Adaptation to milk drinking and evolution of lactase persistence in pastoralist goat herders in central-northern Chile

Nicolas Andres Montalva Rivera

A thesis for the Doctor of Philosophy degree at University College London

Department of Anthropology

UCL

June 2014

I, Nicolas Andres Montalva Rivera, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Milk contains lactose as source of energy, which is digested by intestinal lactase, an enzyme that declines after weaning. The genetic trait of lactase persistence (LP) evolved together with the development of milking and pastoralism in the Old World as an adaptation to milk consumption in adults. The spread of this trait is one of the best examples of positive natural selection in humans. However, the specific mechanisms conferring selective advantages to LP are unknown.

Milk drinking was introduced in the last 500 years to South America. To better understand the relationship between milk drinking, LP, growth, reproduction, and survival, this thesis explores the nature and extent of this dietary adaptation in the pastoralist communities of central–northern Chile, a population of mixed Amerindian and European ancestry, thus with persistent and non–persistent individuals.

Data collected during 10 months of fieldwork consisted of questionnaires about reproduction and diet, DNA samples, and measurements of stature and weight of 450 participants.

The lactase gene enhancer region was sequenced in all samples, and the European *−13,910*T* was the only LP–associated variant found. Phenotypes of 41 participants were established using hydrogen breath tests, showing strong association of this variant with phenotype. The frequency of LP in this population (0.38) is similar to that of non–pastoralist admixed populations of South America.

To evaluate the effects of population structure, DNA analysis was used to study ancestry and relatedness. Controlling for these and other variables, the associations of LP and milk consumption with fertility, mortality, height and weight were assessed. We found no effect of LP on fertility, but a significant effect of LP on BMI and of BMI on fertility. These results suggest additional studies to evaluate the relationship between LP, BMI and fertility as an hypothesis of one of the possible routes to the positive selection for lactase persistence.

Acknowledgements

Interdisciplinarity has its difficulties and frustrations, but can also be extremely rewarding. My project had the good fortune to be part of a collaborative effort between the Department of Anthropology and the Research Department of Genetics, Evolution and Environment at UCL, giving me two homes and the opportunity to interact with two wonderful groups of scientists which I would like to thank for their support in this thesis.

I had the pleasure of being jointly supervised by two formidable researchers. I am very grateful to Prof. Ruth Mace for her patience during those first years of helping me to get used to a completely new environment, her approach to quickly putting out my administrative 'fires', her ability to make me think about the entire perspective of the problems, and her inspiring commitment to the cause of a scientific approach in anthropology. I would like to express my deep gratitude to Prof. Dallas Swallow for taking this project to heart, her detailed reviews of my writing (and her patience in dealing with my particular tendency to write too much), her precise ability to identify key aspects, her patience to teach me a field that was at times too distant from my original background, and for her support in several aspects of my academic and extra–academic life.

I appreciate the feedback given by Dr. Andrea Migliano in the design of this study, the advice offered by Prof. Andrés Ruiz Linares and Dr. Kaustubh Adhikari about the estimation of ancestry proportions and for sharing with me their panel of AIMs and data from parental populations, and the insightful discussions with Dr. Nikolas Maniatis about the use of STRs to study cryptic relatedness.

I also want to thank Ranji Arasaretnam, Mari-Wyn Burley, Rosemary Ekong, Fraser Simpson, and Tyrone Young for their enormous assistance with labwork.

I would like to express my gratitude to former and current students at both departments for their support and their fruitful comments about this project. I have greatly benefited from the discussions at the meetings of the Human Behavioural Ecology Group, specially to Heidi Colleran for advice about design of questionnaires. I am also grateful to Elizabeth Gallagher, Mirna Kovacevic, Adam Powell, Adrian Timpson, Catherine Walker, Ripu Bains, Victor Acuña, Juan Chacón,

and Heather Elding for their support, and particularly to Katherine Brown, Pascale Gerbault, and Anna Rudzinski for their help in the lab and their comments.

Very special thanks to Bryony Jones, who helped me with my firsts steps in labwork, and Anke Liebert, who shared with me her panel of markers and her data to infer *LCT* haplotypes. She was also an invaluable source of knowledge, suggestions, comments, and support.

I would also like to thank Prof. Gloria Gallardo from the Uppsala Centre for Sustainable Development in Sweden for sharing her knowledge about the Agricultural Communities with me.

This work would have been impossible without the help and support of the University of Chile. Prof. Sergio Flores provided suggestions at the earliest stages of this project, and allowed me to use the laboratory of genetics at the Department of Anthropology during the fieldwork, where I was closely assisted by Prof. Sebastian Krapivka. I would also like to thank undergraduate students who helped me in both labwork and fieldwork: Aurea Argomedo, Camila Balcazar, Pamela Cañas, Paloma Contreras, Paulina Contreras, Catalina Fernandez, Sandra Flores, Tomas Gonzalez, Paulina Jara, Evelyn Maldonado, Camila Maripangui, Maria Jose Moraga, Paulina Morales, Evelyn Munzenmayer, Constanza Silva, Paula Tralma, and Gabriela Urrutia. I really enjoyed working with you.

I would also like to express my gratitude to the people of the Agricultural Communities of the Chilean *"Norte Chico"* who are the most important part of this project. I am particularly in debt with Mirtha Gallardo, Jehova Ibacache, Homero Cortes, Rodolfo Villar, Luis Valderrama, Rosa Alvallay and Pedro Toledo.

At a personal level, I am very grateful to Allen, Lisa, and Solomon, who taught me how to live in England, and were incredibly supportive at difficult times. I would like to thank my friends Alvaro, Anibal, Claudia, Emiliano, Gabriel, Hugo, Inge, Isi, Iván, María José, Pancha, Piero, Poly, Raimundo, Ricardo, Roman, and Xime. I always felt your support through these years.

I wish to thank my family: Gonzalo, who drive hordes of anthropology students into the field site, Marcela, who was a constant source of wisdom, love, and care. My sister Angela, who always made me think that she had an eye on everything while I was away. My brother Sebastian, who also contributed to the logistics of this project making queries in the communities while I was busy with other things. And special thanks to my parents, Viviana and Andrés, for their love and support looking after my loved ones these years.

Finally, I wish to thank my daughter, Leonor, and my wife, Gabriela, for their love, support, and sacrifice beyond everything I thought bearable. This was a tough journey that we might have not followed knowing the hardships of the route ahead. At last, it is over. I hope there is a place in the

future from where we shall be able to see that it worth of all this. Meanwhile, I shall help time as much as I can to heal all the wounds. This thesis is for you.

This PhD was funded by the Bicentennial Becas–Chile Scholarship for the Advanced Human Capital Program by the Chilean National Commission for Scientific and Technological Research (CONICYT). Additional funding was provided by The Parkes Foundation Small Grant for Research Students, The Gen Foundation Award, the UCL Grand Challenge of Global Health, and the Annals of Human Genetics. I would like to avail of this opportunity to express my grateful thanks to these funding bodies for their interest in my project.

To Gabriela & Leonor

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This project aims to understand the dietary adaptation to milk consumption in a population of goat herders from South America. The following introduction explains the context of this research question around the broad concept of human adaptation, reviewing documented examples of adaptation to different environments and dietary habits. Adaptation to milk consumption through lactase persistence is one of the main fields of study of human adaptation and as the central topic of this thesis is introduced in its own section, which includes a literature review of lactase persistence in Latin America, a region poorly represented in studies about this trait. Finally, the history and ethnography of the communities of goat herders of Chile is introduced, describing the features that make them an interesting case study for our research questions.

1.1 Human Adaptation

1.1.1 Human Variability and Adaptation

Describing anthropology as the study of human variability is one of the few definitions that finds some degree of agreement in an otherwise factional discipline. It is historically founded as an interest in different human groups that led to an effort to discover the relationship between their behaviour, society, culture, and biology. Variability is a framework that appeals to social and biological anthropologists alike.

The study of human variability has always been related to issues of wider social interest. It played an important role in European expansionism and colonialism from the 15th to early 20th centuries (Harris, 2000), and was a form of scientific pretext to justify social inequalities until recently (Lewontin, 2001). It has also been used to orientate population-specific needs in, for instance, cultural policy, forensic identification, ergonomics, and healthcare (Harris, 1997). Today, the study of human variability faces the challenge of understanding the effects of globalisation and the spread of western behaviour on a wide array of other populations, each with particular biological traits and cultural livelihoods. The study of human variability can contribute towards developing public policies to improve local welfare and inter–cultural relations.

Part of the variation we are interested in is caused by traits gained by human groups as a way of adjusting to specific conditions. Many terms (such as adaptation, acclimatisation, accommodation, habituation, etc.) are used in the literature to classify these processes, which include individual responses, and different forms of variability at genetic, physiological, behavioural, or cultural levels as well as natural selection (see Mazess 1975 and Harrison & Morphy 1998 to compare different classification schemes). Although there is no agreement in terminology, there is general acknowledgement of the difference between heritable long-term modifications at a population level (adaptation) and individual non-heritable short-term responses (acclimatisation). Part of the difficulty in defining the different ways humans adjust to their living conditions resides on the deep interrelationships between them, since the range of responses an individual can have and how they are triggered is also result of an adaptation process. In the same way, variation at physiological, behavioural and cultural levels can also be result of heritable long–term modifications, and therefore considered as adaptations, as is acknowledged for culture by the Dual Inheritance Theory, a framework concerned with the study of the co–heritability of genetic and cultural traits and with the results of their interactions as gene–culture co–evolution (Aoki, 2001; Boyd, 1988; Cavalli-Sforza, 1981; Durham, 1982).

This thesis is about how a pastoralist population living in an adverse environment adapts to a recently introduced dietary behaviour. These three key elements: pastoralism, extreme environments, and changes in diet, have been useful models to study of human adaptation.

1.1.2 Pastoralism as a Model of Human Adaptation

Pastoralism refers to a subsistence pattern based on animal rearing. It is commonly practised in areas of low productivity for agriculture and other resources, and usually requires the movement of herds according to availability of pasture. Pastoralism is classified as exclusive when herding is the only productive activity and other resources are obtained by selling or bartering animal products, and as agro–pastoralism when is complemented with some small–scale agriculture. A further classification is that of transhumance, if movements follow a regular seasonal pattern using (more or less) fixed routes, or as nomadic, if the herders migrate in an opportunistic way according to availability of resources. While economically important in many areas, rearing animals as part of sedentary productive strategies is not normally considered pastoralism (Leonard & Crawford, 2008).

There are several theories about the origins of pastoralism and many possible models are supported based on archaeological, historical, and anecdotal ethnographic evidence. According to Leonard & Crawford (2008) the Evenki of Siberia gradually started to capture and keep alive some of the reindeer they hunted and thus moved from a hunter–gatherer subsistence to pastoralism. A transition in the same direction occurred in Africa, where hunter–gatherers adopted pastoralism, possibly from the Near East, well before the introduction of farming, as shown by archaeological records of domestic animals in several areas by 5,500-6,000 BP, while the first domestic plants appeared around 4,000 BP (Gifford-Gonzalez & Hanotte, 2011; Marshall & Hildebrand, 2002)[1]. However, in most other regions pastoralism appeared after agriculture (Blench, 2001). Based on Middle–East prehistory, archaeologists argued that mixed strategies of dryland agriculture and pastoralism were prevalent initially, until the development of irrigation techniques by some groups who became specialised farmers, leaving animal herding to marginal nomads (Cribb, 1991). All these accounts suggest multiple origins and continuous bidirectional transitions between highly specialised pastoralism and farming, as has been documented in Africa by Mace et al. (1993).

Although scientists and policy makers have predicted the eventual decline of pastoralism as an inefficient mean of production (Blench, 2001; Leonard & Crawford, 2008; Mace, 1991), it is still common today. Well studied clusters of pastoralism in the Old World are located in East Africa, the Sahara, the Middle-East and the Central Asian steppe. In addition, less studied pastoralism is

[1] An Eastern origin is unarguably the case for the introduction of camels, sheep and goats in Africa. For cattle, there are a few proposals of an independent domestication event. See Gifford-Gonzalez & Hanotte, 2011 for a review.

currently practised in the Himalayas, North-east Russia, and in Europe by Basque groups in Spain and France, Saami in North Scandinavia, and in the Swiss and Italian Alps (Leonard & Crawford, 2008).

Surprisingly, pastoralism in South America has remained virtually unexplored by the scientific literature, even though it has been practised for at least 6,000 years and played a central role in the prehistory of Andean cultures and the expansion of the Inca empire (Blench, 2001). Two domestic species of camelids, the llama (*Lama glama*) and the alpaca (*Vicugna pacos*) were widely used as source of wool, meat and as pack animals. While their wild counterparts, guanacos (*Lama guanicoe*) and vicuñas (*Vicugna vicugna*) (Wheeler, 1995) were also trapped and kept captive for meat and wool production by hunter-gatherers during the earlier formative period (Baied & Wheeler, 1993), pastoralism became more important with the establishment of commercial and political networks between groups across the mountains, for which the use of llama caravans for packing was essential, particularly after the appearance of highly centralised political entities such as Tiwanaku and the Inca empire (Baied & Wheeler, 1993; Lynch, 1983).

After Spanish colonisation, species and customs of European pastoralism were incorporated into the practices of Andean herders. Today mobile pastoralism still can be found in the central Andean region, where alpaca wool is an important commodity in Peruvian and Bolivian economies. In this area pastoralism is practised by indigenous communities, and the area remains strongly linked to the Andean species, customs, and heritage (Blench, 2001; Browman et al., 1997; Westreicher et al., 2007). Pastoralism in the *'Atacama puna'* region in Northern Chile and Argentina is less important demographically and economically, but similar in terms of herd composition and indigenous influence (Göbel, 1997; Westreicher et al., 2006).

However, pastoralism today shows a wider range than traditional pre-Hispanic Andean camelid herding. Mobile pastoralism of recent origin exists in the Ecuadorian highlands *'páramos'* where both South American and European species are herded (Hess, 1990), and in the Patagonian region of Argentina and Chile, where sheep and cattle herding are practised by admixed populations in an area without animal domestication before European contact (Westreicher et al., 2006). Pastoralism oriented to milk production is common only in communities of admixed origin in central–northern Chile, where goats are dominant and camelids absent (Gallardo, 2002; Westreicher et al., 2006), and which forms the subject of this thesis.

As a highly specialised livelihood, pastoralist populations have to deal with specific pressures of both biological and cultural origin. Food production based preferentially on animals rather than crops implies certain restrictions as well as novel additions to the diet (e.g. milk and milk

products). By moving from place to place pastoralists avoid seasonal pathogens that could affect both themselves and their herds and this also allows them to allocate herds optimally according to the availability of pastures. Living in close contact with animals puts them at higher risk of zoonoses, and constant movement between places with different climates requires adaptation to different climatic conditions. These various features of pastoralist life have been widely acknowledged as being influential on shaping variability through natural selection, as will be reviewed in the next sections, making pastoralism a good model to study human adaptation in the context of the relationship between biology and culture.

1.1.3 Human Adaptation to Extreme Environments

As humans dispersed throughout the world, they had to face a wide variety of environments which were quite different from their original East African ecoclimate. Although cultural adaptation achieved through innovations in clothing, heating, housing, transport, hunting, and social structure facilitated human migration, biological adaptation achieved through natural selection is responsible of an important part of the phenotypic variability of this species. The following sections are some examples of different environmental challenges faced by human populations, and indications about how humans adapted to cope with them.

1.1.3.1 Temperature

One of the challenges faced by human expansion was posed by the diversity of temperatures around the world. Of particular interest is the hypothesis of variability in body size proportions as result of adaptation to temperature. The hypothesis was suggested first by Bergmann (1848), who introduced the idea known as *'Bergmann's rule'*. According to this rule large animals have lower rate of heat loss than small animals because their exposed surface is small in proportion to their body size. In agreement with this hypothesis, correlations between body size and latitude have been found at different taxonomic levels, such as among clades within mammals (Ashton et al., 2000; Freckleton et al., 2003), and species within primates (Harcourt & Schreier, 2009). Between human populations variation in body size is in general agreement with Bergmann's rule (Katzmarzyk & Leonard, 1998; Roberts, 1953; Ruff, 2002). However, there is little known about the genetic bases for these adaptations, and the assessment of the rule can be affected by some confounding variables such as poverty and malnutrition in tropical areas (Katzmarzyk & Leonard, 1998) or other evolutionary processes, like adaptation to move in the forest, sexual selection, or earlier reproduction (Jobling et al., 2013).

Apart from the study of body size as a thermoregulatory mechanism, other factors have been explored to explain human adaptation to hot and cold temperatures.

