient transportation of milk from

[4]Please note that the current work considers populations, not population density.

the rural countryside to cities is necessary for consumption. Given the high costs associated with milk consumption in cities, we don't expect dairying to have led to larger city shares.

4.1.1 The Importance of Dairying

In a companion work, we show that milk is strongly associated with higher populations densities in 1500 CE, which can be seen as a proxy for pre-colonial development (Cook 2011). Two main reasons are given for this relationship.[5] First, milking provided a quantitative caloric benefit to early farmers. Second, milk provided a qualitative improvement to a farmer's diet. Of primary focus to the current work is the qualitative improvement in diet.

The secondary products revolution allowed for early farmers and pastoralists to obtain more calories from a set herd of livestock (Sherratt 1983). One major secondary product is milk, or dairying. Considering two farmers with identical heads of livestock, if one practices dairying, he is able to obtain a greater quantity of calories compared to the other. This increase in caloric production led to an increase in the carrying capacity of a given parcel of land, thereby increasing population. This idea is confirmed by the rapid natural selection of a gene variant that allows for milk consumption.

The proposed hypothesis in this paper is that milk consumption complemented the nutritionally superior potato. If two farmers adopted the potato, the one who is able to consume milk would be able to support a larger family. The reason for this relationship lies in the nutritional benefit of each food source. Milk is high in fats and proteins, while potatoes contain high amounts of calories relative to other staple crops. For a 30 year old male weighing 150 pounds a quart of milk would contain 69% of daily recommended protein, 46% of daily fat, 19% of daily carbohydrates, and 30% of daily calories (USDA 2011). [6] Additionally, a quart of milk contains 138% of daily calcium, 72% of vitamin A,

[5] A third potential reason is also discussed. Milk provided an obvious substitute to mother's milk. The availability of this substitute led to an increase in the fecundity of women by shortening the weaning period of infants, and thereby, allowing for a shorter postpartum infertility period in women. The shorter infertility period allowed mothers of lactase persistent children to have more children.

[6] The recommended amount of protein for 19-30 year old males is 0.66 grams per kilogram of adult weight.

10% of vitamin D, 150% of riboflavin, and 30% of potassium. In contrast, one large potato (300 grams; baked) contains 17% of daily recommended protein, 0.01% of daily fat, 25% of daily carbohydrates, and 14% of daily calories.[7] One 300 gram potato also contains 38% of vitamin C, 53% of iron, and 26% of folate. This implies a farmer could subsist on one quart of milk and 5 potatoes a day. Furthermore, the amino acids found within milk are a complement to those found within staple crops (WHO 2009).

Milk is an important complement to potatoes. Without milk the efficacy of potatoes in increasing populations is weakened. Potatoes provided calories, while milk provided needed protein and fat. Together milk and potatoes led to larger populations. The next sections outlines the data to be used, while section 3 confirms the role of milk and potatoes in the creation of larger populations.

4.2 Data

4.2.1 The Frequency of Lactase Persistence

Not all people are able to consume milk. The reason for varied consumption is due to differences in the presence of an enzyme in the small intestine, lactase. Lactase is responsible for breaking down lactose, a sugar found in all milks. If lactase is not present, then the lactose will pass to the colon and lead to cramping or diarrhea (Simoons 1969). As is common in the literature, we will refer to lactose tolerance as lactase persistence.

The presence of lactase in some people, and not others, is one of the most famous examples for the continued natural selection in humans (Ingram et al. 2009). The production of lactase is correlated with a particular gene variation. The particular gene variation,

The recommended daily fat intake is 65 grams. Calorie intake is based on a 2,000 calorie diet. A quart of whole milk (without added vitamin A or D) contains 595 calories, 30.74 grams of protein, and 31.92 grams of fat (USDA 2011). Additionally, it contatins 1,100 mg of calcium, 449 mg of vitamin A, 1 mg of vitamin D, 1.65 mg of riboflavin, and 1,288 mg of potasium (USDA 2011)

[7]An unsalted 300 gram baked potato contains 7.47 grams of protein, 0.39 grams of fat, 63.24 grams of carbohydrates, and 278 calories. Additionally, a potato contains 28.7 mg of vitamin C, 3.23 mg of iron, and 84 mg of folate

however, differs across differing populations (Ingram et al. 2009; Tishkoff et al. 2006). Therefore, we base our measure of lactase persistence on the observed ability to consume lactose, not on latent genetic variation. Specifically, the data for frequencies of lactase persistence come from Ingram et al. (2009). The authors collect lactase persistence frequencies for a wide number of ethnic groups from previous studies.[8] The data present two problems: First, data are on the ethnic level, while we are interested in explaining the country level population. Second, the collection dates for the frequency of lactase persistence range from roughly 1960 to 1990; this may not be relevant to explaining population changes of the 18th and 19th century.

In order to aggregate ethnic lactase persistence data to the country level we use ethnic compositions from Alesina et al. (2003).[9] The ethnic data of Alesina et al. (2003), however, give compositions for roughly the mid 1990's. In order to find ethnic compositions for the pre-industrial period, we post multiply the ethnic compositions of Alesina et al. (2003) by the migration matrix of Putterman and Weil (2010). This, in theory, removes all contemporary ethnic groups that are attributed to migration and gives ethnic compositions for the year 1500 CE.[10] This gives a viable proxy for ethnic compositions in the pre-industrial era.

The fact that data on lactase persistence frequencies are collected roughly 200 years after the explanatory variable shouldn't lead to bias. Lactase persistence, and therefore its country wide frequency, is determined by a gene variant. This implies that differences in the ability to consume milk are due to differences in the genome. Variation in the genome is determined by two primary forces: genetic drift and natural selection.[11] Both of which are relatively slow to act, and given that these changes must ocurr over 8 generations, it is

[8]The sampling techniques for collecting lactase persistence frequencies are consistent in all past studies. For collection, either blood glucose or breath hydrogen tests are conducted. For a full discussion see Cook (2011).

[9]Ethnic groups from Ingram et al. (2009) are not a perfect match to ethnic groups in Alesina et al. (2003). The use of ethnic language classifications, however, allows for the matching of related groups.

[10]For a full discussion of this idea, please see Cook (forthcoming).

[11]Genetic drift results from the random changing of gene frequencies due to isolated populations. Natural selection implies the selection of certain traits due to survival advantages of these traits.

highly improbable that the frequency of lactase persistence has been non-monotonic across ethnicities over this period (Hartl and Clark 2006).[12]

The resulting country wide data for lactase persistence gives the percent of a country's population that is able to consume milk. In essence, this is similar to the agricultural suitability measure employed by NQ. The frequency of lactase persistence is an inherent difference that gives the *suitability* for milk consumption for a given country. This inherent measure of milk suitability serves the same role as the agricultural suitability of potatoes in NQ. Data are available for 109 Old World countries. Figure 4.1 plots the relationship of population growth between the 18th and 19th centuries and the frequency of lactase persistence.

4.2.2 Agricultural Suitability

Data for crop suitability come from the Food and Agriculture Organization (FAO)'s Global Agro-Ecolological Zones. The FAO's data set takes into account many environmental factors to determine individual crop yields. The environmental factors come from the Climate Research Unit and include precipitation, frequency of wet days,mean temperature, diurnal temperature range, vapor pressure, cloud cover, sunshine, ground-frost frequency, and wind speed. The FAO data also take into account soil and slope conditions. Potential crop yields are then calculated for within country grids. The grids are 0.5 degrees latitude and 0.5 degrees longitude and span the globe. Suitability is determined by classifying grid yields relative to the maximum yield; e.g., a grid is said to be very suitable if it can produce 80-100% of the maximum possible output.[13]

NQ define land to be suitable for cultivation if it can produce 40% of the maximum output. With this understanding, NQ define country-level crop suitability by the fraction of suitable cells within a country. In addition to the suitability of potatoes, agricultural

[12] A human generation is proposed to be 25 years (Ingram et al. 2009)

[13] Maximum output is obtained from the best possible conditions.