Historical records give evidence of very high mortality rates being associated with heat waves (Kovats & Hajat, 2008), and high temperature is known to be associated with low birth weight (Wells, 2002). Hancock et al. (2011b) have found a strong association between an allele in a gene called *KRT77*, which encodes keratin, and high temperatures, suggesting an adaptation to hot weather due to the expression of this gene in the ducts of eccrine sweat glands. But as a species adapted to the warm East African weather, human adaptation to cold temperatures must have been a significant step in colonising most of the world[2]. A higher basal metabolic rate, contributing to heat production, has been found in Siberian populations (Leonard et al., 2002). This phenotype has been explained as a genetic adaptation, attributed to increased thyroid function, rather than mere acclimatisation, since native populations from North America do not show increased metabolic rate, despite sharing a similar environment, diet and subsistent pattern (Leonard et al., 2005). A strong signal of selection has been found for an allele in *TRIP6* (Thyroid receptor-interacting protein 6) found in populations living in places with low temperatures (Hancock et al., 2011b), which can be interpreted as supporting this hypothesis.

The most important form of heat production in infants, known as non-shivering thermogenesis, is mediated by the uncoupling protein UCP1. UCP1 decreases the proton gradient in the mitochondrial membrane of brown adipose tissue, generating heat. An allele in *UCP1*, as well as alleles in genes of other uncoupling proteins, are particularly prevalent in regions with low temperatures, and show strong signals of natural selection (Hancock et al., 2011a).

A variant of the gene *EDAR* which encodes the ectodysplasin receptor and is involved in signal transduction, has also been interpreted as an adaptation to cold temperatures. One role of the common variant of this gene (T > C, *Val* > *Ala*) is to increase hair thickness and hair straightness, and it is also known that the gene plays a role developing of sweat and mammary glands as has been demonstrated with mouse models (Chang et al., 2009; Kamberov et al., 2013). The nonsynonymous polymorphism affects activity of NF-$\varkappa$B, and shows very high frequencies in Asians and Native Americans, but is very rare everywhere else (Fujimoto et al., 2008a,b). This variant shows very strong signals of natural selection in Asian populations, but how its phenotypic effects contribute to adaptation is not clear and several hypotheses have been formulated in the literature: Thick hair, straight hair, and contribution of sebaceous glands to impeding evaporation have been all interpreted as adaptation to the cold environment of Asia during the ice age (Chang et al., 2009; Tan et al., 2013), yet alternatives based on sexual selection have also been proposed (Kamberov et al., 2013).

[2] Antarctica, the coldest continent on Earth, is indeed a notable exception of the expansion of humans through the world.

1.1.3.2 Latitude and UV radiation

For many years, human pigmentation was erroneously considered as a signal of genealogical relations among human populations (Relethford, 2002). However, now is known to be the result of adaptation to different amounts of UV radiation, as shown by the high correlation between pigmentation and latitude (Jablonski & Chaplin, 2000). Variation in pigmentation is caused by different size and distribution of pigment–production cells located in the bottom layer of the epidermis, known as melanocytes, which are known to be stimulated by exposure to sunlight (García-Borrón et al., 2005).

Many genes have a role in pigmentation, and important allelic differences in their worldwide distribution had been found. Signals of selection of variants favouring light colours in *SLC245A2* (Izagirre et al., 2006; Lamason et al., 2005; Norton et al., 2007; Sulem et al., 2007), *MATP* (Graf et al., 2005; Norton et al., 2007), *MC1R* (Beaumont, 2005; Flanagan et al., 2000; Harding et al., 2000; Sulem et al., 2007), *OCA2* (Norton et al., 2007; Sulem et al., 2007) and *TYR* (Izagirre et al., 2006; Lao et al., 2007; Norton et al., 2007; Sulem et al., 2007) have been found in Europe. Of these variants, only *OCA2* and *TYR* are associated with light skin colour in East Asia, and additional variants have been found (*ASIP*, *OCA2*, *DCT*, *EGFR* and *DRD2*) suggesting independent evolution of the trait in both continents (Lao et al., 2007; Norton et al., 2007). On the other hand, there are suggestions of signals of selection against light pigmentation in Africa, based on distribution of dark skin variants of *TP53BP1* and *RAD50* (Izagirre et al., 2006).

This diversity is explained by hypothetical selective pressures operating in both high and low levels of exposure to sunlight. The selective pressure is specific to each geographic region, favouring the production of adequate levels of sunlight filtering for a given latitude. Therefore, in places where UV radiation is high, pigmentation should protect from sunburn, dehydration, and degradation of micronutrients, while in places where it is low, pigmentation should allow biosynthesis of vitamin D (Jobling et al., 2013). Some studies have suggested selection for light skin colours at greater latitudes (Lamason et al., 2005; Lao et al., 2007; Norton et al., 2007), while others suggest selection for dark colours at lower latitudes (Harding et al., 2000), or both (Izagirre et al., 2006).

1.1.3.3 High altitude

Partial pressure of oxygen in air decreases with elevation, diminishing oxygen content in inhaled air at significant levels at altitudes above ~3,000 metres. Under these conditions, most humans can suffer symptoms of acute mountain sickness, to which respiratory physiology (evolved at sea level) responds by inducing hyperventilation and increasing basal metabolic rate, which can result in acclimatisation after some weeks of exposure (Beall, 2007). However, long–term sustainability

of these responses can result in lethal oedemas and chronic mountain sickness (Beall et al., 1997), and affect foetal and maternal mortality (Julian et al., 2009; Moore et al., 2011), thus constituting an important impediment to the colonisation of these territories.

However, settlements at high altitude have existed for a long time in the Himalayas, the Andes, and the Ethiopian plateau[3], suggesting genetic adaptation in the populations living there. Moreover, phenotypes of physiological adaptation to hypoxia are very different in the three regions. While arterial hypoxaemia is present in both Andean and Himalayan populations, decreased haemoglobin concentration is reported in the Himalayas, and increased haemoglobin with erythrocytosis is shown in the Andes. On the other hand, Ethiopian highlanders have similar levels of oxygen saturation, haemoglobin concentration, and arterial oxygen to those reported for sea level populations (Beall, 2006, 2007; Beall et al., 2002).

Signals of selection have been found in haplotypes of *EPAS1*, *EGLN1*, and *PPARA*, in Tibetan and Sherpa populations. These genes are all in the pathway of hypoxia-inducible transcription factor HIFα2, and are related to haemoglobin concentration, which is decreased in Himalayans, but not in other high altitude populations (Beall et al., 2010; Simonson et al., 2010).

Andean physiology at high altitude is similar to that shown by acclimatised lowlanders, but heritability rates of hypoxaemia and erythocytosis in Andeans born at sea level suggest a genetic basis (Beall, 2006). Some studies have found signals of selection in Andean populations for haplotypes of *ENDRA*, *PRKAA1*, and *NOS2A*, genes that are all involved in the HIF pathway through regulation of vasoconstriction, ATP-deprivation sensing, and blood pressure regulation respectively (Bigham et al., 2009).

1.1.3.4 Local pathogens

Humans have always lived with organisms that cause infectious or parasitic diseases. Distributions of pathogens are importantly associated with environmental factors (Guernier et al., 2004), thus having major consequences in human adaptation to new territories. The study of human adaptation to diseases has been of central importance in the history of the field of human genetics and evolution.

Adaptation to malaria is arguably the most comprehensively studied case of human adaptation. Malaria is responsible for very high mortality rates and it occurs in most subtropical areas in the world (Hedrick, 2011). These features mean that it has been a strong selective pressure in human evolution. The fact that several genetic red cell disorders are particularly prevalent in geographic

[3] Dates for the colonisation of these areas are disputed, but ranges are around 20,000 BP, 10,000 BP and 50,000 BP for Himalayas, Andean and Ethiopian highlands respectively (Alkorta-Aranburu et al., 2012; Beall, 2001)

regions where malaria is or was endemic, suggested the idea of an heterozygous advantage specific to these regions (Weatherall, 2008). For instance, a nonsynonymous substitution in the haemoglobin gene causes a abnormal variant of β–globin, known as HbS, which form chains of polymers after releasing oxygen, causing rigid and fragile sickle–shaped red blood cells that are regularly damaged producing anaemia. Allison (1954) interpreted the high prevalence of sickle-cell anaemia and the excess of heterozygotes in regions with malaria as a result of balancing selection due to a possible advantageous effect of the disorder conferring resistance to the infection. Further studies seem to corroborate this hypothesis, and additional haemoglobinopathies (e.g. other haemoglobin variants, α and β–thalassaemias), isoforms of Duffy antigen receptors, and disorders involved in metabolism of red cells (e.g. G6PD deficiency), have been found in areas affected with malaria (reviewed in Weatherall 2008 and Kwiatkowski 2005), showing strong signals of natural selection in some populations (Hamblin & Di Rienzo, 2000; Tishkoff et al., 2001). It is likely that all these variants modify properties of the erythrocyte that are exploited by the parasite and are crucial for its development, increasing destruction rates in the spleen (thalassaemias), reducing cytoadherence (Duffy antigen receptors), or increasing oxidative stress (G6PD deficiency). However, specific details about how these protective mechanisms work are not fully understood (Kwiatkowski, 2005).

There have more recently been similar findings in a variety of other diseases. A deletion affecting the structure of chemokine receptor CCR5 (*CCR5-*Δ*32*) impedes its attachment to the cell membrane. While chemokines play a role in leukocyte response to infection, it is believed that other chemokines can replace CCR5 function in the carriers of this mutation (de Silva & Stumpf, 2004). This allele, common only in European populations, has attracted attention because CCR5 acts as co-receptor of HIV, and carriers of *CCR5-*Δ*32* show resistance or delayed progression of AIDS (Martin, 1998). However, due to the short time for HIV for act as a selective pressure, it is speculated that selection for this trait has been the result of resistance to other diseases with historical importance, such as plague (Stephens et al., 1998) or smallpox (Galvani & Slatkin, 2003). There is an ongoing debate about the validity of these hypotheses, as similar frequencies to the observed today have been found in ancient DNA of 5,000 years ago, and genetic differentiation and linkage disequilibrium are not very strong in comparison to better-established cases of adaptation (Hedrick & Verrelli, 2006; Sabeti et al., 2005).

Mutations in the *CFTR* gene, which encodes Cl^- ion channels, cause cystic fibrosis in homozygotes. This disease reduces life expectancy significantly and produces infertility in males (Poolman & Galvani, 2007), but despite these evident detrimental effects, the most prevalent mutation, Δ*F508*, shows frequencies around 2% in European populations.

This high incidence has been explained as heterozygous advantage provided by resistance to Cl^-

secreting diarrhoea (Romeo et al., 1989), and hypotheses based on important episodes of cholera (Romeo et al., 1989) and typhoid fever (Pier et al., 1998) have been proposed as the selective pressures promoting current frequencies. The hypothesis of typhoid fever is supported by decreased susceptibility to *Salmonella typhi* in heterozygous mouse models (Pier et al., 1998). These suggestions have been criticised on several grounds, such as ambiguous signals of positive selection (Wiuf, 2001) and the absence of this variant in regions with higher prevalence of typhus or cholera than Europe (Mateu et al., 2002; Poolman & Galvani, 2007). As an alternative, Poolman & Galvani (2007) have proposed resistance to tuberculosis as the selective pressure promoting *CFTR-ΔF508* heterozygosity, based on the historical importance of pandemic tuberculosis in Europe. Further research is needed to corroborate this hypothesis.

Cultural practices have also been involved in exposing humans to different diseases causing adaptation, as exemplified by the role of ritual cannibalism in the spread of resistance to kuru in the Fore of Papua New Guinea (Lindenbaum, 2008). Kuru is a contagious neurodegenerative disease produced by self-replication of abnormally folded protease-resistance proteins (PRNP). Heterozygosity at codon 129 of *PRNP* confers resistance to both kuru and Creutzfeldt–Jakob disease (Mead et al., 2009). Among the Fore, selection favouring heterozygous has been shown through differences in deviations from Hardy–Weinberg equilibrium between different age and sex groups (Mead et al., 2009). Exposure to brain consumption was common only before the ban of cannibalism in 1950, and mainly among females and children, because males had little participation in mortuary feasts and they normally consumed low-risk tissues (Mead et al., 2008). Therefore, a deep knowledge of the history and cultural features of Fore cannibalism was needed to detect the relevance of this comparison. The effect of this practice on genetic adaptation was large enough to allow detection of natural selection using a method as direct as comparing Hardy–Weinberg equilibrium between generations (Mead et al., 2008) and constitutes one of the best examples of fast evolutionary change in humans. Although not as strong as in Papua New Guinea, allele frequencies at STRs linked to *PRNP* variants show signals of strong natural selection in most world populations, with the exception of East Asia (Mead et al., 2003). This could be interpreted as an adaptation to ancient contagion of prion diseases through animal vectors or evidence of resorting to cannibalism in the history of most human populations, and therefore an adaptation to a dietary practice (Mead et al., 2003).

1.1.4 Human Adaptation to Dietary Behaviour

Because different resources are available in different environments, changes in diet can be considered as part of the set of adaptations to different environments discussed above. This includes

cultural adaptations such as avoidance and preference towards certain foods as well as food processing (Harris, 1998; Simoons, 1994), and biological adaptations to particular dietary habits exemplified below[4].

1.1.4.1 Alcoholic beverages

Alcohol dehydrogenases (ADH) and aldehyde dehydrogenases (ALDH) are enzymes involved in alcohol metabolism in many organisms including humans (Mulligan et al., 2003). Gene families encoding both ADH and ALDH are highly variable in all populations, but a few variants with effects on enzymatic activity at genes *ALDH2*, *ADH1B* and *ADH7* are exclusively present at high frequencies in East Asian populations, where signals of strong natural selection have been detected (Evsyukov & Ivanov, 2013; Li et al., 2009; Oota et al., 2004; Osier et al., 2004, 2002; Peng et al., 2010). The strong link of this variants with alcohol abstinence and low alcohol consumption in East Asians has led to some researchers (Chen et al., 1999; Tu & Israel, 1995) to suggest provision of protection against alcoholism as the cause of their high frequencies.

The mechanisms acting as evolutionary advantages for this trait are a matter of debate. Alcoholism is associated with a variety of diseases that can be life–threatening, and the flush reaction suffered associated with alcohol consumption in the carriers of these variants could have discouraged alcohol consumption (Osier et al., 2002; Tu & Israel, 1995). In addition, a particular combination of non–protective variants in Native Americans is highly correlated with binge drinking and alcoholism (Mulligan et al., 2003). Studying several Chinese populations, Peng et al. (2010) have found a marked geographic distribution of ADH1B*47*His* in association with locations of early rice domestication, which would have been used in production of alcoholic beverages. The study dated the origin of the ADH1B*Arg*47*His* polymorphism between 10,000 and 7,000 years ago, in agreement with dates of the origin and expansion of agriculture in China. However, other studies have proposed causes not related to alcoholism, highlighting the role of ADH and ALDH in protection against mycotoxins and anaerobic parasites as more likely causes for the selection of these variants (Han et al., 2007; Oota et al., 2004).

1.1.4.2 Starch and meat

An important determinant in components of diet is given by subsistence strategies in different populations, and the changes in diet promoted by the invention of agriculture are likely to have had an adaptive effect. Copy number variation of the salivary amylase gene (*AMY1*) between different populations have been interpreted as adaptation to the diet rich in starchy foods such as roots, tubers, wheat and rice (Luca et al., 2010; Perry et al., 2007).

[4] Adaptation to milk drinking is not included in this list, but is discussed in detail in section 1.2.

Salivary amylase is involved in starch digestion, and number of copies of the *AMY1* gene are related with levels of amylase production (Fried et al., 1987; Groot et al., 1989; Lebenthal, 1987). Perry et al. (2007) show a strong association between *AMY1* copy numbers in different populations according to the importance of starchy foodstuffs in their diet, suggesting selection as the cause of this distribution.

A polymorphism in the *AGXT* gene (alanine:glyoxylate aminotransferase) is strongly related with primary hyperhoxaluria, a lethal disease; however, this variant is present in high frequencies in some populations. Caldwell et al. (2004) have suggested an explanation for the frequencies of this variant as adaptation to a diet rich in meat. In other mammals, herbivorous and carnivorous diets differ in whether glycoxylate synthesis occurs in the peroxisome or the mitochondria respectively (Birdsey et al., 2004). In humans, AGT normally targets peroxisomes, but this variant is known to produce targeting of AGT to the mitochondria. Caldwell et al. (2004) presented frequencies of this variant at two extremes between Saami and Han Chinese, as supporting their hypothesis. However, another study (Ségurel et al., 2010) could not find this distribution in relation to subsistence or meat consumption using a larger dataset.

1.1.4.3 Taste perception

Sensory response to food has been hypothesised as an adaptation to encourage or dissuade consumption of advantageous or pernicious substances (Jobling et al., 2013; Luca et al., 2010). Since the discovery of variability in the ability to taste phenylthiocarbamide (PTC) (Blakeslee, 1932; Wooding, 2006), research has been focused mainly on the evolution of taste. PTC is an artificial substance that is perceived as either bitter or tasteless by different subjects, and though it is not present in nature, PTC perception is related to the ability to taste other bitter substances (Wooding et al., 2004). The exploration of the genetic causes of PTC perception variability led to the discovery of the *T2R* family of genes, all related to bitter taste perception (Chandrashekar et al., 2000; Shi et al., 2003) and showing deviations from neutrality in many populations (Fischer et al., 2005; Soranzo et al., 2005; Wooding et al., 2004). The main hypothesis to explain these deviations is adaptation to prevent consumption of toxic substances (Chandrashekar et al., 2000; Fischer et al., 2005; Soranzo et al., 2005; Wooding et al., 2004).

There is no agreement on whether this is a variability inherited from adaptations in earlier points of mammal evolution, or an ongoing adaptation to humans diversification in diet. A selective pressure based on avoidance of toxic elements would have applied to other species as well and could be result of an earlier evolutionary process. However, there is evidence of diversity of bitter receptors in mammals (Shi et al., 2003) suggesting parallel evolution, and the genetic basis of bitter taste

perception are not the same in humans and other apes (Fischer et al., 2005; Wooding et al., 2006), suggesting continuous adaptation to bitter taste after the divergence of humans and chimpanzees.

Nevertheless, the evolutionary importance of bitter tasting for survival in modern humans is still discussed. Analysis of Neanderthal DNA suggest variability in *T2R38* (Lalueza-Fox et al., 2009), and relaxation of selective pressures over bitter taste perception are likely to have occurred in recent human evolution (Wang et al., 2004). Alternatives to explain recent evolution of this trait not related to bitterness perception, but to disease susceptibility, have also been proposed (Campbell et al., 2012; Glendinning, 1994; Lee et al., 2012).

The evolution of other tastes has not been as deeply studied. Some candidates have been identified for sweet (Fushan et al., 2010) and umami (Shigemura et al., 2009), both members of the *T1R* family. However, gustatory perception of other substances is still unclear (Bachmanov & Beauchamp, 2006).

1.2 Lactase Persistence

1.2.1 Lactase and Lactose Digestion

Lactose, the main sugar in milk, is a disaccharide that cannot be absorbed from the intestine into the bloodstream until broken down into its constituent monosaccharides: glucose and galactose. This process is catalysed by lactase, an enzyme located in the brush border of the enterocytes in the small intestinal epithelium. In lactose maldigesters, lactase activity is low and lactose passes intact through the small intestine into the colon. Once there, lactose acts as a fermentation substrate for gut bacteria, releasing gas, lowering the pH inside the lumen of the colon, reducing water absorption, and in some cases producing osmotic diarrhoea (Swagerty et al., 2002). In some people this occurs with other gastric symptoms ranging from a mild stomach ache to nausea and vomiting, and hence is referred to as lactose intolerance.

Lactase deficiency can be caused by a very rare recessive disorder (congenital alactasia), but the most common type of deficiency is caused by down-regulation of lactase after weaning. This process is part of the normal development in other mammals and in most humans (around 65% according to Itan et al., 2010), and is not always related with digestive discomfort (Flatz, 1987; Ingram et al., 2009b). Lactase activity persists at high levels in only about 35% of humans (termed lactase persistence), allowing them to drink milk in adulthood without adverse symptoms.

Although they are commonly used as synonyms, the terms lactose maldigestion, lactose intolerance and lactase non-persistence have different meanings. Lactose maldigestion is the lack of capacity to break down the lactose molecule, while lactose intolerance refers to the digestive symptoms associated with maldigestion (which are not present in all maldigesters as mentioned before). Lactase non-persistence is one of the possible causes of lactose maldigestion (Levitt et al., 2013), which can also be due to damage of the intestinal epithelium caused by other digestive problems (secondary lactose intolerance).

Direct methods to measure lactase activity are needed to determine lactase non-persistence but require a sample of intestinal tissue from a biopsy. However, indirect methods based on evaluation of lactose digestion through measurements of levels of blood glucose and concentration of breath hydrogen can be used to infer lactase persistence status, and are less invasive and suitable for population studies. After administration of a dose of lactose to a fasting volunteer, levels of blood glucose increase in lactose digesters but not in non-digesters (due to their capacity to obtain glucose from lactose), and the concentration of breath hydrogen increases in non-digesters but not in digesters (as hydrogen is a product of the bacterial fermentation of undigested lactose, providing

hydrogen–producing bacteria are present in the colon). Error rates of 7% false non–persistent and 9% false persistent are estimated for the blood glucose test, and error rates of 5% false non–persistent and 7% false persistent are estimated for the breath hydrogen test (Mulcare et al., 2004), suggesting the latter to be more accurate. Taking into account the error underlying these indirect methods, it is clear that subjects who are truly persistent may nevertheless feel unwell with milk due to other gastric problems or psychosomatic effects (Briet et al., 1997; Peuhkuri & Vapaatalo, 2000), and subjects who are non–persistent may be able to consume lactose without adverse effects especially if they are drinking milk in reduced quantities, or together with other foods, or consuming processed milk products such as cheese or yoghurt, which contain less lactose, or have a gut flora adapted to milk consumption (Levitt et al., 2013). There is also some confusion with milk protein allergies, another adverse reaction to milk consumption but with very different causes (Crittenden & Bennett, 2005; Ingram et al., 2009a).

Curiously, and despite all the circumstances mentioned under which lactase non–persistent populations can include milk in their diet, lactase persistence has a remarkable geographic and ethnic distribution associated with milk usage, which has been interpreted as an important genetic adaptation for populations with high dietary dependence on milk. This distribution has been the subject of research for more than 50 years (Auricchio et al., 1963; Simoons, 1969, 1970; Swallow, 2003) examining its origins and significance, and provides one of the best examples of the role of culture in generating human biological diversity.

1.2.2 Genetics of Lactase Persistence

Contrary to common belief, lactase persistence is not induced by a diet rich in milk (Leichter, 1973) but is genetically inherited as a dominant trait, as has been shown in family studies (Sahi, 1974; Sahi & Launiala, 1977). Lactase is encoded in the lactase gene (*LCT*), a sequence of ~50 kb located on chromosome 2q21 (Kruse et al., 1988). The enzyme is synthesised as a precursor which is further processed in the endoplasmic reticulum and the Golgi apparatus to become a mature protein, subsequently anchored to the intestinal epithelium (Swallow, 2003).

There are no structural differences in the lactase protein of persistent and non–persistent individuals (Boll et al., 1991) and single polymorphisms in *LCT* and its adjacent promoter region do not cause lactase persistence status (Ingram et al., 2009a). However, combinations of these variants were used to identify four common core haplotypes (A, B, C and U), differentially distributed across populations, with core haplotype A being highly abundant and associated with lactase persistence in Europe (Harvey et al., 1998; Hollox et al., 2001).

Further analyses of this haplotype in the *5′* direction revealed a very large (~1 Mb) undisrupted

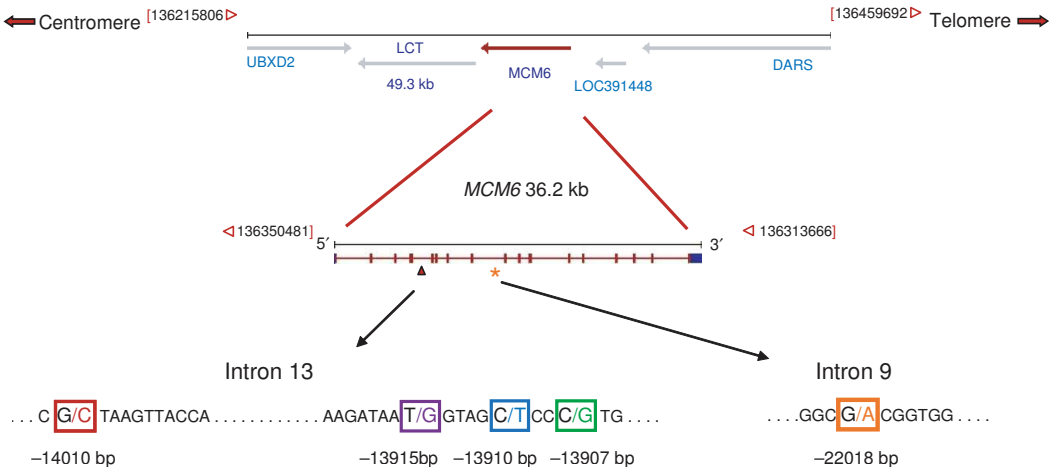

Figure 1.1. Chromosomal context of *LCT* and *MCM6*, showing the *LCT* enhancer region in *MCM6* and the first variants associated with lactase persistence in European and some African populations (Adapted from Tishkoff et al. 2007 by permission from Macmillan Publishers Ltd: Nature Genetics copyright ©2007 Nature Publishing Group. Copyright clearance can be found in Appendix A).

sequence (Poulter et al., 2003), and the association of lactase persistence with a SNP (transition *C* > *T*) located ~14 kb upstream of *LCT* in Finnish families (Enattah et al., 2002). This discovery was followed by the demonstration *in vitro* of the enhancing effect of this allele in the activity of the lactase promoter (Olds & Sibley, 2003), apparently completing the picture of lactase persistence as caused by a *cis*–acting enhancer region of *LCT* located within an intron of the adjacent *MCM6* gene (Olds & Sibley, 2003; Troelsen, 2005). However, while this single variant (*−13,910*T*) is highly associated with lactase persistence in European populations (Enattah et al., 2002), it is not causal of the high levels of lactase persistence in some non–European groups.

Other variants (*−13,907*G*, *−13,915*G*, *−14,010*C*, *−14,009*G*) have been found to explain high frequencies in some African and Middle–eastern populations, some of them with evidence of association with long undisrupted haplotypic backgrounds (Enattah et al. 2008; Imtiaz et al. 2007; Ingram et al. 2007; Jones et al. 2013; Tishkoff et al. 2007, see Figure 1.1). Nevertheless it has been reported that the known variants are not enough to fully account for the worldwide distribution of lactase persistence frequencies (and lactose digesters have been identified who do not carry any of the known mutations according to Ingram et al. 2009b; Jones et al. 2013; Ranciaro et al. 2014). The existence of other undiscovered variants is likely (Itan et al., 2010)

1.2.3 Distribution of Lactase Persistence in the Old World and its link with Pastoralism

The analysis of the frequencies of lactase persistence in the Old World shows a noticeable pattern of geographic distribution, strongly associated with populations that have or have had pastoralist

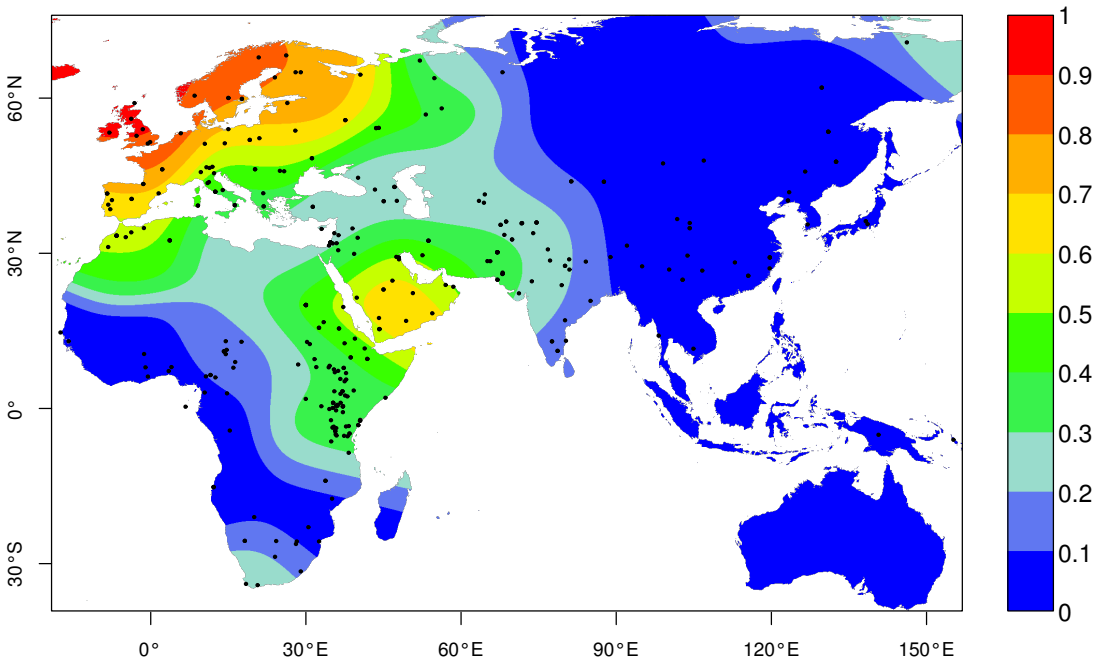

Figure 1.2. Distribution of lactase persistence in the Old World estimated using kernel density interpolation as in Ingram et al. 2009a. Interpolation was inferred from all 5 known functional alleles (*−13,907*G*, *−13,910*T*, *−13,915*G*, *−14,009*G*, *−14,010*C*) using the data from the Global Lactase Association Database (GLAD) updated in 2014 by Anke Liebert, using all recently collected data (Gallego Romero et al., 2012; Ranciaro et al., 2014).

subsistence strategies (see Figure 1.2). This distribution shows low frequencies of lactase persistence in most populations, with important exceptions in most of northern Europe, with the highest levels of lactase persistence in Scandinavia, gradually decreasing towards South-eastern latitudes; and an ethnic distribution in East Africa and the Middle East, with groups with high levels of lactase persistence scattered among groups with very low frequencies.

The association between high frequencies of lactose digestion and the custom of drinking milk from other mammals was pointed out in early studies of lactase persistence (Simoons, 1978, 1970). The custom of drinking milk from other mammals is a practice developed only by few pastoralist groups during the Neolithic, but was absent in most populations until recently (Harris, 1998). Table 1.1 shows frequencies of lactose tolerance (measured by either breath hydrogen or blood glucose) in different populations grouped according to their subsistence pattern (Simoons, 1978). The Table shows high frequencies of lactase persistence only in populations with historical dependency on milk and dairy products. However, this relationship does not hold in the other direction: there are some groups with a remarkable tradition of milking, and high levels of milk consumption, but low levels of lactase persistence, such as the Mongolians (Strickland, 1993; Tserendolgor et al., 1998) or Greek pastoralists (Kanaghinis et al., 1974). The adaptation to a milk–rich diet can also be achieved by other mechanisms: adaptation of milk processing methods to produce cheeses and fermented products reduces the amounts of lactose (and thus is a cultural adaptation), while there

is also evidence of physiological adaptation, which probably operates by adaptation of intestinal bacteria (Swallow, 2003).

Population	Prevalence of Lactose maldigestion
Hunter and Gathering Peoples	
Greenland Inuit	85
Bushmen of Southern Africa	95
Other	82
Non–milking agriculturalist (and unmixed overseas descendants)	
Native Americans from Southwestern U.S.	97
Yoruba and Ibo	90
Thai	98
Chinese and South-east Asia	98
Java Indonesians	82
Other	84
Recent milk–users	
Bantu and related groups	88
Traditional Milk user (and unmixed overseas descendants)	
Nomadic Fulani	22
Bedouin and Urban Saudi Arabs	14
Finns	15
American Whites	19
Czechs in Canada	8
Five groups in India and Pakistan	18
Other	6
Mixed groups of milking/non–milking ancestry	
Hausa/Fulani	67
Yoruba/European	44
American Blacks	66
Amerindian/Anglo–American	51
Mexican/Americans	53
Thai/Northwest Europe	50
Other	54

Table 1.1. Prevalence of Lactose maldigestion in different populations grouped by subsistence pattern. (Adapted from Simoons 1978 by permission from Springer Science and Business Media: The American Journal of Digestive Diseases copyright ©1978 Springer. Copyright clearance can be found in Appendix A).

1.2.4 Evolution of Lactase Persistence

The pattern of distribution of lactase persistence led to the idea that this adaptation had been spread by selection but exactly how this operated was much less clear. Early studies proposed three main hypotheses, known as the culture–historical hypothesis, the solar radiation hypothesis, and the dry environment hypothesis. The culture–historical hypothesis (Simoons, 1970) explains positive selection of lactase persistence as a result of a general nutritional advantage of being able to drink milk and digest it properly. As one of the first proposals, Simoons analysed the distribution of lactase persistence not only from a geographic standpoint, but also in relation to cultural

practices, subsistence patterns, economical dependence on livestock, and milk consumption. He proposed that lactase persistence was selected wherever human groups had livestock suitable for milking, in addition to difficult access to the nutrients available in milk.

The dry environment hypothesis (Cook & Al-Torki, 1975), explains effectively the high frequencies of lactase persistence among groups in low latitudes, for instance, Yöruk, Beja, Tuareg and Jordan Bedouin. According to this hypothesis, lactose tolerance was selected in zones with limited access to water. Individuals able to digest milk could take advantage of its water content. In comparison, lactose maldigesters affected by diarrhoea after drinking milk would increase their risk of dehydration. Despite the fact that some North African groups use their camels' milk as a source of water nowadays (Holden & Mace, 1997), it is not likely that hydration alone would explain the higher frequencies of this trait in Northern Europe, unless we take into account the hypothesis of contaminated water as source of infections.