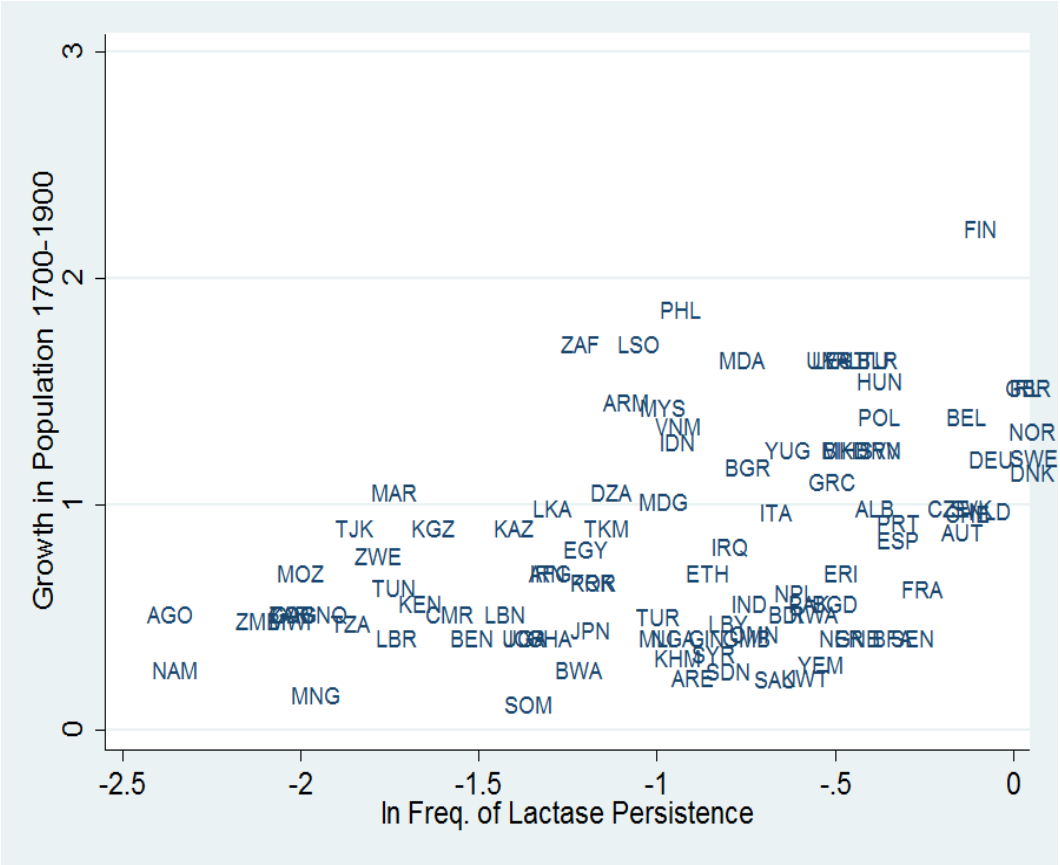

Figure 4.1: The Freq. of Lactase Persistence
and Population Growth between 1700 and 1900 CE

suitability measures are also found for sweet potatoes, silage maize, grain maize, cassava, wetland rice, dryland rice, wheat, and a measure for the suitability of all crops. This gives data for all Old World countries.

A few concerns arise in using crop specific agricultural suitability. First, the measure are obtained in the 1990's, roughly 200 years after the population boom. The suitability measures, however, are based on climatic conditions which have changed little over the last 200 years. Second, a number of differing potato varieties have evolved over time. If farmers were manipulating potatoes to grow in high population areas, this can lead to the proposed relationhsip exhibited in NQ. This brings into question the causative properties of potatoes on population growth. The manipulation of potato varieties, however, has been shown to be mostly for visual reasons, not to supplement high population areas. According to NQ (P. 613), "To the best of our knowledge, the focus was not on developing varieties that could be grown in climates with rapid population growth." Crop suitability measures are an exogenous determinant of crop yields across countries.

4.2.3 Other Variables

The primary dependent variable considered in the current work is country level population between 1000 and 1900 CE. These data come from McEvedy and Jones (1978), which are widely used in related work (see e.g., Acemoglu et al. 2002; Cook *forthcoming*; Putterman 2008). Increases in productivity prior to the Industrial Revolution led to short run improvements in living conditions and long run increases in populations (Ashraf and Galor 2011; Malthus 1798). Therefore, the use of populations is seen as a proxy for levels of development.

Other variables to be used include elevation, ruggedness, and the percent of a country within the tropics. These variables are intended to control for addtional environmental effects that influence agricultural productivity and populations. Elevation and the percent of a country within the tropics come from the Center of International Development at

Harvard. While ruggedness is from Nunn and Puga (2010).

Table 3.1 gives summary statistics for the frequency of lactase persistence, the natural log for potato and other crop suitability, and other controls. Note that the number of Old World countries drops from 130 used in Nunn and Qian to 109; this is due to a lack of data for lactase persistence.[14] The truncation does not significantly alter the sampled mean.

Table 4.1: Summary Statistics

Variable:	N	Mean	Std. Dev.	Min	Max
Freq. of Lactase Persistence	109	0.4155	0.2424	0.0233	0.96
ln *Potato Area*	109	4.0814	3.5209	0	10.4904
ln *Old World Crops Area*	109	7.378	2.4358	0	11.0763
ln *Elevation*	109	6.023	1.0431	2.3228	7.9984
ln *Ruggedness*	109	-0.1731	1.0244	-3.3104	1.8249
ln *Tropical Area*	109	3.6894	4.7774	0	12.3081
ln *All Crops Area*	109	7.8927	2.4811	0	12.0043
ln *Maize Area*	109	5.4567	3.4699	0	11.6632
ln *Silage Maize*	109	3.5499	3.5921	0	11.1043
ln *Sweet Potato Area*	109	3.3483	4.0601	0	10.992
ln *Cassava*	109	2.8917	3.9943	0	11.0466

4.3 Results

The primary estimations to be discussed are a corrollary to those found in NQ. Towards this end, we will first show that the effect of milk and potato consumption is significant after a proposed date for the introduction of potatoes to the Old World. Secondly, we show suitability measures for potatoes and milk are associated with greater populations after this date. In essence, this is a difference-in-difference estimation, in which the suitability measure are continuous measure of receiving the treatment, potatoes. The introduction of potatoes to the Old World is the treatment to be considered. Historically, milk has

[14]The excluded countries are Australia, Bhutan, Central African Republic, Chad, Cote d'Ivoire, Djibouti, Fiji, Iceland, Israel, Mauritania, New Zealand, Qatar, Romania, Russia, Sierra Leone, Solomon Islands, Swaziland, Timor-Leste, Togo, and Western Sahara.

been present throughout the Old World. The goal of estimation is not to measure the sole benefits of milk consumption, but rather, to measure the complementarity with the more productive staple crop of potatoes. It is therefore reasonable to measure the effects of milk *after* the introduction of potatoes.

4.3.1 Flexible Estimation

The purpose of this section is to establish the adoption date of potatoes. Our hypothesis is that milk consumption, measured by the frequency of lactase persistence, complemented the introduction of the potato. Therefore, varied milk consumption should affect populations *after* the introduction of the potato. This is measured by the joint significance of the coefficient of the frequency of lactase persistence for within year estimations. As in NQ, this is defined as flexible estimation, and is given by the following equation:

$$[l]P_{it} = \sum_{j=1100}^{1850} \beta_j^M \ln Freq.of Lactase Persistence_i \cdot I_t^j + \sum_{j=1100}^{1850} \beta_j^P \ln PotatoArea_i \cdot I_t^j + \sum_{j=1100}^{1850} \mathbf{X'}_i \mathbf{I}_t^j \phi_j + \sum_c \gamma_c I_i^c + \sum_{j=1100}^{1850} \rho_j I_j^t + \epsilon_{it} \tag{4.1}$$

where i is a country indicator, t is an indicator for the year, and P_{it} is the population of country i in year$t$. The coefficient of interest is β_j^M, which represents the effect of dairying in the jth year. Our primary hypothesis states that milk aided in the population boom after the introduction of potatoes. Therefore, we expect β_j^M to be positive and significant for $j > 1750$, where as in NQ, we use 1750 as the introduction date of potatoes. β_j^P's represent the within year effect of potatoes, $\sum_{j=1100}^{1850} \mathbf{X'}_i \mathbf{I}_t^j \phi_j$ represents year specific controls, country fixed effects are given by $\sum_c \gamma_c I_i^c$, and year fixed effects are given by $\sum_{j=1100}^{1850} \rho_j I_j^t$.

We perform these estimations with two specifications, focusing solely on the coefficient of the frequency of lactase persistence. First, we estimate year specific coefficients of the frequency of lactase persistence while excluding the suitability of potatoes and including

controls from NQ; this is shown by excluding $\sum_{j=1100}^{1850} \beta_j \ln PotatoArea_i \dot{I}_t^j$ from equation (1). Estimates are given in Table 4.2. Second, we again perform flexible estimations, but now include both the suitability of potatoes and the frequency of lactase persistence, as well as the baseline controls. We suggest that milk supplemented the population effects of potatoes; therefore, dairying should have a significante effect on top of the effect of potatoes. Estimates of equation (1) are found in Table 4.3[15]

Table 4.2 displays the within year effects of dairying for 1100-1850 CE. The estimates of Table 4.2 exclude potato suitability from the flexible estimation. Again, the purpose of flexible estimation is to show dairying had a significant effect *after* the introduction of the potato, which is historically estimated to be 1750 CE. Column (1) gives flexible estimates while excluding all controls. The coefficient of lactase persistence is significant at the 1% level for all years after the proposed introduction of potatoes. After 1750, the joint significance of the coefficients of lactase persistence is significant at the 1% level; this is shown by the F statistic of 13.42.[16] Furthermore, the point estimate of the coefficient is larger at later dates. As proposed by NQ, the larger coefficients represent a more widespread diffusion of the potato. Column (1) supports our hypothesis: Dairying complemented the potato in the production of larger populations. This is corroborated by the frequency of lactase persistence having a significant effect on populations after the introduction of potatoes, and a mostly insignificant effect before potatoes.

Columns (2) and (3) include additional variables related to population. An additional measure for agricultural suitability is included in column (2). It is possible that the complementarity associated with dairying may be the result of a more fundemental determinant of Old World populations, the suitability of Old World crops. Potatoes supplanted many of the staple crops of the Old World. Milk complements all staple crops; however, the marginal

[15]The coefficient of the flexible estimates of the frequency of lactase persistence are displayed.

[16]The coefficient of lactase persistence becomes signficant after 1700, the period before the proposed introduction date. Given that the introduction date varied across the Old World, it isn't a surprise the coefficient is significant in this period.

Table 4.2: Flexible Estimates: Excluding Potato Suitability

Dependent Variable: Population			
	(1)	(2)	(3)
ln *Freq. of Lactase Persistence* ×1100	0.0175	0.0237	0.0109
	(0.0495)	(0.0461)	(0.0548)
ln *Freq. of Lactase Persistence* ×1200	0.0652	0.0788*	0.0409
	(0.0495)	(0.0461)	(0.0548)
ln *Freq. of Lactase Persistence* ×1300	0.1331***	0.1541***	0.1287**
	(0.0495)	(0.0461)	(0.0548)
ln *Freq. of Lactase Persistence* ×1400	0.0093	0.0285	0.0582
	(0.0494)	(0.0459)	(0.0547)
ln *Freq. of Lactase Persistence* ×1500	0.0383	0.0656	0.0835
	(0.0494)	(0.0459)	(0.0547)
ln *Freq. of Lactase Persistence* ×1600	0.0865*	0.1156**	0.1018*
	(0.0494)	(0.0459)	(0.0547)
ln *Freq. of Lactase Persistence* ×1700	0.1302***	0.1636***	0.1467***
	(0.0494)	(0.0459)	(0.0547)
ln *Freq. of Lactase Persistence* ×1750	0.1661***	0.2034***	0.1708***
	(0.0494)	(0.0459)	(0.0547)
ln *Freq. of Lactase Persistence* ×1800	0.2328***	0.2763***	0.2152***
	(0.0494)	(0.0459)	(0.0547)
ln *Freq. of Lactase Persistence* ×1850	0.2836***	0.3321***	0.2236***
	(0.0494)	(0.0459)	(0.0547)
ln *Freq. of Lactase Persistence* ×1900	0.3279***	0.3830***	0.2220***
	(0.0494)	(0.0459)	(0.0547)
Baseline Controls (× Year fixed effects):			
ln *Potato-suitable Area*	N	N	N
ln *Old World Crops*	N	Y	Y
ln *Elevation*	N	N	Y
ln *Ruggedness*	N	N	Y
ln *Tropical Area*	N	N	Y
N	1304	1304	1304
R^2	0.7928	0.8248	0.8399
F-stat of Joint Significance for LPF[a] 1750-1900	13.42	21.09	6.08

Standard errors in parentheses. All estimations include country and year fixed effects, and all baseline controls are interacted with year indicators.
a. LPF = Lactase Persistence Frequency

effect of potatoes should be larger with the more productive staple of potatoes. Differences in the production of Old World crops can lead to differences in the efficaciousness of dairying. Therefore, it is necessary to control for variations in Old Crops. The yearly coefficients of lactase persistence again become highly significant after 1750 and the point estimates increase in later years. The effects of lactase persistence are unaffected by the inclusion of Old World crops. Column (3) includes all baseline controls from NQ. These include Old World crop suitability, mean elevation, mean ruggedness, and the fraction of a country within the tropics. The inclusion of all controls does not influence the significance of lactase persistence after 1750. At all dates beyond 1750, the coefficient of the frequency of lactase persistence is significant at the 1% level, while in years before 1700 CE, the coefficient of interest is mostly insignificant.

Table 4.2 shows that the frequency of lactase persistence did play a role in the population boom of the 19th and 20th century. In order to isolate the effect of dairying, we need to include potato suitability into the flexible estimation. This is given by estimating equation (1) with results displayed in Table 4.3.

Column (1) of Table 4.3 includes within year estimations for both the frequency of lactase persistence and the suitability of potatoes. [17] With the inclusion of potato suitability, the estimated effect of lactase persistence becomes significant in 1750 CE and remains significant for all future periods. In addition to the effect of dairying turning on after the introduction of potatoes, the point estimate of the coefficient of lactase persistence is growing over time. This is indicative of a greater diffusion of the potato, and therefore, a more complementary role for dairying. Although the point estimates are suppressed, potatoes have a large effect on populations after the proposed introduction date; this is shown by the test of joint significance. The estimates of column (1) corroborate the hypothesis that both dairying and potato consumption led to the population boom of the 19th and 20th century.

As in Table 4.2, columns (2) and (3) introduce the suitability for Old World crops and all

[17] The estimates of the coefficient of potato suitability are suppressed.

Table 4.3: Flexible Estimates: Including Potato Suitability

Dependent Variable: Population			
	(1)	(2)	(3)
ln *Freq. of Lactase Persistence* ×1100	-0.0012	0.0060	-0.0023
	(0.0489)	(0.0496)	(0.0555)
ln *Freq. of Lactase Persistence* ×1200	0.0252	0.0419	0.0154
	(0.0489)	(0.0496)	(0.0555)
ln *Freq. of Lactase Persistence* ×1300	0.0830*	0.1171**	0.0996*
	(0.0489)	(0.0496)	(0.0555)
ln *Freq. of Lactase Persistence* ×1400	-0.0194	0.0216	0.0353
	(0.0488)	(0.0494)	(0.0555)
ln *Freq. of Lactase Persistence* ×1500	-0.0073	0.0495	0.0526
	(0.0488)	(0.0494)	(0.0555)
ln *Freq. of Lactase Persistence* ×1600	0.0338	0.0919*	0.0698
	(0.0488)	(0.0494)	(0.0555)
ln *Freq. of Lactase Persistence* ×1700	0.0802	0.1554***	0.1155**
	(0.0488)	(0.0494)	(0.0555)
ln *Freq. of Lactase Persistence* ×1750	0.1007**	0.1786***	0.1335**
	(0.0488)	(0.0494)	(0.0555)
ln *Freq. of Lactase Persistence* ×1800	0.1465***	0.2306***	0.1705***
	(0.0488)	(0.0494)	(0.0555)
ln *Freq. of Lactase Persistence* ×1850	0.1754***	0.2613***	0.1700***
	(0.0488)	(0.0494)	(0.0555)
ln *Freq. of Lactase Persistence* ×1900	0.1970***	0.2897***	0.1620***
	(0.0488)	(0.0494)	(0.0555)
Baseline Controls (× Year fixed effects):			
ln *Potato-suitable Area*	Y	Y	Y
ln *Old World Crops*	N	Y	Y
ln *Elevation*	N	N	Y
ln *Ruggedness*	N	N	Y
ln *Tropical Area*	N	N	Y
N	1304	1304	1304
R^2	0.8166	0.8310	0.8452
F-stat of Joint Significance for LPF[a] 1750-1900	5.12	10.79	3.43
F-stat of Joint Significance for PSA[b] 1750-1900	25.14	6.91	7.16

Standard errors in parentheses. All estimations include country and year fixed effects, and all baseline controls are interacted with year indicators.
a. LPF = Lactase Persistence Frequency
b. PSA = Potato-suitable Area

baseline controls, respectively. The inclusion of Old World crop suitability does not weaken the coefficient of lactase persistence after 1750. The coefficient of the frequency of lactase persistence remains highly significant after the introduction of the potato and is growing over time. Additionly, dairying has a mostly insignificant effect on populations before 1700 CE. Column (3) includes all controls as well as yearly potato suitability estimates. The frequency of lactase persistence remains significant after the introduction of the potato. However, the point estimate of the coefficient no longer becomes greater over time.

Tables 4.2 and 4.3 perform within year estimations to show that dairying did have a significant effect on populations after the introduction of the potato. This relationship holds whether or not we control for potato suitability and other relevant determinants of population variations. The baseline estimates of NQ are similar to a difference in difference approach. The rationale is based on the effect of potatoes before and after introduction from the New World. In the next section, we will take this same approach in regards to dairying. The flexible estimates of Tables 4.2 and 4.3 give creedence to this approach, where it is shown that dairying too turns on after the introduction of the potato.

4.3.2 Baseline Estimation: Complementarity between Milk and Potatoes

The flexible estimations give support to exploring the effect of both milk and potatoes after potatoes were introduced to the Old World. For the baseline estimation, our primary focus is on the relationship of milk and potatoes after 1750. This is done by interacting both the frequency of lactase persistence and the suitability of potatoes with an indicator variable for the post-introduction period. This is given by the following equation:

$$\begin{aligned}[l]P_{it} = \quad & \beta^M \ln Freq.of Lactase Persistence_i \cdot I_t^{Post} + \beta^P \ln PotatoArea_i \cdot I_t^{Post} \\ & + \textstyle\sum_{j=1100}^{1850} \mathbf{X'}_i \mathbf{I}_t^j \phi_j + \sum_c \gamma_c I_i^c + \sum_{j=1100}^{1850} \rho_j I_j^t + \epsilon_{it}\end{aligned} \tag{4.2}$$

where I_t^{Post} is the post 1750 indicator and all other variables are identical to those found in equation 1. β^M and β^P give the respective effect of milk and potatoes after introduction. The estimates of β^M and β^P can be seen as exogenous suitability measures that are able to capture the benefits of the potato. Potato suitability measures the ease in which some countries were fully able to adapt the more substantial staple crop, while the frequency of lactase persistence captures the additional health benefits from the complementarity of milk and potatoes.