The solar radiation hypothesis (Flatz & Rotthauwe, 1973) suggests that lactase persistence and fair skin pigmentation were selected in Northern Europe as an adaptation to high latitudes with low solar radiation. The lack of sunshine made difficult the synthesis of vitamin D, increasing the risk of rickets. A decrease in the amounts of melanin would allow the synthesis of vitamin D in the skin despite low levels of solar radiation, and lactase persistence would allow calcium uptake from fresh milk, preventing rickets. This hypothesis explains effectively the selection of this trait in Northern Europe, but does not explain why some groups outside Europe have high levels of lactase persistence.

There are problems with these three selection models: the culture–historical hypothesis is challenged by the existence of geographic areas with extended milking practices and low levels of lactase persistence (e.g. Southern Italy and Greece, Central Asia), and also by the possibility of nutritional benefits of fresh milk not fully replaced by milk consumption of other milk products (Almon et al., 2010; Greenwald et al., 1963). Aridity is not related to distribution of lactase persistence in Africa controlling for genealogical relationships using the comparative method (Holden & Mace, 1997), and computer simulations based on large databases of lactase persistence frequencies in Europe do not require to take into account latitude to reach current frequencies (Itan et al., 2009). A fourth approach has been developed based on the ideas of the culture–historical hypothesis applying the framework of the gene–culture co–evolution theory (Aoki, 1987, 2001). Though this approach provides insights into the relationship of lactase persistence and dairying, it does not address the problems regarding lactase persistence being selected specifically in these populations and why it is geographically continuous in Europe but disrupted in Africa.

With the hypotheses mentioned above, other alternatives have been proposed: Current frequencies of lactase persistence reached by demographic effects and without selection (Brines, 2004), adoption of milking practice among those populations who were already lactase–persistent (Burger et al., 2007; Nei & Saitou, 1986; Simoons, 1970), and selection favouring lactase non-persistence in Africa (Anderson & Vullo 1994, reviewed in Ingram et al. 2009a). However because of the extensive evidence showing positive natural selection favouring lactase persistence, none of these alternatives have much credence, but are mentioned here for the sake of completeness.

Gerbault et al. (2009) evaluated the importance of pastoralism and latitude to explain distribution of lactase persistence using a dataset of frequencies of persistence in relation to other genetic traits, archaeological data on domestication dates, and computer simulations. They concluded that pastoralism could explain distribution of frequencies in Africa, but that the scenario was less clear in Europe, where genetic drift alone can explain frequencies in south-east Europe, but is not enough to account for the high frequencies in the north-west.

Part of their discussion is of particular importance for the present research: In south-west Europe frequencies of lactase persistence are higher than would be expected by their latitude, though lower than the expected by importance of pastoralism (Gerbault et al., 2009). This was indeed pointed out in an earlier study (Sahi, 1994), but attributed to questionable methods and when only one publication was available. A recent study (Sverrisdóttir et al., 2014) could not find the $-13{,}910^{*}T$ variant in neolithic Iberian remains (in agreement with the findings by Burger et al. 2007 for northern and eastern Europe), concluding from computational simulations that current frequencies cannot be reached without strong selection, despite the high solar radiation at this latitude. This will be discussed in Section 1.2.5.

Regardless of its causes and origins, the observations on lactase persistence draw attention to the importance that milk consumption must have had in the evolutionary history of different independent groups. As has been shown by Jones et al. (2013), the differences in the pattern of distribution of lactase persistence between Europe and Africa can be explained as result of selection acting on a larger starting population of more diverse genetic background in Africa, causing a soft selective sweep (i.e. positive selection acted on previous standing variation), than on an expanding homogeneous one in Europe, resulting in the better known hard sweep (i.e. positive selection acted on a new variant). Lactase persistence is not only unevenly distributed worldwide, but it is also strongly linked to pastoralism, and it is caused by multiple genetic variants leading to the same trait through a process of convergent or parallel evolution (Enattah et al., 2007; Ingram et al., 2009b; Jones et al., 2013; Tishkoff et al., 2007). At least three of these variants are located on extended undisrupted haplotypes (Bersaglieri et al., 2004; Ranciaro et al., 2014; Tishkoff et al.,

2007), that were absent in early Neolithic populations (Burger et al., 2007), and have reached current frequencies in the period of 5,000 – 10,000 years since animal domestication (Evershed et al., 2008; Itan et al., 2009). These data support the hypothesis of strong positive natural selection favouring lactase persistence in those groups who developed milking practices soon after animal domestication.

It would appear then that to have reached current frequencies in the allotted time, milk consumption must have been decisive in terms of survival and/or fertility (Aoki, 1986; Brines, 2004; Ingram et al., 2009a). If this is really the case, it may be possible to observe this selection in action in living people. In this project, we aim to see whether any evidence for differential weight, height, fertility and survival can be seen in an agro–pastoralist population with heterogeneous persistence status due to mixed ancestry, under constant threat of environmental stress, and high dependency on livestock and dietary milk.

1.2.5 Lactase Persistence in Latin American Mixed Groups and its Parental Populations

Studies of lactase persistence in Latin American populations are scant and unsuitable for our purposes: The samples are usually from hospital patients who are likely to be lactose intolerant (either due to non–persistence or secondary loss of lactase) (e.g. Bulhões et al., 2007; Escoboza et al., 2004; Morales et al., 2011; Teves et al., 2001), the studies usually include people of all ages (e.g. Escoboza et al., 2004; Lacassie et al., 1978; Lisker et al., 1974; Pretto et al., 2002), and their blood glucose (BG) and hydrogen breath (HB) lactose tolerance tests were often performed with non–standard doses of lactose or with fresh milk (e.g. Pretto et al., 2002; Rosado et al., 1994a,b; Vettorazzi et al., 1992). Published data of frequencies of lactase digester status are shown in Table 1.2. The inclusion criteria presented by Itan et al. (2010) were employed, and studies based on hospital patients suspected of having lactose intolerance, children under 12, or family members were excluded [5].

An interpolation map of the distribution of the frequencies of lactase persistence in Table 1.2 is shown in Figure 1.3. This map is based on the design of Figure 1.2 (which is based on the map presented in Ingram et al. 2009a) and uses the same scale and colour scheme to allow comparison with the Old World, resulting in a more homogeneous scenario of frequencies ranging from mid-low to mid levels and consistent with the model of admixture distinctive of this region (mainly Native American and Iberian, see Table 1.3). Differences between the individual countries could be caused by differences in the historic admixture, the contribution from African populations, and the

[5] It is worth mentioning studies on frequencies of lactase persistence in mixed Mexican-Americans in the United States (Dill et al., 1972; Sowers & Winterfeldt, 1975; Woteki et al., 1977), reporting results ranging from 45 to 53% of lactose digesters based on blood glucose. These were not included due to their geographic location.

Country and Population	Digest. (%)	N	Test	Reference
Brazil				
Whites Porto Alegre	50	20	HB	Bulhões et al. 2007
Whites São Paulo	33	9	Biopsy	Escoboza et al. 2004
Non-whites São Paulo	0	5	Biopsy	Escoboza et al. 2004
Whites Campinas	55	40	BG	Sevá-Pereira et al. 1983
Black Campinas	15	20	BG	Sevá-Pereira et al. 1983
Chile				
Santiago Mixed (army recruits)	24	98	BG	Lacassie et al. 1978
Santiago Mixed (army cadets)	36	97	BG	Lacassie et al. 1978
Santiago Mixed (hospital patients*)	42	51	HB	Morales et al. 2011
Colombia				
Bogotá (university students)	44	98	HB	Ángel et al. 2005
Mexico				
Huamantla Mixed (high school students*)	29	193	BG	Lisker et al. 1974
Huamantla Mixed (factory workers)	17	100	BG	Lisker et al. 1974
Inxtenco (high school students*)	23	108	BG	Lisker et al. 1974
Peru				
Lima (hospital patients*)	0	9	BG	Paige et al. 1972
Lima (hospital staff)	33	30	BG	Figueroa et al. 1971
Lima (university students)	37	44	BG	Calderón 1971
Uruguay				
Montevideo (hospital patients*)	47	109	HB	Maggi et al. 1987
Guatemala				
Guatemala City (hospital patients*)	48	46	HB	Vettorazzi et al. 1992

Table 1.2. Percentages of lactose digesters in some Latin American mixed Populations.*: High school students over 13 years. Hospital patients were not referred for gastric symptoms.

continuity of migration from Europe (particularly in the Atlantic coast), and factors such as drift and inbreeding, which could have had an important effect in small and isolated populations in some regions.

Studies reporting frequencies of lactase persistence based on lactose tolerance tests in Spanish and Native American populations are shown in Table 1.3. As in the previous table, studies including young subjects (e.g. Carnicer, 1993; Leichter & Lee, 1971; Vázquez et al., 1975), subjects referred for gastric symptoms (e.g. Casellas & Malagelada, 2003; Casellas et al., 2009), or of unclear methods for determining ethnicity (e.g. Morales et al., 2011; Murthy & Haworth, 1970) were excluded.

Fewer studies have been done based on genotype frequencies. Table 1.4 reports results of $-13{,}910C/T$ genotype frequencies in Latin America and Spain. Two of these studies (Bulhões et al., 2007; Morales et al., 2011) have reported significant association ($p < 0.001$) between $-13{,}910^*T$ and negative breath hydrogen in Brazilian and Chilean mixed populations, and three

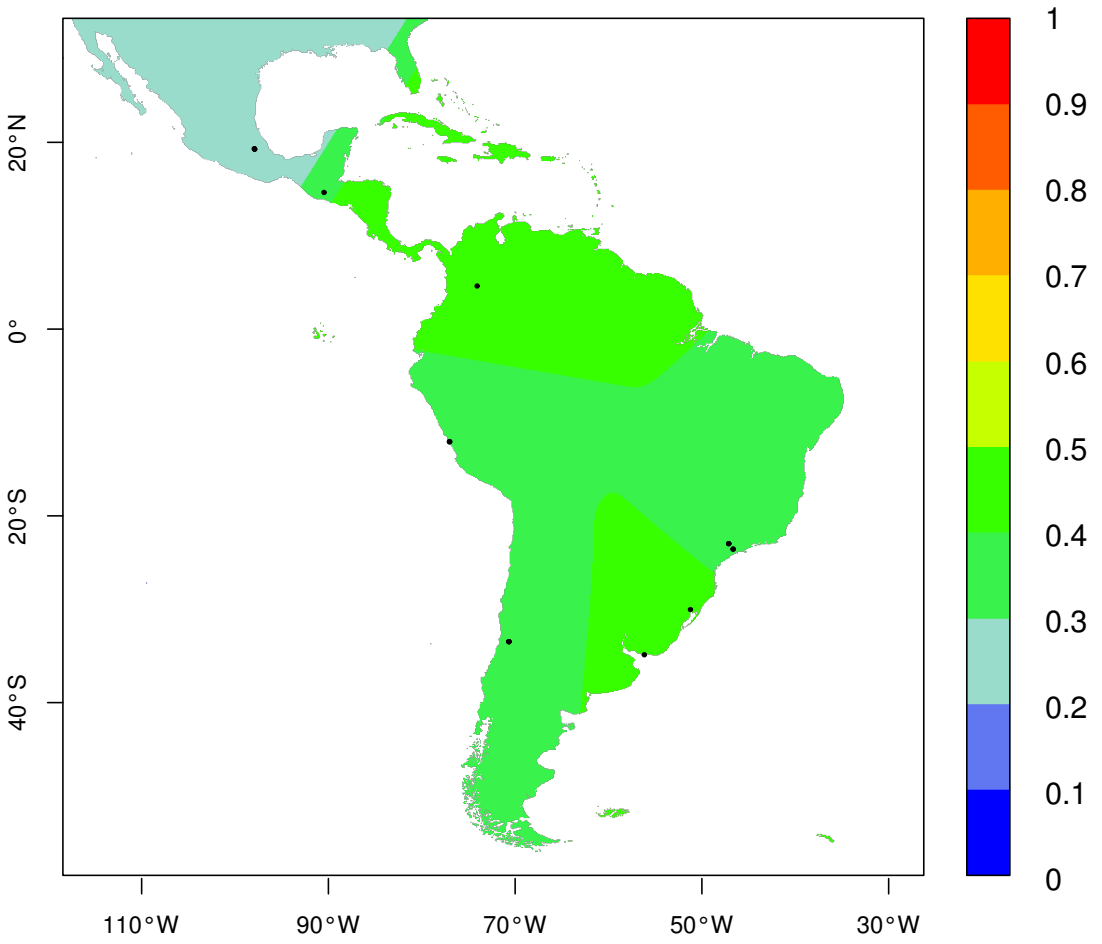

Figure 1.3. Distribution of lactase persistence in Latin America (mainly urban) estimated using kernel density interpolation as in Ingram et al. 2009a. Interpolation was inferred from literature reporting results of lactose tolerance tests in Table 1.2. The scale and colour scheme used in Figure 1.2 has been retained to allow comparison.

have confirmed its status in a block of linkage disequilibrium with *−22,018G/A* (Bulhões et al., 2007; Friedrich et al., 2012b; Morales et al., 2011). Both associations are distinctive features of the European variant of lactase persistence.

Despite the scarcity of studies, the results are consistent with the idea that lactase persistence was absent in Native American populations until its introduction by Europeans, and frequencies of lactase persistence in Spain are low in comparison with Northern Europe, but high compared with both indigenous and mixed populations in Latin America. The influence of populations other than Native Americans and Spanish to Chilean historical demographic structure is not clear and is the subject of discussion among historians (Lizcano, 2005; Medina & Kaempffer, 1979). African ancestry is a major demographic component in many Latin American countries (Lizcano, 2005), but has traditionally been underestimated in Chile. According to Medina & Kaempffer (1979) *"The black population (in Chile) was always scant, reaching a peak of 25,000 during colonial periods. (...) its final contribution to the race is not over 1%"*.

Country and Population	Digest	N	Testing Method	Reference
Colombia				
Chami	0	24	BG	Alzate et al. 1969
United States				
Great Basin Natives	8	100	BG	Johnson et al. 1978
Oklahoma Natives	7	35	BG	Bose & Welsh 1973
Oklahoma Natives	8	40	HB	Caskey et al. 1977
Spain				
Galicia	65	546	HB	Leis et al. 1997
Valencia*	72	119	Unknown	Guix García et al. 1974 (from Sahi 1994)
Southern Spain*	81	265	Unknown	Yanez et al. 1971 (from Bloom & Sherman 2005)

Table 1.3. Percentages of lactose digestion in Amerindian and Spanish Populations. *: Data cited in other sources. Testing method not specified in reference.

In contrast to historical sources, Acuña et al. (2000) have estimated 5 to 10% of African ancestry in a study of the genetic structure of the populations in the main three valleys of the Region of Coquimbo by analyses of phenotype frequencies of 7 blood group systems, one plasma protein, 6 erythrocyte enzymes, and 3 HLA groups. All valleys have the haptoglobin phenotype 2–1 and high frequencies of Dce haplotype in the Rh blood system, both used as markers of African ancestry.

This study was conducted in the same geographic area studied in this thesis (to be introduced in section 1.3), and thus constitutes an important antecedent. However, because most African slavery in the Americas came from West Africa, it is reasonable to think that a greater contribution of non–European individuals to the foundation of the populations of goat herders in the Coquimbo region would lead to an increased proportion of non-digesters, in the absence of other factors, such as selection. Nevertheless, few cases of variants others than *−13,910*T* have been reported in Brazilian Afro-descendant populations (*−13,779*C* n=2, *−13,937*A* n=2, *−14,010*C* n=1, *−14,011*T* n=4, while *−13,910*C* n=360 [Friedrich et al. 2012b]), though the role of some of these variants on lactase persistence has not been completely established, and the magnitude of the African contribution in Brazil is much larger than that in Chile.

On the other hand, other populations might have left their genetic mark on the Spanish parental population prior to the migration event, and this might have affected gene frequencies differentially in Spain and in the Americas after contact due to demographic effects.

As commented in Section 1.2.4, frequencies of lactase persistence are higher in the Iberian Peninsula than in other parts of Southern Europe. With the values for Spain, presented in Tables 1.3 and 1.4, in Portugal frequencies of persistence range from 47.7% to 67.2% predicted digesters from

Country and Population	Pred. Digest.	*−13,910*T %*	N	Reference
Brazil				
White Porto Alegre Mixed	51	30	337	Friedrich et al. 2012b
Black Porto Alegre Mixed	32	18	182	Friedrich et al. 2012b
Belem Mixed	31	18	200	Friedrich et al. 2012b
Recife Mixed	37	20	262	Friedrich et al. 2012b
Guarani-Kaiowá (Amerindian)	1.2	0.6	84	Friedrich et al. 2012a
Guarani-Ñandeva (Amerindian)	15	7.6	59	Friedrich et al. 2012a
Kaingang (Amerindian)	8	4.6	72	Friedrich et al. 2012a
Xavante (Amerindian)	0.9	0.5	101	Friedrich et al. 2012a
Chile				
Santiago Mixed	43	22	216	Morales et al. 2011
Urban Mapuche (Amerindian)	12	6	43	Morales et al. 2011
Rural Mapuche (Amerindian)	10	5	29	Fernández & Flores 2014
Colombia				
Medellin Mixed	48	31	94	1,000 Genomes 2012
Peru				
Lima Mixed	19	11	85	1,000 Genomes 2012
United States				
Mexican in Los Angeles	42	24	64	1,000 Genomes 2012
Puerto Rico	39	21	104	1,000 Genomes 2012
Spain				
Canary Islands	62	36	551	Almon et al. 2010
Barcelona	60	39	944	Agueda et al. 2010
Valencia	60	39	940	Corella et al. 2011
Catalonia	68	45	163	Rasinperä et al. 2005
Spain	69	46	107	1,000 Genomes 2012

Table 1.4. Percentage of predicted digesters according to *−13,910*T* genotype in Latin American and Spanish populations.