Table 4.4 recreates the estimates of Table IV from NQ. This corresponds to columns (1)-(5) in Table IV of NQ. The purpose of recreating estimates from NQ is due to sample mismatch with the inclusion of the frequency of lactase persistence (N = 1304 in the current work and 1552 in NQ). With the lower sample, the point estimates are insignificantly different than those found with the larger sample of NQ. Column (1) estimates the effect of potato suitability on populations after the introduction of the potato while including no controls. From column (1), a 1% increase in suitability for potatoes leads to an increase of 0.052% increase in populations; this relationship is significant at the 1% level. In regards to column (1), column (2) includes Old World crop suitability, column (3) includes all baseline controls, column (4) includes the baseline controls, excluding Old World crop suitability, and an agricultural suitability measure for all crops, and column (5) includes all baseline controls as well as agricultural suitability for other New World crops. In all estimations potato suitability has a positive and significant effect on populations after the introduction date. The use of the lower sample in this work produces roughly identical estimates of those in NQ.

We will now expand on the results found in Table 4.4 by including the frequency of lactase persistence after the introduction of potatoes. The equation to be estimated in Table 4.5 is given by equation 2. The introduction of the potato should have been more beneficial to those countries that have a history of dairying. In other words, those countries which were able to digest milk should have gained additional population benefits from the

Table 4.4: Baseline Estimation: Recreating Table IV of Nunn and Qian (2011)

Dependent Variable: Population: 1100-1850 CE					
	(1)	(2)	(3)	(4)	(5)
ln *Potato Area* × Post	0.0523***	0.0408***	0.0353***	0.0346***	0.0413***
	(0.0044)	(0.0049)	(0.0060)	(0.0056)	(0.0062)
Baseline Controls (× Year fixed effects):					
ln *Old World Crops*	N	Y	Y	N	Y
ln *Elevation*	N	N	Y	Y	Y
ln *Ruggedness*	N	N	Y	Y	Y
ln *Tropical Area*	N	N	Y	Y	Y
Other Controls (× Year fixed effects):					
ln *All Crops Area*	N	N	N	Y	N
ln *Maize Area*	N	N	N	N	Y
ln *Silage Maize Area*	N	N	N	N	Y
ln *Sweet Potato Area*	N	N	N	N	Y
ln *Cassava Area*	N	N	N	N	Y
N	1304	1304	1304	1304	1304
R^2	0.7979	0.8114	0.8380	0.8394	0.8526

Standard errors in parentheses. All estimations include country and year fixed effects, and all controls are interacted with year indicators.

adoption of potatoes. This idea is tested in Table 4.5, which includes both the suitability for potatoes and the suitability of milk consumption—measured by the frequency of lactase persistence—after the introduction of the potato from the New World.

Column (1) includes the area of potato suitability and the frequency of lactase persistence both interacted with the post adoption indicator. The coefficients of each variable are highly significant with a 1% increase in potato suitability leading to a 0.04% increase in populations and a 1% increase in the frequency of lactase persistence leading to a 0.13% increase in populations.[18] The positive and significant coefficients of two variables of interest corroborates our hypothesis: Both potatoes and milk played a role in the population boom of the 18th and 19th centuries.

Columns (2)-(5) of Table 4.5 include geographic and other crop suitability measures. In all specifications, the coefficients of potato suitability and the frequency of lactase persistence are significant at the 1% level and are similar in magnitude to the estimates of column (1). Column (2) controls for Old World crop suitability. The inclusion of this measure is intended to control for an inherent advantage in agriculture before the introduction of the potato. Additionally, the effect of dairying may be driven by a complementarity to Old World staple crops. The inclusion of Old World crop suitability to the estimation of column (1) leads to a slight reduction in the point estimate of the coefficient of potato area and a slight increase in the point estimate of the coefficient of the frequency of lactase persistence. Neither coefficient change results in an altered statistical significance. Column (3) includes Old World crop suitability as well as other geographic controls. The inclusion of these measures leads to a reduction in the measured effect of dairying, while increasing the estimated effect of potatoes. Column (4) replaces Old World crop suitability in column (3) with suitability for all crops. This leads to trivial difference in the size of the coefficients but does not alter significance. Column (5) includes geographic and Old World and New World crop suitability measures. Again, the coefficients of potatoes and milk remain

[18]Please note I'm not talking about a percentage point increase in the frequency of lactase persistence.

Table 4.5: Baseline Estimation: Including the Frequency of Lactase Persistence

Dependent Variable: Population: 1100-1850 CE					
	(1)	(2)	(3)	(4)	(5)
ln *Potato Area* × Post	0.0443***	0.0238***	0.0291***	0.0300***	0.0358***
	(0.0045)	(0.0052)	(0.0061)	(0.0057)	(0.0063)
ln *Freq. of Lactase Persistence* × Post	0.1304***	0.1795***	0.1106***	0.1035***	0.1051***
	(0.0216)	(0.0218)	(0.0241)	(0.0239)	(0.0235)
Baseline Controls (× Year fixed effects):					
ln *Old World Crops*	N	Y	Y	N	Y
ln *Elevation*	N	N	Y	Y	Y
ln *Ruggedness*	N	N	Y	Y	Y
ln *Tropical Area*	N	N	Y	Y	Y
Other Controls (× Year fixed effects):					
ln *All Crops Area*	N	N	N	Y	N
ln *Maize Area*	N	N	N	N	Y
ln *Silage Maize Area*	N	N	N	N	Y
ln *Sweet Potato Area*	N	N	N	N	Y
ln *Cassava Area*	N	N	N	N	Y
N	1304	1304	1304	1304	1304
R^2	0.8040	0.8217	0.8410	0.8420	0.8552
F	372.8545	224.9309	105.5836	106.3831	63.9901

Standard errors in parentheses. All estimations include country and year fixed effects, and all controls are interacted with year indicators.

positive and significant at the 1% level and differ slightly in magnitude from the estimates without controls.

The estimates of Table 4.5 again confirm that milk consumption did play some role in the acquisition of larger populations after the introduction of the potato. Our hypothesis, however, is that the role of milk is a complement to potatoes. In other words, the population benefits of milk consumption seen in Table 4.5 are dependent upon potato consumption. This relationship is explored in Table 4.6.

Table 4.6 includes an interaction of potato suitability and the frequency of lactase persistence after the introduction of the potato. The inclusion of the interaction term causes the coefficient of lactase persistence to become insignificant. The coefficient on the interaction term, however, is highly significant in all specifications. With the inclusion of the interaction term, potato suitability remains a significant explanatory variable for population. These results imply that dairying alone is an insufficient explanatory variable. However, dairying and potato suitability explain a significant portion of population growth after the introduction of the potato. This corroborates our hypothesis of dairying complementing potato consumption in the role of increased populations. The effect of dairying is tied to potato suitability.

Post adoption estimates of the effect of potato suitability, the frequency of lactase persistence, and the interaction between the two variables of interest are found in column (1) of Table 4.6. The marginal effect of dairying, evaluated at the mean of the natural log of potato suitability, indicates that a 1% increase in the frequency of lactase persistence corresponds with an increase in populations of 0.14% after the introduction of the potatoes. This is similar to the estimate of column (1) in Table 4.5, or the effect of lactase persistence when excluding the interaction. The marginal effect of potato suitability, evaluated at the mean of the natural log of the frequency of lactase persistence, is also similar to the estimate found in Table 4.5.

The complementarity between potatoes and milk can be seen in the interaction term.

Table 4.6: Baseline Estimation: Interaction of Milk and Potatoes

Dependent Variable: Population: 1100-1850 CE					
	(1)	(2)	(3)	(4)	(5)
ln *Potato Area* × Post	0.0830***	0.0599***	0.0841***	0.0803***	0.0838***
	(0.0076)	(0.0081)	(0.0106)	(0.0096)	(0.0115)
ln *Freq. of Lactase Persistence* × Post	-0.0100	0.0506	-0.0169	-0.0170	-0.0024
	(0.0310)	(0.0311)	(0.0312)	(0.0301)	(0.0318)
ln *Potato Area* × ln *Freq. of Lactase Persistence* × Post	0.0355***	0.0315***	0.0386***	0.0379***	0.0331***
	(0.0057)	(0.0055)	(0.0062)	(0.0059)	(0.0067)
Baseline Controls (× Year fixed effects):					
ln *Old World Crops*	N	Y	Y	N	Y
ln *Elevation*	N	N	Y	Y	Y
ln *Ruggedness*	N	N	Y	Y	Y
ln *Tropical Area*	N	N	Y	Y	Y
Other Controls (× Year fixed effects):					
ln *All Crops Area*	N	N	N	Y	N
ln *Maize Area*	N	N	N	N	Y
ln *Silage Maize Area*	N	N	N	N	Y
ln *Sweet Potato Area*	N	N	N	N	Y
ln *Cassava Area*	N	N	N	N	Y
N	1304	1304	1304	1304	1304
R^2	0.8101	0.8266	0.8463	0.8475	0.8584
F	359.9711	223.0699	107.9378	108.9365	64.9786

Standard errors in parentheses. All estimations include country and year fixed effects, and all controls are interacted with year indicators.

As the frequency of lactase persistence approaches unity, the interaction term approaches zero. More importantly, the marginal effect of potatoes increases. The frequency of lactase persistence is between 0 and 1, implying the natural log is negative. As this frequency increases, the reduction in the marginal effect of potatoes increases. When evaluated at the mean frequency of lactase persistence the, marginal effect of a 1% increase potato suitability in column (1) is a 0.044% increase in population.[19] This is identical to the coefficient estimate of column (1) in Table 4.5. If, however, we assume the maximum frequency of lactase persistence, the marginal effect of a 1% increase in potato suitability leads to a 0.082% increase in population after the introduction of the potato. The higher frequency of lactase persistence leads to a greater effect of potatoes.

Of prime importance to Table 4.6 is that the effect of the frequency of lactase persistence is dependent on the log of the suitability of potatoes not being zero. As potato suitability approaches zero, the marginal effect of milk consumption on post-1750 populations also approaches zero. Dairying alone is unable to explain larger populations. Dairying is merely a complement of potatoes in the production of populations.

4.3.3 Identification

As mentioned in Cook (2011), the development of lactase persistence may not be completely exogenous. The selection of lactase persistence is dependent upon the cultural practice of dairying, implying lactase persistence itself may not be the cause of larger populations. Rather, it is the byproduct of culture. To address this issue, we use estimate the effect of lactase persistence with the use of an environmentally given exogenous instrument, solar radiation.

Solar radiation serves as an instrument through the added advantage of milk in low sunlight areas. Sunlight, or solar radiation, is responsible for synthesizing vitamin D within the body. In areas with low levels of solar radiation, the body is unable to produce adequate

[19]The mean frequency of lactase persistence is roughly 42%.

amounts of vitamin D, which may lead to the disease rickets. Rickets results in the softening of bones, which can be offset with a diet heavy in calcium rich milk (Flatz and Rotthauwe 1973). Therefore, the ability to consume milk has an added advantage in areas with low levels of solar radiation, resulting in a negative relationship between the frequency of lactase persistence and solar radiation. This relationship is exogenous to any cultural benefits associated with milk consumption, and therefore serves as a valid way to measure a clear complementarity between the ability to consume milk and the agricultural suitability of potatoes.

The measure of solar radiation comes from the Atmospheric Science Data Center of NASA and constitutes a 22 year average of solar radiation for a horizontal surface at the representative lattitutde and longitude of a country, given by the CIA World Factbook (2011). As shown in Figure 2.5, the relationship between solar radiation and the frequency of lactase persistence appears to be non-linear.[20] Solar radiation and its square are the primary instruments used for the frequency of lactase persistence, while solar radiation interacted with potato suitability and the square of solar radiation interacted with potato suitability are used as instruments for the variable used to measure the complementarity between milk and potatoes.

Table 4.7 gives the baseline IV estimates, which repeats the OLS estimations of Table 4.6. The complementarity between potatoes and milk, measured by $\frac{\partial^2 Population^{1100-1850}}{\partial PSA \partial LPF}$, remains positive, implying the complementarity, and significant at the 1% level in all estimations. The IV estimates, however, show a significant direct effect of milk consumption, a finding not shown in the OLS estimates of Table 4.6. In particular, the estimated effect of a standard deviation increase in the frequency of lactase persistence evaluated at the mean agricultural suitability of potatoes in column (1) of Table 4.7 is associated with a 0.1% increase in population between 1100 CE and 1850 CE. This is a modest effect; the remaining columns of Table 4.7 estimate similar marginal effects of added lactase persistence to the growth in

[20]This corresponds to a critical level of solar radiation for advantageous milk consumption.

Table 4.7: Complementarity: IV Estimates

Dependent Variable: Population: 1100-1850 CE					
	(1)	(2)	(3)	(4)	(5)
ln *Potato Area* × Post	0.1822***	0.1654***	0.2501***	0.2028***	0.2987***
	(0.0184)	(0.0211)	(0.0250)	(0.0191)	(0.0344)
ln *Freq. of Lactase Persistence* × Post	-0.4801***	-0.4149***	-0.5349***	-0.4011***	-0.6982***
	(0.0995)	(0.1058)	(0.1018)	(0.0843)	(0.1263)
ln *Potato Area* × ln *Freq. of Lactase Persistence* × Post	0.1185***	0.1124***	0.1490***	0.1268***	0.1723***
	(0.0142)	(0.0143)	(0.0154)	(0.0127)	(0.0215)
Baseline Controls (× Year fixed effects):					
ln *Old World Crops*	N	Y	Y	N	Y
ln *Elevation*	N	N	Y	Y	Y
ln *Ruggedness*	N	N	Y	Y	Y
ln *Tropical Area*	N	N	Y	Y	Y
Other Controls (× Year fixed effects):					
ln *All Crops Area*	N	N	N	Y	N
ln *Maize Area*	N	N	N	N	Y
ln *Silage Maize Area*	N	N	N	N	Y
ln *Sweet Potato Area*	N	N	N	N	Y
ln *Cassava Area*	N	N	N	N	Y
N	1304	1304	1304	1304	1304
R^2	0.7691	0.7889	0.7972	0.8147	0.7877
First Stage F-Statistic	63.04	34.4	40.197	51.849	28.504

Standard errors in parentheses. All estimations include country and year fixed effects, and all controls are interacted with year indicators. The frequency of lactase persistence after the introduction of the potato and the interaction of this variable with the agricultural suitability of potatoes are treated as endogenous. A proposed exogenous instrument is solar radiation and its square. Instrumental variables estimation takes use of this instrument and an interaction with the agricultural suitability of potatoes.

population between 1100 and 1850 CE.

Of more importance to this study is the estimated effect of potato suitability when controlling for the complementarity associated with milk consumption. The estimates of column (1) suggest that a 1% increase in potato suitability evaluated at the mean lactase persistence frequency is associated with an 0.08% increase in Old World populations over the specified period. This is nearly double the OLS estimate of column (1) in Table 4.6. As milk consumption approaches unity, the effect of a 1% increase in potato suitability leads to a 0.18% increase in Old World populations; an estimated effect more than double the OLS estimate. The larger marginal effect of the IV estimates evaluated at the mean frequency of lactase persistence is seen in the remaining estimations of Table 4.7, from which a 1% increase in potato suitability leads to a 0.06% increase in Old World populations in column (2), a 0.11% increase in population in column (3), a 0.09% increase in column (4), and a 0.15% increase in Old World populations when all Old World and New World crop controls are included.

The IV specification estimates a larger direct effect of the agricultural suitability of potatoes, a larger complementarity between this agricultural suitability and the frequency of lactase persistence, as well as a direct positive, significant effect of the ability to consume milk. These effects may be attributed to a correction of measurement error in the effect of milk consumption. The estimated complementarity between milk consumption and potato suitability, the focus of this work, is highly significant. This significance remains while controlling for baseline factors of Old World population growth and controls for additional crop suitability measures. The IV estimations give further credence to the main hypothesis of this work: The introduction of the potato to the Old World had divergent effects on population growth due to differences in the ability to consume milk; milk provided an important complementarity within the diet that allowed the adoption of the potato to have a greater effect on population growth between the 18th and 19th centuries.

4.4 Conclusion

The effect of potatoes on Old World populations is well documented (Nunn and Qian 2011). Potatoes provided nutritional advantages to the staple crops of the Old World, thereby raising the carrying capacity of potato suitable countries. This paper questions whether the effect of potatoes was augmented by the ability to consume milk—measured by the country level frequency of lactase persistence, or lactose tolerance. Milk provided necessary nutrients not found within potatoes. In other words, milk consumption complemented the nutritionally superior potato.

We show that the frequency of lactase persistence has a positive effect on populations after the introduction of the potato to the Old World. And this positive effect is tied directly to the complementarity with potatoes. It is shown that as the fraction of lactase persistence rises to one, the effect of potato suitability on population growth doubles the estimate found in Nunn and Qian (2011). This verifies the theoretical complementarity between the two food sources and shows that both potatoes and milk consumption contributed to the large growth in Old World population between the 18th and 19th century.

Chapter 5

Conclusion

This dissertation is a collection of essays that show the role of genetic adaptations to agriculture in leading to varied economic outcomes, past and present.

In Chapter 2 I show that the frequency of lactose tolerance, or lactase persistence, measured at the country level, has a positive and statistically strong relationship with population densities in the year 1500 CE. Given the Malthusian economy of the time, this implies the ability to consume milk, which is conferred by a gene, is associated with historic levels of development in Old World states. Of prime importance is the reason as to why lactose tolerance varies across countries. Lactose tolerance is the result of natural selection since the domestication of animals, which corresponds with the Neolithic Revolution. This echoes the theme of this dissertation: Agriculture constituted a major environmental shift, and beneficial adaptations to this environment resulted in beneficial economic effects.