*−13,910*T* (with a frequency of *−13,910*T* from 0.27 to 0.38) (Coelho et al., 2005; Manco et al., 2013). In contrast, places at the same latitude in Greece show values ranging from 16% of predicted digesters from *−13,910*T* (*−13,910*T* allele 0.09), to 25% of digesters according to blood glucose test (Anagnostou et al., 2009; Ladas et al., 1982). At Italy frquencies range from 15.3% of digesters in the south to 42.9% in the north predicted from *−13,910C/T* (with *−13,910*T* allele from 0.08 to 0.24, and lactose tolerance test between 14% and 49%) (Anagnostou et al., 2009; Burgio et al., 1984; Coelho et al., 2005; Meloni et al., 2001).

The relatively high frequency of lactase persistence in the Iberian peninsula has been pointed out in several other studies (Gerbault et al., 2009; Itan et al., 2009; Sahi, 1994). Moreover, the population history and migration patterns between Africa and the Iberian peninsula could explain both anomalous levels of genetic variants on the continental context (Bertranpetit & Cavalli-Sforza,

1991; Flores et al., 2004) and the degree of admixture with populations from North–east Africa and the Middle East (Adams et al., 2008). It is tempting to think about the influence of populations of Arab origin in the Iberian Peninsula between the 7th and the 15th centuries as another source of lactase persistence associated variants. In Berber nomads from North Africa, frequencies of LP range from 0.14 in Amizmiz Shilha (Mulcare et al., 2004; Myles et al., 2005) to 0.41 in Saharawi pastoralists (Enattah et al., 2008), accounted by *−13,915*G*, *−13,910*T*, and *−13,907*G*. However, the small amount of published data on Iberian populations shows 95% of association between the European *−13,910*T* allele and digester status in Portuguese (Coelho et al., 2005) and 96% in Chileans (Morales et al., 2011).

1.3 Study Population: The Agricultural Communities of Chile

> *The inhabitants of the cottage were freeholders, which is not very common in Chile; they support themselves on the produce of a garden and little field, but are very poor. (…) The Country becomes more and more barren; the valleys have so little water that there is scarcely any irrigation; of course the intermediate country is quite useless and will not even support goats. In the Spring after the winter rains there is a rapid growth of thin pasture and cattle are then brought down from the Cordilleras to graze.*
>
> Charles Darwin – The Voyage of the Beagle (1835).
> Chapter XVI: Northern Chile and Peru.

1.3.1 Overview

The semi–arid region in the North Central part of Chile, locally known as *'Norte Chico'*, is inhabited by agro–pastoralist groups distinguished by a communitarian system of land ownership and management of large extensions of common land. Although there are some of these communities in the neighbouring regions, most of them are scattered unevenly in various rural villages in the Region of Coquimbo. These 180 communities sum up to around 30,000 people owning collectively approximately 1,000,000 hectares with scarce sources of irrigation, little arable land, and small carrying capacity (Instituto Nacional De Investigaciones Agropecuarias, 2005). The region where these 'Agricultural Communities'(as they are legally known) are settled is a transitional zone between the Atacama Desert, one of the driest places on Earth (Clarke, 2006), and the Chilean central valleys. The zone has an average annual rainfall of 100–200 mm (Bahamondes, 2003) and constant threat of droughts. These groups had developed a set of practices that have been described by social scientists as a notable adaptation to their ecological conditions (Alexander, 2008; Avendaño & Gallardo, 1986; Gallardo, 2002). Among these, their system of land management and territorial organisation, their reliance in multiple sources of income from temporary wage labour migration, and transhumant pastoralism are particularly important to understand the lifestyle of these communities.

1.3.2 Geographic setting

The Region of Coquimbo is a narrow area between the Pacific Ocean and the Andes Mountain Range, covering around 40,000 km^2 extending from 29°S to 32°S around meridian 71°W. Terrain is defined by a pronounced slope of incremental altitude towards the Andes, which reaches between

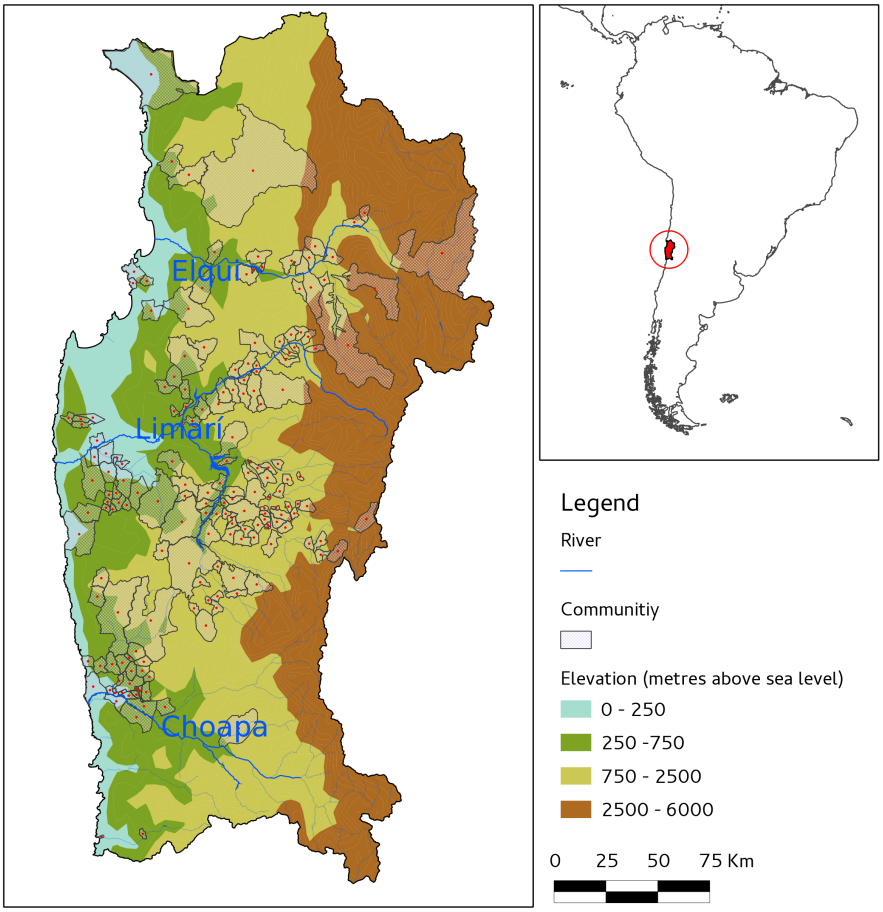

Figure 1.4. *Left:* Map of the Coquimbo Region showing the location of the Agricultural Communities, elevation, and the three main rivers with their associated valleys. *Right:* Location of the Coquimbo Region in South America

3,500 and 6,000 metres in the region (Novoa & López, 2001). Next to the coast, a lower mountain range impedes the flow of winds and humidity from the west and captures most of the moisture, resulting in the dry weather of the area between both mountain chains. This area has scattered west-east oriented mountain ridges below 3,000 metres, and is cut by the three main east to west rivers, named, from north to south: Elqui, Limarí and Choapa. These rivers are associated with three surrounding valleys with the same names (see Figure 1.4).

Most of the population of the region live in settled towns at the coastline. The valleys around the main rivers are occupied by large estates of plantations and agroindustrial companies. The Agricultural Communities are mostly located in the marginal and low–productive areas in the middle and high mountains. These mountains are covered by scarce vegetation consisting of cactus, low acacias, and dryland shrubs, with few sources of water other than the scant rain.

Most of the crop production and vegetation on grazing fields depends on rainfall, which is distributed erratically due to the effect of the El Niño Southern Oscillation (ENSO) in the region (Fiebig-Wittmaack et al., 2011; Young et al., 2009). The distribution of rainfall is usually described as

an anomalous pattern of long periods of drought followed by short periods of excessive rainfall causing floods and land erosion. Further grazing and logging contribute to erosion and it has been estimated that 85% of the regional surface shows signs of erosion, and 18.3% of the regional surface shows no signs of vegetation (Instituto Nacional De Investigaciones Agropecuarias, 2005).

In summary, these communities are settled in areas where available subsistence strategies are very limited by the scarcity of resources: few sources of irrigation and little arable land make large–scale agriculture impossible, and thus agriculture is limited to domestic consumption and irregular dryland cropping, while the paucity and quality of pastures, the irregular terrain, and the low carrying capacity of the grazing fields make goats better suited for the area than other domestic species, despite their significant contribution to land degradation.

1.3.3 History

The peopling of the Americas has been one of the most controversial topics in archaeology, with no consensus about the dates of arrival of the first settlers, number of migratory waves entering the continent or their migration routes. However, it is generally agreed that this was the last major continental mass to be populated by humans, and specifically by one or more small groups of Asian settlers who crossed to the new lands through Beringia, and since then remained isolated from the rest of the world until the European colonisation of the Americas[6] (Fix, 2002; Hey, 2005; Wang et al., 2007).

The site of Monte Verde in southern–central Chile has played a central role in the debate of the time of arrival of the first settlers, with evidence of occupation dated around 14,800 calibrated years ago (Dillehay & Collins, 1988). Although this site is located 1,300 km south of the Coquimbo Region (the distance from London to Madrid), there is evidence of late Pleistocene occupation in the Coquimbo region as old as 13,000 calibrated years before present in creeks Quereo and Santa Julia near Los Vilos (Jackson et al., 2007).

Continuing occupation of hunter–gatherers from both Paleoindian and Archaic periods[7] has been documented archaeologically, followed by the appearance of sites with evidence of early stages of agriculture (corn, beans and quinoa), animal rearing (llama) and pottery between years 0-800 CE known as cultural complexes *El Molle* and *Las Ánimas* (Troncoso & Pavlovic, 2013). There is no agreement regarding the origins of these practices, which are archaeologically similar to those ob-

[6] There is evidence of previous contacts with populations from Scandinavia (Ingstad, 1970; Ingstad & Friis, 1969) and Polynesia (Roullier et al., 2013; Storey et al., 2007) proposed for North and South America respectively. However, systematic contacts with effects on gene flow are unlikely.

[7] Terminology used to divide prehistory of the Americas differs to that used in the Old World: Paleoindian refers to the era of hunter-gatherers of extinct fauna before 10,000 BP. This is followed by the Archaic period characterised by hunter–gatherers of modern fauna. The period when domestic species appeared is referred generally as Formative. This schema is attributed to Willey and Phillips (1985, reprinted by Phillips 2001).

served in the Argentinian Northwest, and could have been adopted by hunter–gatherers or product could be product of population replacement (Troncoso & Pavlovic, 2013).

From 900 CE to 1500, the region was inhabited by a group known as Diaguita (Ampuero et al., 1989), which is archaeologically defined by a particular pottery style and burial pattern. This group occupied a continuous area corresponding to today's Northwest Argentina and the Atacama and Coquimbo Regions in Chile. According to archaeological records (Rodríguez et al., 2004) farming, llama and alpaca herding were their main subsistence activities.

From 1470 until the arrival of Europeans in 1537, local native groups in the Coquimbo Region were invaded, and eventually conquered, by the Inca Empire. Archaeological records show an increase in defensive structures (*pukaras*) followed by the adoption of Inca styles in pottery (Ampuero et al., 1989).

Although the native population was likely to be diminished after the Inca invasion, there were still around 30,000 people by the time of the first contacts with Spanish groups. Shortly after Spanish arrival and the beginning of the pandemic spread of diseases among Native Americans, the native population in the area started to decline (Lorandi et al., 1988). According to the Spanish chronicles of the time, there were no more than 3,000 natives by 1544 (Ampuero, 1978).

In those times, mining was the main activity in the southern part of the Spanish Empire. The *'Norte Chico'* was bordered by some of the most important mines to the north in the middle of the desert. The Region of Coquimbo then became the main source of food and fuel to be used by people working in the mines, leading to an intensive use of lands as a source of firewood and for grain and meat production. It is in this period when livestock and cattle were introduced (Gallardo, 2002). These processes led to land deterioration due to intensive farming, overgrazing, and logging (Dubroeucq, 2004; Santander, 1993).

Economic activities were developed in large estates of plantations and animal herding, known as *'Haciendas'* and *'Ranchos'* respectively. The economy of the Spanish colonies in the Americas was based on land and administrative rights granted by the Crown to high–rank soldiers who were in charge of these estates. In most parts of Spanish America, the workforce contained varying proportions of African slaves and a system of indigenous slavery known as *'encomienda'*: the land grantee was responsible of conversion of the natives to the Christian faith, and they were obliged to work in return for the Christianising efforts on the part of the slave–owner. These two forms of workforce were complemented by paid work of low–rank Spanish soldiers, especially in places with reduced native population. However, the specific composition of the workforce in the estates of the Region of Coquimbo is unknown: while historians argue high importance of a Spanish workforce

due to the lack of native population (Castro & Bahamondes, 1984; Mellafe, 1981), a study on protein markers in these populations supports a trihybrid European–Amerindian–African model (Acuña et al., 2000). This research will try to address this problem in Chapter 4.

In most of Latin America, rules of indivisible inheritance through primogeniture (majorat) and payment of dowries in form of land, led to the concentration of property in few enormous private estates, known as *'Latifundios'*. But this was not the case in the Region of Coquimbo where a communal system of collective land ownership emerged. The origins of this land management system are unknown, but there are three main hypotheses. Santander (1993) supports a Pre–Hispanic origin, based on the Andean system of cooperative work known as *Minka* in Quechua, and still practised today among Quechuas and Aymaras in Peru and Bolivia. Another proposal is that it emerged as a new inheritance system of land grants, replacing the principle of primogeniture with a form of non–partible multigeniture, as a way of coping with the poor conditions for agriculture and the scarcity of workforce (Gallardo, 2002). Finally, Castro & Bahamondes (1986) propose this system as a new form of labour organisation among abandoned estate workers, because of the retirement of landlords once the land was useless for profitable large-scale agriculture.

Whatever the case may be, it is generally agreed that land deterioration made the system of large-scale production counterproductive by the end of the 18th century. Since then, agriculture gradually lost its economic importance in favour of livestock rearing; and through replacement and transformation, the colonial system of land tenure, economy and subsistence changed to the organisation distinctive of the Agricultural Communities today.

During the last two centuries the Agricultural Communities have faced challenges such as further land deterioration due to economic booms in both mining and wheat production during the 19th century, struggles to gain legal recognition of their system of land tenancy (achieved in 1968), and multiple changes in their access to land usage due to political reforms, such as the first Agrarian Reform from 1965 to 1970 by a Christian Democrat government, the second Agrarian Reform from 1970 to 1973 lead by a Socialist government, and the Counter-Reform from 1973 to 1989 led by a right–wing dictatorship, which affected the control of peasants over productive land in rural areas (Alexander, 2008; Gallardo, 2002). After the return to democracy in 1990, commoners have been gradually registering their land usage rights (now fully recognised and regulated by a special law) with the Land Registry authority. However, sustained changes in wage labour conditions, regulations on dairy production, and water usage, are the main threats to the survival of the lifestyle of the Agricultural Communities today (Alexander, 2004).

1.3.4 Economy and Subsistence

The Agricultural Communities have been shaped by the multiple ecological constraints and the particular history of the region in which they are settled. These elements have promoted the development of traits that have attracted great interest of researchers from different disciplines: their particular form of land management and the social organisation related to it, the diversity of activities performed as part of a complex network of temporary labour migration, and the development of transhumant pastoralism.

Communal systems of land management are far from unique, and interestingly, are usually associated with animal rearing and pastoralism. Gallardo (2002) classifies the system of these communities as 'semi–communal' due to its mixture of different forms of private rights and temporary grants within the common, and describes it as more related to the medieval English open–field system and the Swiss Alps Commons rather than communitarian systems of land management in the Americas such as indigenous communities in South America, Indian reservations in the United States, or the Mexican *'ejidos'* (Gallardo, 2002).

This system has been certainly encouraged by the lack of suitable arable land. According to previous studies (Alexander, 2008; Castro & Bahamondes, 1986; Instituto Nacional De Investigaciones Agropecuarias, 2005), only 6% of the area covered by a community is irrigated. This land is divided into individual plots, which are used to grow crops for domestic consumption. These plots are known as *'hijuelas'*, from the Spanish word for 'son', *'hijo'*, and in relation to their agnatic–cognatic inheritance. Due to this inheritance system, some people have more than one *hijuela*, while others have none. The ownership of a *hijuela* classes a person as a 'commoner', giving him or her rights to use the common fields to feed his or her livestock, and to vote about grazing quotas and other issues in communal elections where each right (instead of each person) counts as one vote. A schematic drawing of the system of land usage is shown in Figure 1.5.

Hijuelas, and the commoner's rights associated with them, cannot be sold to people outside the community, yet they can be sold to another commoner upon the approval of the rest of the community by means of election. The commoner's relatives are usually allowed to settle and build a house in the community after paying a small fee. They are able to use the common field for livestock rearing by paying a bit more, but they cannot vote in elections for community representatives, grazing quotas, water usage, communitarian infrastructure, etc. Regional emigration rates are very high (Morales & Parada, 2005) and this is probably due to the massive migrations of junior siblings without land rights.

Commoners have rights to ask for temporary grants of land in the common fields, to be used for

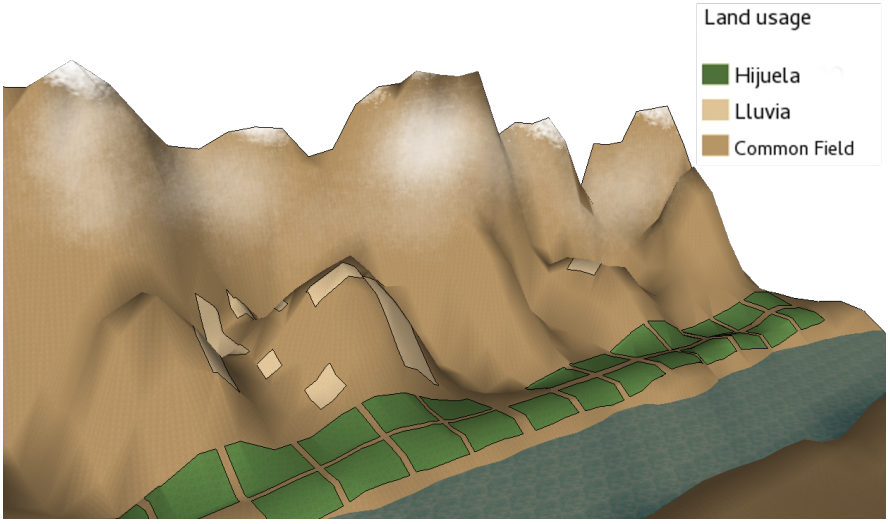

Figure 1.5. Land tenancy and management in the Agricultural Communities. *'Hijuelas'*: individual irrigated plots. *'Lluvias'*: Communal non-irrigated land granted to a commoner in temporary concession for individual use. Common: Grazing fields of collective ownership and usage.

dryland agriculture. The commoner's proposal is voted upon, and if granted, a plot of land in the common field is fenced. These plots are known as *'lluvias'*, the Spanish word for 'rain', because this is the only source of water to feed grains sowed there. *Lluvias* are used mainly to grow barley and wheat, which is used to feed animals and to be sold outside for flour production, which is more profitable than selling cheese. However, dryland agriculture is a risky activity considering that all the investment of a year could be lost if rainfall is insufficient.

In addition to the occasional sales of wheat, families increase their income through seasonal labour in agro-industrial private estates, in mines on the northern part of the country, or through remittances sent by family members living and working permanently outside the community. Though mining is considered by commoners as a very dangerous activity due to the risk of accidents and diseases related with mining (e.g. silicosis, pneumoconiosis, etc.), salaries are extremely high in comparison with the income of selling cheese, and so most young men spent a few years in the northern part of the country working in the mines. Part-time work in nearby industrial plantations or labour migration to other rural areas is also very common, particularly during extensive episodes of drought when the production of milk and crops is drastically diminished in the communities, but its not as severe in irrigated industrial plantations.

However, the most important part of the economy and the main means of subsistence for the communities is livestock rearing. Between 80–90% of the total lands of a community are unenclosed common fields, used collectively for livestock grazing on wild pastures. Goats are by far the commonest and most important animal reared (72%), followed by sheep (15%). Cattle are found only in some of the southernmost communities (7%). In addition, herders usually have a few horses

(5%), and mules and donkeys (1%) for pack and transport (Instituto Nacional de Estadisticas, 2007). The Coquimbo Region concentrates 57% of the total number of goats in the country (404,562 out of 705,527), and local authorities estimate an excess of 263% over the maximum recommended herd size according to the available resources per hectare (Instituto Nacional De Investigaciones Agropecuarias, 2005).

Goat meat, skin, and cheese are destined for the market, mainly to be sold informally at the roadside of the main highways during holidays when high traffic is expected. Cheese selling is an important source of money for the communities, used to buy things that they cannot produce themselves (e.g. clothes, school supplies, cooking oil, matches, artificial rennet, etc.).

Although variable, most herders plan their animals' births once a year, generally around September. Traditionally, during dry years with few pastures, the whole family used to cross all the way through the Andes to the Argentinian side of the mountains with their livestock. The peak of the milking season, from November to March, was usually spent there, when the snow in the mountains has melted and new pastures are available for grazing. During this journey, goats are milked and cheese is made at different stations used as summer dwellings known as *'veranadas'*, to be sold on their way back to the winter settlements (*'invernadas'*). During this trip, meals comprised mainly *'churrasca'* (flat unleavened bread made with flour, butter and water, roasted in the coals of a camp fire), goat's meat, cheese, and milk. In addition to pure fresh milk, milk is also consumed with *'mate'* (a common herbal infusion in South America), and as *'cocho'* (toasted wheat flour mixed with milk and eaten as porridge).

Transhumant pastoralism was the norm in the past, but it is unusual nowadays for many reasons: the selling and production of cheese is becoming more difficult since the authorities in charge of hygiene are starting to forbid it, arguing lack of hygienic standards (Alexander, 2004). Moreover, since school is compulsory, children cannot travel with their parents, preventing them from learning how to produce cheese by the time they inherit the animals and the land. Today, fewer people practice transhumance herding at all, and of those who still practice it, most go less far. This is variable according to the rainfall on a given year, and the whole trip to Argentina is made only in extremely dry years. Nonetheless, and until recently, milk and dairy products were a dominant part of the daily intake of food in these groups for around five months every year. This diet is surprising in the context of the evolution of lactase persistence and its frequencies in Latin America, presented previously in section 1.2, and pose the main research questions of this study, introduced in the following section.

1.4 Research Questions and aims

The main objective of this thesis is to better understand the cultural and genetic adaptation to a milk–rich diet of goat herders from the Agricultural Communities in the Chilean *Norte Chico*. Given the background of their lifestyle and ancestry, their dietary habits led us to pose several questions in relation to their adaptation to milk consumption. The questions, and the sections of the thesis where these are addressed are presented below:

- **How frequent is lactase persistence in these communities and which of the lactase persistence–associated alleles (nucleotide changes reported in the lactase enhancer region), are present?**

  The *LCT* enhancer region, causal of lactase persistence, was sequenced in samples obtained from these communities, in order to determine whether these alleles occur at different frequencies within and between communities. These results are presented in Chapter 3.

- **Are lactase persistence frequencies affected by geography, ancestry, inbreeding, or other confounding causes?**

  Further genetic tests were performed to check that the results were not confounded by geographic clustering, ancestry and relatedness. These and other problems of population structure are explored in Chapter 4.

- **How much milk and milk products are consumed by people in these communities and are any differences in milk consumption associated with lactase persistence? Are there symptoms from milk consumption that are correlated with persistence status?**

  This information was collected through ethnography and questionnaires, to determine whether there is any correlation of behaviour with genotype. General results are presented in Chapter 3, and are analysed in detail in Chapter 5.

- **Are there any differences in height, weight, and Body Mass Index associated with lactase persistence?**

  Height and weight were measured with portable instruments on the field. Associations of these variables with milk consumption and lactase persistence are explored in Chapters 5.

- **Are there any differences between number of children's births and children's deaths associated with the genotype of the parents?**

  Data on reproductive behaviour, fertility, and child mortality is analysed in Chapter 5.

If any or all of these associations can be established, our results might suggest recent natural selection in humans, and improve the understanding about the mechanisms that resulted in the selection for lactase persistence in early milking populations.

Chapter 2

Materials and Methods

This chapter contains a detailed description of the Materials and Methods used in this thesis, and can be consulted as the thesis is read. The Description of Materials starts with a description of the locations included in this study, and features a full page map of the Coquimbo Region and the location of the sites included in this study (Figure 2.1). This is followed the specifications of consumables, equipment and software used in the fieldwork, the laboratory, and in analyses. Description of methods is divided into a section of methods of data collection and methods for analysis of data. The data collection section includes our fieldwork and laboratory protocols as well as description of the genetic markers used in the thesis. The data analysis section describes statistical techniques used in the following chapters.

2.1 Materials

2.1.1 Overview of study sites

The following section introduces the locations where data collection was carried out. Villages were not randomly chosen, but selected according to population size (to allow sampling of individuals as unrelated as possible), population density (to allow cost-effective response rates), and accessibility (due to transport costs). Unfortunately, communities in the north are underrepresented, and those in the high mountains were not included.

Population sizes for each site were obtained from a special bulletin of the Chilean National Institute of Statistics (INE) devoted to the Agricultural Communities of the Coquimbo region (Vergara et al., 2005). These data are based on the Chilean National Census of 2002, which although outdated, is the most recent reliable data source available [1].

Fieldwork was organised in two phases: A pilot study between May and July 2011 and the main fieldwork between February and November 2012. A total of 10 sites were visited, and 449 persons (33.9% men and 66.1% women, all adults) took part in this study. With few exceptions, each participant answered a questionnaire, donated a DNA sample, and had his/her height and weight measured. In addition, a lactose tolerance test was performed by all the participants at one of our study sites. For details about these methods see Section 2.2.

A map showing the location of these sites (Figure 2.1) depicts surrounding urban areas and geographic landmarks as reference [2].

[1] A national census was conducted in 2012, but has been seriously questioned on methodological grounds and is likely to be repeated or corrected (BBC News 2013. `http://www.bbc.co.uk/news/world-latin-america-23611210`.)

[2] Such as the local system of dams and reservoirs started in 1923, and still under development, as a response to the mentioned constant threat of droughts.

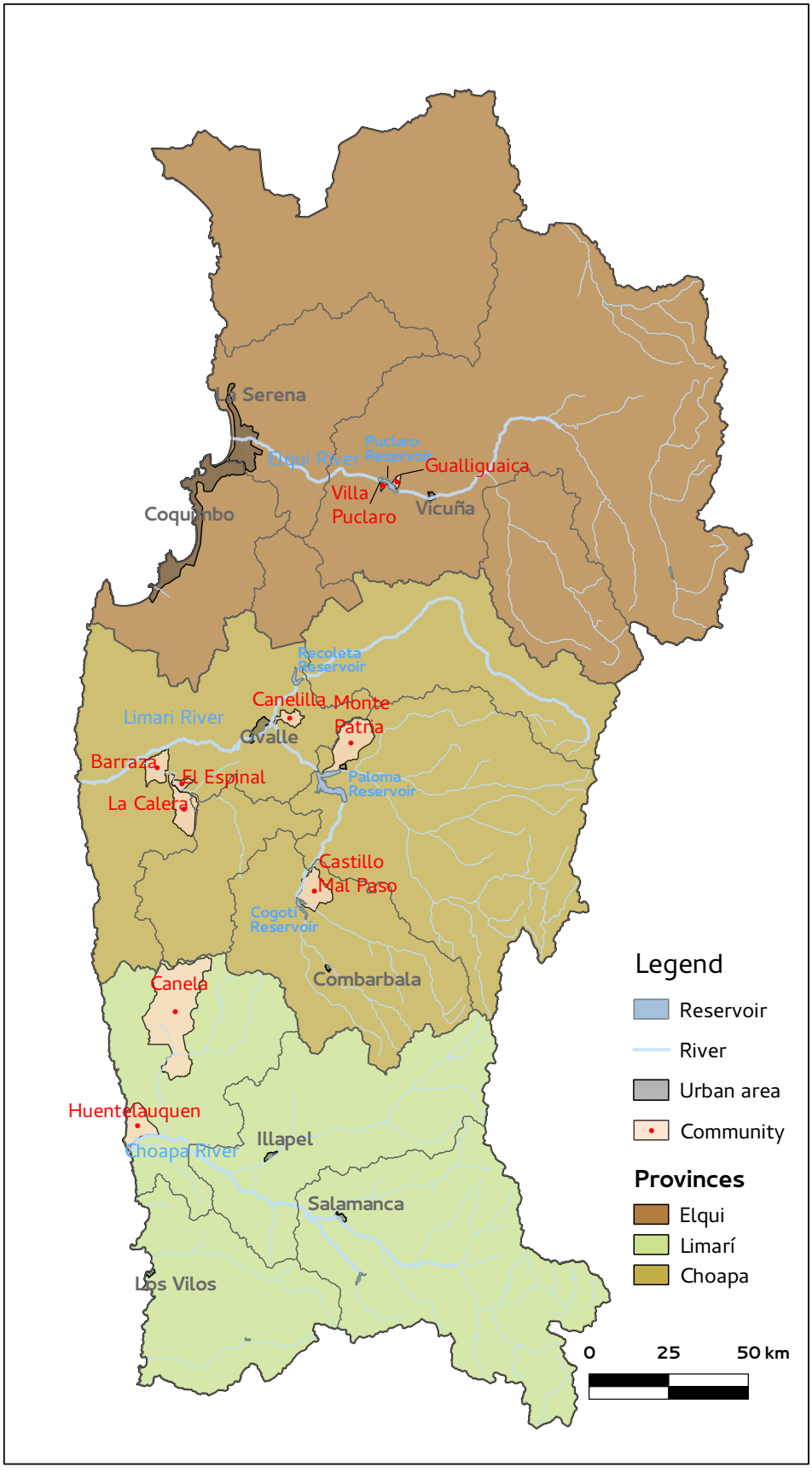

Figure 2.1. Map of hamlets, villages and towns in the Coquimbo Region where data collection was carried out. As reference, other landmarks such as urban areas, main rivers, and reservoirs are included.

2.1.1.1 Gualliguaica and Villa Puclaro

These two villages are located on the northernmost province of the region, the Elqui Valley, between the town of La Serena (pop. 160,000 and the largest urban area in the region) and Vicuña (pop. 12,000), on both sides of the Puclaro Reservoir (see Figure 2.1). Originally the villages of the Agricultural Communities of Gualliguaica and La Polvada were the area flooded by the dam in 1999. Commoners of La Polvada were relocated together with the employees of a cattle-rearing real estate (Punta Azul) to the new Villa Puclaro, and commoners from Gualliguaica were relocated to the other side of the reservoir.

According to Vergara et al. (2005), population sizes of Gualliguaica and Villa Puclaro are 335 and 2,400 inhabitants respectively, but both were of similar size (around 300) at the time of our fieldwork (2011), a difference likely to be attributable to emigration out of Villa Puclaro.

Relocation conditions were very different for the two settlements. People from Gualliguaica managed to keep a village for themselves, and they claim to be happy with their current housing conditions, while Villa Puclaro looks impoverished, and there are constant conflicts between those who were originally from La Polvada and those from Punta Azul. People in Villa Puclaro claim to have worse conditions than what they had before the relocation.

This area was visited as part of our pilot study on June 2011. The Elqui valley, because of its reputation as a tourist attraction and its distance from the lab in Santiago, was the most expensive area for working, in terms of transport and accommodation. It was also the area with the lowest response rates of all the sites included in this study: only 14 people took part, 2 men and 5 women in Gualliguaica, and 3 men and 4 women in Villa Puclaro. Therefore we decided, from the pilot experience, to concentrate our efforts in other parts of the region during the main fieldwork, and to combine these two villages in the analyses[3].

2.1.1.2 Barraza, El Espinal, and La Calera

These three sites are located 30 km west of the town of Ovalle (pop. 100,000), the main urban settlement on the Limarí Province, and are accessible from there through dirt roads. The three places were visited during the main fieldwork in 2012.

In Barraza, settlements were dispersed throughout a big area, and work was divided into three zones: Barraza itself is a village of 350 inhabitants, but surrounded by smaller and more scattered hamlets known as Barraza Alto and Barraza Bajo, of around 150 inhabitants each. A total of 41 people (36 women, 5 men) participated in Barraza, and lactose tolerance tests were performed in

[3] Hereinafter, both villages are grouped under the label *"Puclaro"*.

this site.

La Calera is a large agricultural community distributed into small settlements, the largest of which is a village called Chalinga, were our study was conducted. Out of a population of around 300 people, 49 took part in the study (33 women, 16 men).

Fifteen participants (12 women, 3 men) were recruited in El Espinal, a hamlet of 85 inhabitants.