The relationship between milk consumption and population densities in 1500 CE is robust to a large number of estimations. Truncations accounting for continental, environmental, and cultural effects do not alter the positive and significant relationship found in the baseline estimation. Additionally, the inclusion of a wide array of potentially omitted variables has little effect on the baseline estimated effect of the frequency of lactase persistence. As a further step to establish the relationship between lactase persistence and pre-colonial population densities, I exploit the natural relationship between solar radiation

the ability to consume milk. Low levels of solar radiation result in lowered production of vitamin D and ultimately rickets. Milk consumption partly alleviates this effect by providing high levels of calcium, which offset the deleterious effect of rickets, providing an added benefit of lactose tolerance and an exogenous way to measure differences in the frequency of lactose tolerance. Use of instrumental variables estimation does not alter the estimated effect of lactose tolerance on 1500 CE population densities and provides further proof of a causative relationship between the ability to consume milk and pre-colonial development.

The third chapter explores a different adaptation to the agricultural environment: resistance to a group of infectious diseases. The advantage of an earlier and more widespread adoption of agriculture, along with a large number of domesticate animals, led to the establishment of infectious pathogens into human hosts. This implies peoples of Eurasia have had greater exposure to certain diseases, while peoples of other areas have not. The third chapter argues that this historical exposure has shaped a certain part of the genome, providing a greater inherent resistance to peoples from Eurasia.

Inherent genetic resistance is measured through differences within the human leukocyte antigen (HLA) system. The HLA system is associated with the recognition and disposal of foreign bodies within the human body. The development of infectious Eurasian diseases altered variation within the HLA system. Due to the large number of infectious diseases, prolonged exposure to these diseases, and the fact that pathogens have an advantage to adaption conferred by shorter generations, variation within the HLA system was naturally selected; this is referred to as balancing selection.

Using a genetic measure for variation within the HLA system, I show that states with greater HLA variation have longer life expectancy in 1960, a year used to represent health outcomes before the widespread distribution of effective medicines and vaccines. This relationship holds while controlling for other relevant determinates of health: GDP per capita, latitude, and other measures of genetic differences. This relationship is given further strength through the use of instrumental variables estimation. According to the "Out

of Africa" hypothesis, all modern humans are derived from the highlands of East Africa. From this point, modern humans migrated to occupy the entire planet, excluding Antarctica. During this migration, population bottlenecks occurred from which only a small portion of genetic variation was accounted for in the relatively small number of migrants. This is known as the "serial founder effect" and resulted in a negative, linear relationship between genetic variation and the geographic distance from East Africa. In a recent work, Ashraf and Galor (2012) exploit this natural relationship in order to identify the effect of genetic variation. I too use this relationship in order to confirm the effect of HLA variation, from which it is found that HLA variation has a positive and statistically significant effect on life expectancy in 1960.

The effect of milk consumption is revisited in Chapter 4. Milk is a complement in the diet to many staple crops. Chapter 4 shows the complementarity with the introduction of the potato to the Old World. Potatoes were domesticated in the Andes of South America; after European contact with the New World, the potato was brought to Eurasia for use. The potato is nutritionally superior to other Old World staple crops and led to a boom in population in 18th and 19th centuries (Nunn and Qian 2011). The effect of potatoes is found through the use of exogenous soil conditions that permitted easy growth of potatoes. With this idea in mind, I consider the frequency of lactose tolerance as an exogenous determinate of the availability of milk consumption, from which the high levels of fats, proteins, and vitamins within milk increased the efficaciousness of the potato, which is high in calories, in increasing populations. This idea is confirmed, where it is shown that as the frequency of lactose tolerance rises to unity, the effect of the potato on Old World population growth doubles the previously found effect in Nunn and Qian (2011).

In summary, this dissertation shows that adaptations to the agricultural environment have led to differences in economic outcomes. In my view, contemporary disparities in wealth are not solely determined by contemporary factors. Instead, historical differences, both biological and cultural, have contributed to differences in the factors of economic

growth and therefore form a more ultimate source of differences in the wealth of nations.

Bibliography

[1] Nasa surface meteorology and solar energy, September 2011.

[2] D. Acemoglu and S. Johnson. Disease and development: the effect of life expectancy on economic growth, 2006.

[3] Daron Acemoglu, Simon Johnson, and James Robinson. The colonial origins of comparative development: An empirical investigation. *The American Economic Review*, 91(5):1369–1401, 2001.

[4] Daron Acemoglu, Simon Johnson, and James Robinson. Reversal of fortune: Geography and institutions in the making of the modern world income distribution. *Quarterly Journal of Economics*, 117(4):1231–1294, 2002.

[5] Daron Acemoglu, Simon Johnson, and James Robinson. The rise of europe: Atlantic trade, institutional change, and economic growth. *The American Economic Review*, 95(3):546–579, 2005.

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

[7] B. Anderson and C. Vullo. Did malaria select for primary adult lactase deficiency? *Gut*, 35(10):1487, 1994.

[8] R.M. Anderson and RM May. Infectious diseases of humans: dynamics and control. *New York*, page 757, 1991.

[9] M.A. Arroyo, W.B. Sateren, D. Serwadda, R.H. Gray, M.J. Wawer, N.K. Sewankambo, N. Kiwanuka, G. Kigozi, F. Wabwire-Mangen, M. Eller, et al. Higher hiv-1 incidence and genetic complexity along main roads in rakai district, uganda. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 43(4):440, 2006.

[10] Q. Ashraf and O. Galor. Malthusian population dynamics: Theory and evidence. *The American Economic Review*, 2011.

[11] Q. Ashraf and O. Galor. The 'out of africa' hypothesis, human genetic diversity, and comparative economic development. *American Economic Review*, 2012.

[12] W. Bank. World development indicators. *World Bank*, 2010.

[13] Ian Barnes, Anna Duda, Oliver G Pybus, and Mark G Thomas. Ancient urbanization predicts genetic resistance to tuberculosis. *Evolution*, 65(3):842–8, Mar 2011.

[14] T Bersaglieri, P Sabeti, N Patterson, T Vanderploeg, S Schaffner, J Drake, M Rhodes, D Reich, and J Hirschhorn. Genetic signatures of strong recent positive selection at the lactase gene. *The American Journal of Human Genetics*, 74(6):1111–1120, 2004.

[15] K.K. Bhatia, F.L. Black, T.A. Smith, M.L. Prasad, and G.N. Koki. Class i hla antigens in two long-separated populations: Melanesians and south amerinds. *American journal of physical anthropology*, 97(3):291–305, 1995.

[16] F.L. Black. Measles endemicity in insular populations: critical community size and its evolutionary implication. *Journal of Theoretical Biology*, 11(2):207–211, 1966.

[17] F.L. Black. An explanation of high death rates among new world peoples when in contact with old world diseases. *Perspectives in biology and medicine*, 37(2):292, 1994.

[18] F.L. Black, W.J. HIERHOLZER, F.D.K.P. PINHEIRO, A.S. EVANS, J.P. WOODALL, E.M. OPTON, J.E. EMMONS, B.S. WEST, G. EDSALL, W.G. DOWNS, et al. Evidence for persistence of infectious agents in isolated human populations. *American Journal of Epidemiology*, 100(3):230–250, 1974.

[19] J.M. Blackwell, S.E. Jamieson, and D. Burgner. Hla and infectious diseases. *Clinical microbiology reviews*, 22(2):370, 2009.

[20] H. Bleakley. Disease and development: comments on acemoglu and johnson (2006). *Remarks delivered at the NBER Summer Institute on Economic Fluctuations and Growth, July*, 16:73–117, 2006.

[21] G. Bloom and P.W. Sherman. Dairying barriers affect the distribution of lactose malabsorption. *Evolution and Human Behavior*, 26(4):301–312, 2005.

[22] V. Bockstette, A. Chanda, and L. Putterman. States and markets: The advantage of an early start. *Journal of Economic Growth*, 7(4):347–369, 2002.

[23] V. Bockstette, A. Chanda, and L. Putterman. States and markets: The advantage of an early start. *Journal of Economic Growth*, 7(4):347–369, 2002.

[24] J.S. Brownstein, C.J. Wolfe, and K.D. Mandl. Empirical evidence for the effect of airline travel on inter-regional influenza spread in the united states. *PLoS medicine*, 3(10):e401, 2006.

[25] J. Burger, M. Kirchner, B. Bramanti, W. Haak, and MG Thomas. Absence of the lactase-persistence-associated allele in early neolithic europeans. *Proceedings of the National Academy of Sciences*, 104(10):3736, 2007.

[26] L.L. Cavalli-Sforza, P. Menozzi, and A. Piazza. *The history and geography of human genes*. Princeton Univ Pr, 1994.

[27] A. Chanda and L. Putterman. Early starts, reversals and catch-up in the process of economic development. *Scandinavian Journal of Economics*, 109(2):387, 2007.

[28] Gregory Clark. *A Farewell to Alms: A Brief Economic History of the World*. Princeton University Press, 2008.

[29] M. Coelho, D. Luiselli, G. Bertorelle, A.I. Lopes, S. Seixas, G. Destro-Bisol, and J. Rocha. Microsatellite variation and evolution of human lactase persistence. *Human genetics*, 117(4):329–339, 2005.

[30] M.N. Cohen, G.J. Armelagos, Wenner-Gren Foundation for Anthropological Research, and State University of New York College at Plattsburgh. *Paleopathology at the Origins of Agriculture*. Academic Press New York, NY, USA:, 1984.