2.1.1.3 Canelilla and Monte Patria

These two sites are located east of Ovalle, and are easily accessible through paved roads. Canelilla, next to the Recoleta Reservoir, the oldest in the region, has around 300 inhabitants and 23 participants were recruited here (15 women and 8 men) as part of our main fieldwork in 2012.

Monte Patria was visited during the pilot study in 2011. Next to the Paloma Reservoir, the population is concentrated in the town of Monte Patria, a small town of 3,000 people, with access to commercial and administrative services of its own, and with many people employed in work not related to livestock rearing (i.e. services, mining, and agro-industrial work). Despite its size, only 40 participants (24 women and 16 men) were recruited in this town.

2.1.1.4 Castillo Mal Paso

A group of large agricultural communities (Jimenez y Tapia, Manquehua, and Castillo Mal Paso) uses the grazing fields surrounding the Cogotí Reservoir, only 22 km north of the town of Combarbalá (pop. 15,000). Our work was carried out in the village of La Ligua, at the border of the Agricultural Community of Castillo Mal Paso and a population around 600 of whom 63 (43 women and 20 men) were recruited as part of the main fieldwork in 2012.

2.1.1.5 Canela and Huentelauquén

The commune of Canela (pop. 9,000) is a big cluster of 25 Agricultural Communities where most of the domestic economy relies on goat herding. The town of Canela (pop. 3,700 and an Agricultural Community itself) acts as the centre of the area providing access to services and commerce to the surrounding smaller villages and hamlets (Huentelauquén being one of them).

Unlike Monte Patria, most people in Canela work in goat herding, at least seasonally. The town is also well known for its deep involvement in communal activities, local politics, and a large array of community–based organisations. This might explain the high response rates obtained here (in comparison to Monte Patria). As part of the fieldwork in 2012, a total of 164 participants were recruited in both Canela Baja (the main part of the town) and Canela Alta (a small peripheral village), of which 97 were women and 67 men.

The area known as Huentelauquén (pop. 800) is divided in two different settlements by the Choapa River. South of the river is a private real–estate agro–industrial company which rears cattle and produces cheese and avocado. Most of its workers are originally from the Agricultural Community itself, which is located North of the river. Since weather is more benign here, grazing fields are more productive than in the other communities and herders can manage to have some cattle and sheep as well as goats. Huentelauquén was visited on our pilot study in 2011, 42 participants (29 women and 13 men) were recruited in both sides of the river.

2.1.2 Chemicals, equipment and software

2.1.2.1 In-house prepared solutions

- **Generic solutions prepared in lab**
  - **Slagboom buffer:** 10 mM EDTA (pH 8.0), 100 mM NaCl, 10 mM Tris-HCl (pH 8.0), 0.5% (w/v) SDS and 0.2 mg/ml Proteinase K (Sigma [UK], US Biological [Chile]).
  - **TE buffer (1x):** 10 mM Tris-HCl (pH 8.0), 1 mM EDTA (pH 8.0).
  - **STE (10x):** 100 mM Tris-HCl (pH 8.0), 1 M NaCl, 10 mM EDTA, pH 8.0.
  - **Salt solution (for extraction):** NaCl 5 M.
  - **Agarose gel loading buffer:** 15% (w/v) Ficoll (PM400) in H_2O, traces (about 0.25% w/v) of Xylene cyanol FF (Sigma) and Bromphenol blue (Bio-Rad).
  - **Stretch marker:** Agarose gel loading buffer containing about 50 ng 200, 400 and 800 bp PCR products.
  - **2/3 *'Microclean'* (HM-MC):** 26.7% (v/v) Polyethylene glycol (PEG 8000), 0.7 M NaCl, 1.3 mM Tris-HCL (pH 7.5), 0.13 mM EDTA (pH 8.0), 2.3 mM $MgCl_2$.
- **Commercial solutions**
  - **Lactose:** Lactose monohydrate powder. Merck Millipore KGaA.
  - **AutoGen Yeast Reagent 3:** Deproteinization reagent. AutoGen Bioclear Ltd.
  - **Isopropanol:** Analytical reagent grade. Fisher Scientific Ltd.
  - **Ethanol:** Ethanol Ethyl alcohol 99.86% (v/v) min. Hayman Ltd.
  - **Agarose:** Analytic grade. Fisher Scientific Ltd.
  - **Ethidium Bromide:** 10 mg/mL in H_2O. Sigma-Aldrich Co. LLC.

- **Reaction Buffer IV (10x):** 750 mM Tris-HCL (pH 8.8), 200 mM $(NH_4)_2SO_4$, 0.1% (v/v) Tween 20. Advanced Biotechnologies Ltd.
- **dNTP mix (1x):** (0.2 mM each of dATP, dCTP, dGTP, and dTTP). Thermo Scientific.
- **Taq DNA polymerase in 1x buffer:** 100 mM KCl, 20 mM Tris-HCl, pH 8.0, 0.1 mM EDTA, 1 mM DTT, 0.5% (v/v) Tween 20, 0.5% (v/v) Nonidet P40, 50% (v/v) glycerol. Thermo Scientific.
- **PowerPlex 16HS:** STR multiplex System. Promega Corp.
- **PowerPlex 4-dye Matrix Standards:** Matrix Standards for Promega Powerplex 16HS. Promega Corp.

2.1.2.2 Equipment

- **Field Equipment**
  - **Dry cotton swabs:** 75 mm long Plastic Stick, 10 in pouch, cotton bud. WWR International Ltd.
  - **H_2 Cylinder Gas:** QuinGas-1, 100 ppm H_2. QuinTron Instrument Company, Inc.
  - **Stadiometer:** Portable Height Measure SECA 213. Seca GmbH & Co. KG.
  - **Breath hydrogen monitor:** MicroH_2, Micromedical Ltd.
- **Laboratory Equimpment**
  - **Centrifuges:**
    * Biofuge Pico, Heraeus.
    * Hettich Mikro 120 microcentrifuge. Hettich Zentrifugen. Hettich Holding GmbH & Co. oHG.
    * ALC PK-120. ALC Centrifuges.
    * Eppendorf 5430. Eppendorf AG.
  - **Water baths:**
    * Grant OLS200. Grant Instruments (Cambridge) Ltd.
    * Digital Waterbath. Daihan Labtech Co. Ltd.
  - **Incubators:**
    * Kelvitron T. Heraeus.

    * Incublock. Denville Scientific Inc.

  - **Spectrophotometers:**

    * Nanodrop 8000. Thermo Scientific.

    * Optizen Popbio. Laxco, Inc.

  - **Electrophoresis:**

    * Electro-Fast Stretch electrophoresis ELE-872-010U. Applied Biosystems Corp.

    * Microcomputer electrophoresis power supply E443. Consort bvba.

    * Benchtop UV Transluminator M26 UVP. Ultra-Violet Products Ltd.

    * Video graphic printer UP-895ce. Sony Corp.

  - **Thermocyclers:**

    * ABI GeneAmp PCR System 9700. Applied Biosystems.

    * ABI Veriti 96 well Thermal Cycler. Applied Biosystems.

  - **Others:**

    * ABI 3730xl DNA analyzer. Applied Biosystems.

    * Luckham 4RT Rocking platform.

    * Vortex Genie 2. Scientific Industries Inc.

2.1.2.3 List of suppliers

- Autogen Bioclear UK Ltd: Holly Ditch Farm, Mile Elm, Calne SN11 0PY, UK
- Bio-Rad Laboratories: Symbion Science Park, Fruebjergvej 3, 2100 København ø, DK
- CareFusion Health UK 232 Ltd (Micro Medical Ltd): The Crescent, Jays Close, Basingstoke RG22 4BS, UK
- Consort bvba: Parklaan 36 B-2300, Turnhout, Belgium.
- Daihan Labtech Co. Ltd: Ranjit Nagar, Near Pusa Gate 110008, New Delhi, India.
- Denville Scientific Inc: P.O. Box 4588, Metuchen, NJ 08840-4588, USA
- Fisher Scientific (Fermentas, Lonza Bioscience, Macherey-Nagel): Postboks 60, Industrivej 3, 3550 Slangerup, DK
- Fisher Scientific UK Ltd: Bishop Meadow Road, Loughborough LE11 5RG, UK

- Gastrotec Chile: Av. Vitacura 3568 Of.907, Vitacura, Santiago, Chile.
- Grant Instruments (Cambridge) Ltd: Shepreth, Cambridgeshire SG8 6GB, UK
- Hettich Holding GmbH & Co. oHG. Hettich UK, Unit 200, Metroplex Business Park, Broadway, Salford, Manchester M50 2UE, UK
- Laxco, Inc. 22121 17th AVE SE, Suite 114, Bothell WA98021, USA
- LGC Genomics Ltd (KBioscience): Units 1 & 2, Trident Industrial Estate, Pindar Road, Hoddesdon EN11 0WZ, UK
- Life Technologies Europe BV (Invitrogen): filial Danmark, PO Box 37, 2850 Naerum, DK
- Life Technologies Ltd (Applied Biosystems): 3 Fountain Drive, Inchinnan Business Park, Paisley PA4 9RF, UK
- Millipore (U.K.) Limited: Suite 3 & 5, Building 6, Croxley Green Business Park, Watford WD18 8YH, UK
- Promega Corporation: 2800 Woods Hollow Road, Madison WI 53711, USA
- Promega UK Branch Office: Delta House, Southampton Science Park, Southampton SO16 7NS, UK
- QuinTron Instrument Company, Inc: 2208 South 38th Street, Milwaukee WI 53215, USA
- Seca UK: 40 Barn Street, Birmingham B5 5QB , UK
- Sigma–Aldrich Company Ltd: The Old Brickyard, New Road, Gillingham SP8 4XT, UK
- Thermo Fisher Scientific (ABgene): Abgene House Blenheim Road, Epsom KT19 9AP, UK
- UCL Centre for Comparative Genomics: UCL Research Department of Genetics, Evolution and Environment Darwin Building, Gower Street, London WC1E 6BT, UK
- VWR International Ltd: Hunter Boulevard, Magna Park, Lutterworth LE17 4XN, UK

2.1.2.4 Internet resources

- **1000 Genomes:** The 1000 Genomes Project Consortium.
  `www.1000genomes.org`
- **Ensembl Genome Browser:** The European Bioinformatics Institute – Welcome Trust Sanger Institute.
  `www.ensembl.org`

- **HSPH Food Frequency Questionnaires:** Harvard School of Public Health, Nutrition Department's Download site.
  `regepi.bwh.harvard.edu/health/nutrition.html`
- **OSGeo:** The Open Source Geospatial Foundation.
  `www.osgeo.org`
- **OTCA Database of registered commoners rights:** Oficina Técnica Comunidades Agrícolas. Chilean Ministry of National Assets.
  `www.comunidadesagricolas.cl`
- **Structure Harvester:** Visualising STRUCTURE output and implementing Evanno method. Taylor Lab, Department of Ecology and Evolutionary Biology, University of California, Los Angeles.
  `taylor0.biology.ucla.edu/structureHarvester`
- **UCSC Genome Browser:** Genome Bioinformatics Group, University of California Santa Cruz.
  `genome.ucsc.edu`
- **Universidad de la Frontera Map Library:** Cartography Rulamahue. Universidad de La Frontera.
  `www.rulamahue.cl`
- **Short Tandem Repeat DNA Internet Databases:** National Institute of Standards and Technology. U.S. Department of Commerce.
  `www.cstl.nist.gov/strbase`

2.1.2.5 Software

When applicable, references in brackets are provided for publications associated with each software, or when a specific referencing format is suggested by the package documentation.

- **Analyses and editing of sequences and chromatograms:** The following packages were used for aligning, editing, and comparing sequences of the *LCT* enhancer region.
  - **ChromasPro 1.7:** Sequence editing and analysis. Technelysium Pty Ltd.
    `technelysium.com.au`
  - **Unipro UGENE 1.13:** Free open-source cross-platform bioinformatics software. The UGENE team, Unipro. (Fursov & Novikova, 2008; Okonechnikov et al., 2012).
    `ugene.unipro.ru`
- **Visualisation of genotypes:** SNP Genotypes were visualised using LGC Genomics proprietary package SNPviewer, a software with integrated comparison of samples and controls and some

tools for basic analyses (i.e. HWE, scatter plots, and F_{st}). STR Genotypes were analysed with GeneMapper, a system compatible with the allelic ladders and bins provided by Promega, allowing visualisation of peaks in chromatograms and easy identification of repeat numbers.

  - **SNPviewer 2 3.2.2.16:** Scatter plot visualization of genotyping data. LGC Genomics Ltd. `www.lgcgroup.com/products/genotyping-software/snpviewer`
  - **GeneMapper 4.0:** Fragment analysis. Applied Biosystems. `lifetechnologies.com/uk/en/home/industrial/human-identification/genemapper-id-x-software.html`

- **Haplotype inference and comparison:** Haplotypes were inferred using the Bayesian algorithm implemented in PHASE, which were compared using EHH test in Sweep, as described in section 2.3.3.3.
  - **PHASE 2.1.1:** Software for haplotype reconstruction, and recombination rate estimation from population data. Department of Human Genetics – Department of Statistics, University of Chicago (Crawford et al., 2004; Li & Stephens, 2003; Stephens & Donnelly, 2003; Stephens & Scheet, 2005; Stephens et al., 2001). `stephenslab.uchicago.edu/software.html`
  - **Sweep 1.1:** Large-scale analysis of haplotype structure. The Broad Institute (Sabeti et al., 2002). `www.broadinstitute.org/mpg/sweep`
- **Population genetics:** Admixture and STRUCTURE were used to evaluate ancestry proportions and population structure. CLUMPP and Distruct provide functions to customise graphic parameters of the bar plots produced by STRUCTURE. Genepop was used to calculate genetic distances and HWE for both SNP and STR analyses[4].
  - **Admixture 1.23:** Fast model-based estimation of ancestry in unrelated individuals. Cold Spring Harbor Lab. (Alexander et al., 2009) `www.genetics.ucla.edu/software/admixture`
  - **STRUCTURE 2.3.4:** Software for population genetics inference and clustering. Pritchard Lab, Stanford University (Falush et al., 2003, 2007; Hubisz et al., 2009; Pritchard et al., 2000). `pritchardlab.stanford.edu/structure.html`
  - **CLUMPP 1.1.2:** Cluster matching and permutation program for dealing with label

[4] Other analyses were done using population genetics packages for R, which are described in Appendix B.

switching and multimodality in analysis of population structure. Rosenberg lab, Stanford University. (Jakobsson & Rosenberg, 2007).
web.stanford.edu/group/rosenberglab/clumpp.html

- **Distruct 1.1:** Graphically display results produced by STRUCTURE. Rosenberg lab, Stanford University. (Rosenberg, 2003).
web.stanford.edu/group/rosenberglab/distruct.html

- **Genepop 4.2.2:** Population genetics. Raymond and Rousset, Laboratiore de Genetique et Environment, Université de Montpellier (Raymond & Rousset, 1995; Rousset, 2008).
http://kimura.univ-montp2.fr/~rousset/Genepop.htm

- **Miscellaneous genetic tools:** PLINK was used to flip DNA strand for SNPs when required and to check linkage disequilibrium between SNPs. PGPSpider was used to convert genotype data to specific formats required for each software listed in the Population Genetics section.

  - **PLINK 1.07:** A toolset for whole–genome association and population–based linkage analysis (Purcell et al., 2007).
  pngu.mgh.harvard.edu/~purcell/plink

  - **PGDSpider 2.0.5.1:** An automated data conversion tool for connecting population genetics and genomics programs. Computational and Molecular Population Genetics lab (CMPG), Institute of Ecology and Evolution (IEE), University of Berne (Lischer & Excoffier, 2012).
  cmpg.unibe.ch/software/PGDSpider

- **Geospatial tools:** Maps in Figures 2.1 and 1.4 were made using QGis and GDAL[5].

  - **QGis 2.0:** Quantum GIS Geographic Information System. Quantum GIS Development Team, Open Source Geospatial Foundation Project (Quantum GIS Development Team, 2011).
  qgis.org

  - **GDAL 1.10:** Geospatial Data Abstraction Library. GDAL Development Team. Open Source Geospatial Foundation (GDAL Development Team., 2011).
  http://www.gdal.org

- **Data entry, management and statistics:** A database of questionnaire responses was made using EpiData, a software for data entry that provides integrity checks and encryption. Data

[5] Other analyses were done using cartographic and spatial capabilities provided by R packages, which are described in Appendix B.

management and statistic analyses were done in GNU R. Various analyses in R where performed using its extended capabilities through different packages. A list describing the references to these packages can be found in Appendix B.

- **EpiData 3.1:** A comprehensive tool for validated entry and documentation of data. EpiData Association (Lauritsen & Bruus, 2008).
  `www.epidata.dk`
- **GNU R 3.0.2:** A Language and Environment for Statistical Computing. R Core Development Team (R Core Development Team, 2013).
  `www.r-project.org/`

2.2 Data collection methods

2.2.1 Ethnography

Local people and dignitaries from each community were approached during the first days of fieldwork in each village before any recruiting, in order to introduce ourselves and let them know about our work there. This phase included advertising with posters around the village. Afterwards suitable prospective participants were approached, the details of our project were explained, and information sheets were provided (see D. Beforehand, suitable (i.e. safe, clean, and comfortable) meeting space was arranged in each village to interview those who agreed to participate. Most villages had communal venues and sports clubs, and they were usually suitable and available for our purposes upon request. Throughout this phase informal interviews were conduced to collect and record general information about the area, trends in migration and demography, dietary habits, goat rearing practices, milking, and milk processing.