[31] D.A. Comin, W. Easterly, and E. Gong. Was the wealth of nations determined in 1000 bc?, 2006.

[32] GC Cook and MT Al-Torki. High intestinal lactase concentrations in adult arbs in saudi arabia. *British Medical Journal*, 3(5976):135, 1975.

[33] G.S. Cooke and A.V.S. Hill. Genetics of susceptibility to human infectious disease. *Nature Reviews Genetics*, 2(12):967–977, 2001.

[34] M.O. Cooper and W.J. Spillman. *Human food from an acre of staple farm products*. US Dept. of Agriculture, 1917.

[35] A.W. Crosby. *Ecological imperialism: the biological expansion of Europe, 900-1900*. Cambridge Univ Pr, 2004.

[36] D.M. Cutler, A.S. Deaton, and A. Lleras-Muney. The determinants of mortality, 2006.

[37] R. Dawkins. *The extended phenotype: the gene as the unit of selection*. 1982.

[38] R. Dawkins. *The Selfish Gene: –with a new Introduction by the Author*. Oxford University Press, USA, 2006.

[39] P.I.W. de Bakker, G. McVean, P.C. Sabeti, M.M. Miretti, T. Green, J. Marchini, X. Ke, A.J. Monsuur, P. Whittaker, M. Delgado, et al. A high resolution hla and snp haplotype map for disease association studies in the extended human mhc. *Nature genetics*, 38(10):1166, 2006.

[40] J. Diamond. *Guns, Germs, and Steel: The Fates of Human Societies*. W.W. Norton, 1998.

[41] A.P. Dobson and E.R. Carper. Infectious diseases and human population history. *Bioscience*, pages 115–126, 1996.

[42] N.S. Enattah, T.G.K. Jensen, M. Nielsen, R. Lewinski, M. Kuokkanen, H. Rasinpera, H. El-Shanti, J.K. Seo, M. Alifrangis, I.F. Khalil, et al. Independent introduction of two lactase-persistence alleles into human populations reflects different history of adaptation to milk culture. *The American Journal of Human Genetics*, 82(1):57–72, 2008.

[43] S.L. Engerman and K.L. Sokoloff. Factor endowments: institutions, and differential paths of growth among new world economies: a view from economic historians of the united states, 1994.

[44] S.L. Engerman and K.L. Sokoloff. Factor endowments, inequality, and paths of development among new world economics, 2002.

[45] FAO. Global agro-ecological zones, 2002.

[46] G. Flatz and H.W. Rotthauwe. Lactose nutrition and natural selection. *The Lancet*, 302(7820):76–77, 1973.

[47] World Health Organization. Borrow Dental Milk Foundation and KW Stephen. *Milk fluoridation for the prevention of dental caries*. World Health Organization, 1996.

[48] J.L. Gallup, J.D. Sachs, and A.D. Mellinger. Geography and economic development, 1998.

[49] O. Galor and O. Moav. Natural selection and the origin of economic growth. *The Quarterly Journal of Economics*, 117(4):1133, 2002.

[50] O. Galor and O. Moav. The neolithic origins of contemporary variation in life expectancy. *Brown University Department of Economics Working Paper*, 14:2007, 2007.

[51] P. Gerbault, C. Moret, M. Currat, and A. Sanchez-Mazas. Impact of selection and demography on the diffusion of lactase persistence. *PLoS One*, 4(7):e6369, 2009.

[52] H.J. Greenfield, J. Chapman, AT Clason, A.S. Gilbert, B. Hesse, and S. Milisauskas. The origins of milk and wool production in the old world: A zooarchaeological perspective from the central balkans [and comments]. *Current Anthropology*, 29(4):573–593, 1988.

[53] L. Gueguen and A. Pointillart. The bioavailability of dietary calcium. *Journal of the American College of Nutrition*, 19(suppl 2):119S, 2000.

[54] L. Guiso, P. Sapienza, and L. Zingales. Cultural biases in economic exchange, 2004.

[55] L. Guiso, P. Sapienza, and L. Zingales. Does local financial development matter?*. *Quarterly journal of Economics*, 119(3):929–969, 2004.

[56] M. Harrison. *Disease and the modern world: 1500 to the present day*. Polity, 2004.

[57] D.L. Hartl and A.G. Clark. *Principles of population genetics*, volume 116. Sinauer associates Sunderland, MA, 4 edition, 2006.

[58] J. Hawks, E.T. Wang, G.M. Cochran, H.C. Harpending, and R.K. Moyzis. Recent acceleration of human adaptive evolution. *Proceedings of the National Academy of Sciences*, 104(52):20753, 2007.

[59] D.A. Hibbs and O. Olsson. Geography, biogeography, and why some countries are rich and others are poor. *Proceedings of the National Academy of Sciences of the United States of America*, 101(10):3715, 2004.

[60] A.V.S. Hill. The immunogenetics of human infectious diseases. *Annual review of immunology*, 16(1):593–617, 1998.

[61] E.J. Hollox, M. Poulter, M. Zvarik, V. Ferak, A. Krause, T. Jenkins, N. Saha, A.I. Kozlov, and D.M. Swallow. Lactase haplotype diversity in the old world. *The American Journal of Human Genetics*, 68(1):160–172, 2001.

[62] C. Hoppe, C. Mølgaard, and K.F. Michaelsen. Cow's milk and linear growth in industrialized and developing countries. *Annu. Rev. Nutr.*, 26:131–173, 2006.

[63] R. Horton, L. Wilming, V. Rand, R.C. Lovering, E.A. Bruford, V.K. Khodiyar, M.J. Lush, S. Povey, C.C. Talbot, M.W. Wright, et al. Gene map of the extended human mhc. *Nature Reviews Genetics*, 5(12):889–899, 2004.

[64] A.L. Hughes and M. Yeager. Natural selection at major histocompatibility complex loci of vertebrates. *Annual review of genetics*, 32(1):415–435, 1998.

[65] C.J.E. Ingram, C.A. Mulcare, Y. Itan, M.G. Thomas, and D.M. Swallow. Lactose digestion and the evolutionary genetics of lactase persistence. *Human genetics*, 124(6):579–591, 2009.

[66] C.J.E. Ingram, C.A. Mulcare, Y. Itan, M.G. Thomas, and D.M. Swallow. Lactose digestion and the evolutionary genetics of lactase persistence. *Human genetics*, 124(6):579–591, 2009.

[67] C.J.E. Ingram, T.O. Raga, A. Tarekegn, S.L. Browning, M.F. Elamin, E. Bekele, M.G. Thomas, M.E. Weale, N. Bradman, and D.M. Swallow. Multiple rare variants as a cause of a common phenotype: several different lactase persistence associated alleles in a single ethnic group. *Journal of molecular evolution*, 69(6):579–588, 2009.

[68] M.C. Inhorn and P.J. Brown. The anthropology of infectious disease. *Annual review of Anthropology*, 19:89–117, 1990.

[69] Y. Itan, A. Powell, M.A. Beaumont, J. Burger, and M.G. Thomas. The origins of lactase persistence in europe. *PLoS computational biology*, 5(8):e1000491, 2009.

[70] A.K. Jain, TC Hsu, R. Freedman, and MC Chang. Demographic aspects of lactation and postpartum amenorrhea. *Demography*, 7(2):255–271, 1970.

[71] K.J.M. Jeffery and C.R.M. Bangham. Do infectious diseases drive mhc diversity? *Microbes and Infection*, 2(11):1335–1341, 2000.

[72] HB Kettlewell. D. 1956. further selection experiments on industrial melanism in the lepidoptera. *Heredity*, 10:287–301, 1956.

[73] A. Kiszewski, A. Mellinger, A. Spielman, P. Malaney, S.E. Sachs, and J. Sachs. A global index representing the stability of malaria transmission. *The American journal of tropical medicine and hygiene*, 70(5):486, 2004.

[74] J. Klein et al. *Natural history of the major histocompatibility complex*. Wiley New York, 1986.

[75] R. La Porta. Lopez-de-silanes, f., shleifer, a., vishny, r., 1998. law and finance. *Journal of Political Economy*, 106(6):1113–1155, 1998.

[76] T.R. Malthus. An essay on the principle of population. 1798. *Reprint. Amherst, NY: Prometheus Books*, 1998.

[77] R.D. McCracken. Lactase deficiency: an example of dietary evolution. *Current Anthropology*, 12(4/5):479–517, 1971.

[78] C. McEvedy and R. Jones. *Atlas of world population history.* Penguin Books Ltd, Harmondsworth, Middlesex, England., 1978.

[79] W.H. McNeill. *Plagues and peoples*. Anchor, 1976.

[80] W.H. McNeill. How the potato changed the world's history. *Social research*, 66(1):67–83, 1999.

[81] T. Meloni, C. Colombo, G. Ruggiu, M. Dessena, and GF Meloni. Primary lactase deficiency and past malarial endemicity in sardinia. *Italian journal of gastroenterology and hepatology*, 30(5):490, 1998.

[82] D. Meyer, R.M. Single, S.J. Mack, H.A. Erlich, and G. Thomson. Signatures of demographic history and natural selection in the human major histocompatibility complex loci. *Genetics*, 173(4):2121, 2006.

[83] S. Michalopoulos. The origins of ethnolinguistic diversity: Theory and evidence. *Department of Economics, Tufts University*, 2008.

[84] CA Mulcare. The evolution of the lactase persistence phenotype. *London: University of London*, page 311, 2006.

[85] C.A. Mulcare, M.E. Weale, A.L. Jones, B. Connell, D. Zeitlyn, A. Tarekegn, D.M. Swallow, N. Bradman, and M.G. Thomas. The t allele of a single-nucleotide polymorphism 13.9 kb upstream of the lactase gene (lct)(c-13.9 kbt) does not predict or cause the lactase-persistence phenotype in africans. *The American Journal of Human Genetics*, 74(6):1102–1110, 2004.

[86] N. Nunn. The long-term effects of africa's slave trades*. *Quarterly Journal of Economics*, 123(1):139–176, 2008.

[87] N. Nunn. The importance of history for economic development, 2009.

[88] N. Nunn and D. Puga. Ruggedness: The blessing of bad geography in africa, 2009.

[89] N. Nunn and N. Qian. The potato's contribution to population and urbanization: Evidence from a historical experiment. *The Quarterly Journal of Economics*, 126(2):593–650, 2011.

[90] U.S. Department of Agriculture. Usda national nutrient database for standard references, December 2011.

[91] SB Piertney and MK Oliver. The evolutionary ecology of the major histocompatibility complex. *Heredity*, 96(1):7–21, 2005.

[92] R.L. Porta, F. Lopez-de Silane, A. Shleifer, and R.W. Vishny. Law and finance. Technical report, National Bureau of Economic Research, 1996.

[93] F. Prugnolle, A. Manica, and F. Balloux. Geography predicts neutral genetic diversity of human populations. *Current Biology*, 15(5):R159–R160, 2005.

[94] F. Prugnolle, A. Manica, M. Charpentier, J.F. Guégan, V. Guernier, and F. Balloux. Pathogen-driven selection and worldwide hla class i diversity. *Current Biology*, 15(11):1022–1027, 2005.

[95] L. Putterman. Agriculture, diffusion and development: Ripple effects of the neolithic revolution. *Economica*, 75(300):729–748, 2008.

[96] L. Putterman and D.N. Weil. Post-1500 population flows and the long-run determinants of economic growth and inequality*. *Quarterly Journal of Economics*, 125(4):1627–1682, 2010.

[97] H. Rajeevan, M.V. Osier, K.H. Cheung, H. Deng, L. Druskin, R. Heinzen, J.R. Kidd, S. Stein, A.J. Pakstis, N.P. Tosches, et al. Alfred: the alelle frequency database. update. *Nucleic Acids Research*, 31(1):270, 2003.

[98] S. Ramachandran, O. Deshpande, C.C. Roseman, N.A. Rosenberg, M.W. Feldman, and L.L. Cavalli-Sforza. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in africa. *Proceedings of the National Academy of Sciences of the United States of America*, 102(44):15942, 2005.

[99] P C Sabeti, S F Schaffner, B Fry, J Lohmueller, P Varilly, O Shamovsky, A Palma, T S Mikkelsen, D Altshuler, and E S Lander. Positive natural selection in the human lineage. *Science*, 312(5780):1614–20, Jun 2006.

[100] A. Sherratt. The secondary exploitation of animals in the old world. *World Archaeology*, 15(1):90–104, 1983.

[101] A. Sherratt. The secondary exploitation of animals in the old world. *World Archaeology*, 15(1):90–104, 1983.

[102] T. Shiina, H. Inoko, and JK Kulski. An update of the hla genomic region, locus information and disease associations: 2004. *Tissue Antigens*, 64(6):631–649, 2004.

[103] K. Shillington. *History of Africa*. Bedford Books, 1989.

[104] F.J. Simoons. Primary adult lactose intolerance and the milking habit: a problem in biologic and cultural interrelations. *Digestive Diseases and Sciences*, 15(8):695–710, 1970.

[105] F.J. Simoons. Primary adult lactose intolerance and the milking habit: a problem in biologic and cultural interrelations. *Digestive Diseases and Sciences*, 15(8):695–710, 1970.

[106] Frederick J. Simoons. The geographic hypothesis and lactose malabsorption. *Digestive Diseases and Sciences*, 23:963–980, 1978. 10.1007/BF01263095.

[107] RW Slade and HI McCallum. Overdominant vs. frequency-dependent selection at mhc loci. *Genetics*, 132(3):861, 1992.

[108] E. Spolaore and R. Wacziarg. The diffusion of development*. *Quarterly Journal of Economics*, 124(2):469–529, 2009.

[109] S.A. Tishkoff, F.A. Reed, A. Ranciaro, B.F. Voight, C.C. Babbitt, J.S. Silverman, K. Powell, H.M. Mortensen, J.B. Hirbo, M. Osman, et al. Convergent adaptation of human lactase persistence in africa and europe. *Nature genetics*, 39(1):31–40, 2006.

[110] S.A. Tishkoff, F.A. Reed, A. Ranciaro, B.F. Voight, C.C. Babbitt, J.S. Silverman, K. Powell, H.M. Mortensen, J.B. Hirbo, M. Osman, et al. Convergent adaptation of human lactase persistence in africa and europe. *Nature Genetics*, 39(1):31–40, 2006.

[111] J.A. Traherne, R. Horton, A.N. Roberts, M.M. Miretti, M.E. Hurles, C.A. Stewart, J.L. Ashurst, A.M. Atrazhev, P. Coggill, S. Palmer, et al. Genetic analysis of completely sequenced disease-associated mhc haplotypes identifies shuffling of segments in recent human history. *PLoS genetics*, 2(1):e9, 2006.

[112] M.E. Wilson. Travel and the emergence of infectious diseases. *Emerging infectious diseases*, 1(2):39, 1995.

[113] N.D. Wolfe, C.P. Dunavan, and J. Diamond. Origins of major human infectious diseases. *NATURE-LONDON-*, 447(7142):279, 2007.

[114] S. WRIGHT. *Evolution and the genetics of population (vol 4) variability within and among natural populations, paper*. 1984.

Appendix—Country Level Lactase Persistence Frequency

Country	Frequency	Country	Frequency	Country	Frequency
Afghanistan	0.249344	Indonesia	0.36	Oman	0.437636
Albania	0.621989	Iran, Islamic Rep.	0.250012	Pakistan	0.514856
Algeria	0.296263	Iraq	0.415687	Philippines	0.36
Angola	0.08533	Ireland	0.96	Poland	0.627091
Armenia	0.30708	Italy	0.476237	Portugal	0.66
Austria	0.792949	Japan	0.28	Rwanda	0.517425
Bangladesh	0.550441	Jordan	0.231748	Saudi Arabia	0.469029
Belarus	0.623756	Kazakhstan	0.225231	Senegal	0.687169
Belgium	0.803062	Kenya	0.172536	Slovakia	0.815073
Benin	0.199309	Korea, Rep. (South)	0.28	Slovenia	0.628075
Botswana	0.267943	Kuwait	0.504716	Somalia	0.232573
Bulgaria	0.431158	Kyrgyzstan	0.178806	South Africa	0.272754
Burkina Faso	0.652184	Laos	0.023333	Spain	0.66
Burundi	0.487714	Latvia	0.551405	Sri Lanka	0.251405
Cambodia	0.353603	Lebanon	0.22	Sudan	0.409244
Cameroon	0.186025	Lesotho	0.318861	Sweden	0.96
China	0.08	Liberia	0.16218	Switzerland	0.800882
Congo, Rep.	0.12004	Libya	0.411466	Syrian Arab Republic	0.392525
Croatia	0.622128	Lithuania	0.609445	Tajikistan	0.144223
Czech Republic	0.761653	Macedonia	0.564921	Tanzania	0.142488
Denmark	0.96	Madagascar	0.339264	Thailand	0.048417
Egypt, Arab Rep.	0.274193	Malawi	0.12004	Tunisia	0.16
Estonia	0.571863	Malaysia	0.340161	Turkey	0.335788
Ethiopia	0.387309	Mali	0.338345	Turkmenistan	0.290095
Finland	0.845248	Moldova	0.424611	Uganda	0.230639
France	0.707426	Mongolia	0.127771	Ukraine	0.541916
Gabon	0.120649	Morocco	0.16	United Arab Emirates	0.371411
Gambia	0.42899	Mozambique	0.12259	United Kingdom	0.946144
Germany	0.854067	Myanmar	0.082052	Uzbekistan	0.056004
Ghana	0.247387	Namibia	0.086346	Vietnam	0.354475
Greece	0.546041	Nepal	0.495627	Yemen	0.53
Guinea	0.389424	Netherlands	0.854067	Zambia	0.109302
Guinea-Bissau	0.587258	Niger	0.561083	Zimbabwe	0.152507
Hungary	0.625469	Nigeria	0.350444		
India	0.439754	Norway	0.95551		

Vita

Justin Cook was born in Natchitoches, Louisiana. In 2002, he earned a Bachelor of Science degree in economics from L.S.U. He continued on to graduate school at L.S.U. earning a Master of Science degree in economics in 2009. The second chapter of the dissertation is currently being revised for resubmission to the *Journal of Economic Growth*.