2.2.2 Sampling technique and ethics

As is normally the case in studies involving human subjects, villages were selected strategically according to sample size, accessibility, and previous estimations of inbreeding (see section 4.1.1); and participants were all volunteers. Therefore, biases from sampling relatives and self–reported lactose intolerant individuals could not be avoided at this stage, but other methods were adopted to account for their effects (see section 2.3.4.2 and Chapter 4).

Participants were informed about the project and their personal data were securely protected at all times, in fulfilment of regulations and ethics of research involving living humans. Full ethical approval was obtained both in Chile (University of Chile, Social Sciences Research Ethics Committee CEDEA, reference number 078/2010) and the UK (UCL Research Ethics Committee, reference number 2967/001). In addition, this project was registered with the UCL Data Protection Officer stating that the research project is compliant with the UK Data Protection Act 1998 (registration number Z6364106/2011/11/02, Section 19), and the relevant Chilean legislation (*Ley 19.628 Protección de Datos de Carácter Personal*). Written informed consent was obtained for each participants who agreed to have his or her data analysed as part of this study (English translations of the information sheet and the informed consent forms are provided in Appendix D).

2.2.3 Collection of biological samples

Samples of buccal cells were collected to perform DNA extractions. Each participant was asked to rub 10 cotton swabs on the inner side of their mouth. Swabs were stored in 15 ml conical tubes

containing 2.5 ml of Salgboom buffer (see section 2.1.2.1 for details). Each tube was marked with a distinctive code for each participant, sealed with parafilm, and stored in an insulated thermal bag to avoid precipitation of SDS out of solution in low temperatures. This method was economically convenient compared with commonly used commercial kits such as Oragene (DNA Genotek, Inc.)

During the first season of fieldwork (2011), the tubes containing the swabs were centrifuged and their liquid contents divided into two 1.5 ml conical tubes in a Laboratory at the Department of Anthropology of University of Chile in Santiago, bringing one set of copies of each sample to the UK to do extractions there, and leaving one set of copies in Chile as a backup. During the second season of fieldwork (2012) extractions were done immediately in Chile, and each sample of purified DNA was divided into two with the same purpose.

2.2.4 DNA extraction

Different extraction methods were used in the laboratories at UCL and University of Chile, consisting of modified versions of methods based on phenol/chloroform (Freeman et al., 2003) and salting out precipitation (Quinque et al., 2006) respectively.

At UCL, 0.6 ml of each sample were transferred to a 1.5 ml conical tube with 0.2 ml of the buffer described previously, and incubated in a water bath at 65 °C for 3 hours. Afterwards 100 µl of a phenol/chloroform based commercial reagent (Yeast Reagent 3, Autogen Bioclear, diluted 1:1 with 100% ethanol, see section 2.1.2.1) were added, and the supernatant transferred after vigorous mixing and 10 minutes of centrifuging at 13,000 rpm. This step was repeated once again, after which the supernatant was transferred to clean labelled tubes with 0.6 ml of 100% isopropanol.

These new tubes were gently mixed by inversion, centrifuged for 10 minutes at 13,000 rpm, cooled down in an ice tray for 2 minutes, and centrifuged again for 5 minutes at 13,000 rpm. After centrifugation, supernatant was discarded and DNA pellet was gently washed with 0.6 ml of 70% (v/v) ethanol. Ethanol was immediately discarded and tubes were left open to air dry for 15 minutes. After drying, DNA was suspended in 150 µl of warm (57 °C) TE buffer (see section 2.1.2.1), and left at room temperature overnight. 75 µl of each sample were stored at −80 °C and the remainder at 4 °C.

At University of Chile, we added 1 ml of sample into a 2 ml conical tube containing 1 ml of lysis buffer , and incubated it at 65 °C for 3 hours. Next, 200 µl of NaCl 5 M were added and samples with salt were cooled down on ice for 10 minutes. Tubes were centrifuged at 13,000 rpm for 10 minutes and supernatant was transferred to a new tube with 800 µl of cold isopropanol (4 °C). After gently mixing by inversion, tubes were centrifuged at 13,000 rpm for 15 minutes, supernatant was

Primer Name	Sequence
MCM6778:	5′ - CCT GTG GGA TAA AAG TAG TGA TTG - 3′
MCM6i13:	5′ - ATT CCA AAG AGT CAG AGT CAG AGG ACT TC - 3′

Table 2.1. PCR primers

Stage	Temperature C°	Time (minutes)	Cycles
Initialisation	95	5	–
Denaturation	95	0.5	
Annealing	56	0.5	$\times$39 *cycles*
Elongation	72	1	
Final Elongation	72	5	–

Table 2.2. PCR Cycling conditions.

discarded and tubes were air–dried for 15 minutes. DNA pellets were washed with 1 ml of 75% (v/v) ethanol and centrifuged at 13,000 rpm for 5 minutes before discarding the ethanol and leaving tubes open to air dry for 15 minutes. After drying, DNA was suspended in 150 µl of warm (57 °C) TE buffer (see section 2.1.2.1), and left at room temperature overnight. Purified DNA was transferred to sealed cryotubes to prevent leaking. For protection, tubes were secured in plastic tube storage boxes with ice packs in a polystyrene box, secured inside a hard cardboard box before shipping to London by courier.

2.2.5 PCR and Sequencing of *LCT* enhancer region

A segment of 706 bp of the *LCT* enhancer region (*MCM6*, intron 13) was amplified by PCR using primers *MCM6i13* and *MCM6778* described by Ingram et al. (2007) (see Table 2.1 and Table 2.2). Before PCR, DNA was diluted in ten parts of distilled water. A 20 µl PCR was performed containing 0.5 µM of both *MCM6i13* and *MCM778* primers, 0.2 mM dNTPs, 2.5mM $MgCl_2$, 0.25 µl Taq DNA polymerase and approximately 25 ng of DNA. Please refer to Section 2.1.2 for details of commercial reagents.

Presence of DNA after PCR was checked by electrophoresis using 1 µl of PCR product mixed with 2 µl of loading buffer on a 2% (w/v) agarose gel with 2 µl of Ethidium Bromide (50 ng/ml).

Before sequencing, PCR products were cleaned up by a polyethylene glycol precipitation method using three times the volume of the PCR product of precipitation solution and centrifuging at 4,000 rpm for 60 minutes. After that, plates were inverted on tissue and centrifuged slowly (20 rpm) for 1 minute to remove supernatant. The DNA pellet was gently washed with 150 µl of 70% (v/v)) ethanol, centrifuged at 4,000 rpm for 10 minutes, and then ethanol eliminated by centrifugation

of inverted plates as described before. After allowing DNA to air–dry for 15 minutes, pellets were suspended in 15 µl of water for 10 minutes at room temperature.

Samples were sequenced in both directions using primers in Table 2.1, in an ABI 3730xl DNA Analyzer (Applied Biosystems) by the UCL Centre for Comparative Genomics, using a modified version of the Sanger Method (Sanger & Nicklen, 1977) and a Dye–Terminator Kit. Sequence editing, alignment and visualisation was done using ChromasPro 1.7 (Technelysium Pty Ltd.) and Unipro UGENE 1.13 (Fursov & Novikova, 2008; Okonechnikov et al., 2012).

2.2.6 SNP Genotyping

rs number (dbsnp)	Location band	Ref/Alt	Ancestral	Minor allele (worldwide)	Continental MAF		
					Africa	America	Europe
rs1544450	1p13.1	G/T	T	T	0.947	0	0.089
rs1834619	2p24.2	G/A	G	A	0	0.975	0.037
rs356652	2q11.2	T/G	T	G	0	0.933	0.067
rs260690	2q12.3	C/A	C	C	0.642	0.963	0.045
rs2176046	2q37.3	G/A	G	A	0.015	0.934	0.06
rs10510511	3p24.3	G/T	G	T	0	0.916	0.023
rs3870336	3p21.31	G/A	G	A	0.089	0.936	0.085
rs10935320	3q23	T/C	T	C	0.154	0.979	0.104
rs11725412	4p14	A/G	A	A	0.21	1	0.06
rs10037656	5p15.32	A/G	A	G	0.337	0.98	0.099
rs4145160	5q33.2	G/A	G	A	0.095	0.912	0.067
rs1559163	5q33.2	A/G	A	G	0	0.853	0.023
rs2042314	5q35.1	C/T	C	T	0.139	0.999	0.146
rs12662498	6p12.1	G/A	G	A	0.012	0.980	0.07
rs17086231	6q25.3	C/T	C	T	0.018	0.943	0.117
rs6464749	7q35	A/G	A	G	0.893	0	0.05
rs7018273	8q21.13	A/G	G	G	0.858	0	0.017
rs12347078	9p24.3	A/C	A	C	0.876	0	0.035
rs734241	10q25.3	G/A	G	A	0.044	0.989	0.065
rs174570	11q12.2	C/T	C	T	0.006	0.997	0.111
rs7134749	12q13.12	T/C	T	C	0.21	0.898	0.027
rs2052386	12q15	G/A	G	A	0.077	0.929	0.095
rs1849384	12q21.31	A/C	C	C	0.973	0	0.082
rs4769128	13q12.11	C/T	C	T	0.154	0.988	0.13
rs1243370	14q11.2	T/C	T	C	0.24	0.919	0.055
rs2719921	15q11.2	G/A	A	A	0.876	0	0.033
rs1197062	17q23.2	T/G	G	G	0.891	0	0.057
rs717225	19q13.2	A/G	G	G	0.885	0	0.008
rs6119879	20q11.21	C/T	C	T	0.686	0	0.84
rs2426552	20q13.2	C/T	T	T	0.834	0	0.008

Table 2.3. Thirty Ancestry Informative Markers (AIM), genotyped for ancestry estimations. Minor allele frequencies (MAF) showed in the last three columns refer to frequencies in African, Native American and European populations, and were provided by the laboratory of Professor Andrés Ruiz Linares at UCL GEE. Other data obtained from Ensembl Genome Browser (Flicek et al., 2014) and 1,000 Genomes Project (The 1000 Genomes Project Consortium, 2012)

Samples were genotyped for two sets of SNPs: A panel of 30 Ancestry Informative Markers (AIM) distributed across the genome developed by the laboratory of Professor Andrés Ruiz Linares at UCL GEE (Table 2.3), and 27 SNPs surrounding *LCT* enhancer region in chromosome 2 for inference of haplotypes, as a subset of a panel developed by Anke Liebert (Table 2.4). Genotyping was

rs number (dbsnp)	Position (build 37)	Ref/Alt	Ancestral	Minor allele (worldwide)	MAF (worldwide)
rs1446525	135637847	G/A	A	G	0.288
rs4954209	135737908	G/T	T	T	0.348
rs2874739	135818907	C/T	T	T	0.347
rs1869829	135877562	A/G	G	T	0.385
rs2305248	135928312	A/G	G	G	0.343
rs1900741	136002500	C/T	T	C	0.431
rs1561277	136092061	C/A	A	C	0.264
rs6709132	136232572	A/G	G	G	0.211
rs3806502	136288273	C/T	T	A	0.327
rs4954265	136324225	A/G	G	G	0.27
rs961360	136393658	A/G	A	C	0.315
rs4954278	136408291	C/T	C	T	0.181
rs2278544	136546110	A/G	A	G	0.492
rs2304370	136561735	G/A	G	T	0.254
rs3754689	136590746	C/T	C	T	0.339
rs182549	136616754	C/T	C	T	0.234
rs309152	136657252	T/C	C	G	0.321
rs309137	136765951	T/C	C	T	0.376
rs2090660	136818719	C/T	C	A	0.269
rs12691874	136880474	G/A	G	A	0.339
rs953387	136907170	A/C	A	T	0.46
rs12465599	137074850	A/G	G	G	0.439
rs6715450	137121731	G/A	A	A	0.346
rs543721	137161557	G/T	G	T	0.411
rs12618749	137205474	C/T	C	T	0.228
rs580879	137314139	C/T	T	A	0.257
rs6711718	137407012	T/C	T	C	0.451

Table 2.4. Twenty-seven SNPs surrounding *LCT* enhancer region on Chromosome 2, genotyped for haplotype inference. Minor allele frequencies and other data obtained from Ensembl Genome Browser (Flicek et al., 2014) and 1,000 Genomes Project (The 1000 Genomes Project Consortium, 2012).

outsourced to LGC Genomics (formerly KBioscience), which uses KASP[6] genotyping assays consisting in two allele–specific primers (one for each SNP allele), each containing a unique unlabelled tail sequence at the 5′ end and one common reverse primer, so only one allele–specific primer matches the target SNP (LGC Genomics, 2013). Both positive and negative controls were sent with the samples to LGC Genomics, and resulting genotypes were checked using the software developed by LGC Genomics (SNPviewer 2 3.2.2.16). Genoyping of these SNPs over genome–wide microarrays was preferred since assays were already developed for the mentioned labs, reducing the costs of this project considerably.

2.2.7 STR Genotyping

A kit for forensic identification (Promega PowerPlex 16 HS) was used to genotype 15 highly variable autosomal STR loci (Table 2.5), using a modified version of the manufacturer's instructions to reduce the used volume of reagents to 50%, as described by Ingram et al. (2009b). After amplification, fragments were detected using an ABI 3730xl DNA Analyzer (Applied Biosystems) by the UCL Centre for Comparative Genomics. Data quality was assured using the allelic ladder, positive, and negative controls provided by Promega. Genotypes were analysed and typed using GeneMapper

[6]Kompetitive Allele Specific PCR

STR Locus	Location band	Total n° of alleles (in Promega Ladder)	Alleles (by number of repeats)
TPOX	2p25.3	8	6-13
D3S1358	3p21.31	9	12-20
FGA	4q28	28	16-18, 18.2, 19, 19.2, 20, 20.2, 21, 21.2, 22, 22.2, 23, 23.2, 24, 24.2, 25, 25.2, 26-30, 31.2, 43.2, 44.2, 45.2, 46.2
D5S818	5q23.2	10	7-16
CSF1PO	5q33.1	10	6-15
D7S820	7q21.11	9	6-14
D8S1179	8q24.13	12	7-18
TH01	11p15.5	13	4-9, 9.3, 10-11, 13.3
vWA	12p13.31	13	10-22
D13S317	13q31.1	9	7-15
Penta E	15q26.2	20	5-24
D16S539	16q24.1	9	5, 8-15
D18S51	18q21.33	22	8-10, 10.2, 11-13, 13.2, 14-27
D21S11	21q21.1	25	24, 24.2, 25, 25.2, 26-28, 28.2, 29, 29.2, 30, 30.2, 31, 31.2, 32, 32.2, 33, 33.2, 34, 34.2, 35, 35.2, 36-38
Penta D	21q22.3	14	2.2, 3.2, 5, 7-17

Table 2.5. Fifteen highly variable autosomal STR, genotyped for estimations of relatedness. Data obtained from Short Tandem Repeat DNA Internet Database (Butler et al., 2014).

4.0 (Applied Biosystems) and files for detection of number of repeats for each allele were obtained from Promega.

2.2.8 Lactose Tolerance Test

Lactose digestion phenotypes were determined by lactose tolerance testing using the method described by Peuhkuri et al. (1998) on 40 subjects from Barraza. Volunteers were recruited beforehand and agreed to meet on a specified date and time after fasting for 12 hours.

Breath hydrogen levels were measured using a MicroH_2 meter (see section 2.1.2.2). After measuring baseline levels, subjects had a load of 50 g of lactose in 250 ml of water (see section 2.1.2.1). Afterwards, measures of breath hydrogen were recorded at intervals of 30 minutes. A given test was stopped after two sustained increments of 20 ppm above baseline, and the subject being tested was classified as a lactose non–digester. Individuals showing no substantial rise in breath hydrogen after 3 hours were classified as lactose digesters. The phenotypes of subjects showing fluctuating levels of breath hydrogen were classified as indeterminate.

2.2.9 Anthropometry

Height and weight of participants were measured using a portable scale and stadiometer (see section 2.1.2.2). Instruments were checked between fieldwork trips and replaced when necessary. Participants where asked to remove their shoes, and their head was positioned with the Frankfort Plane parallel to the floor[7]. Sometimes, participants refused to remove their shoes, hats, hair style, heavy jackets, belts, etc. In those cases measurements were considered unreliable and not recorded.

2.2.10 Questionnaires

2.2.10.1 Field procedure

Questionnaires where answered under various different conditions according to the participants' preferences and needs, including their own houses, communal venues, and outside in the field. The necessity of this flexibility to allow higher response rates became clear early in the pilot study. Interviews were always conducted in safe and comfortable places, though complete privacy cannot be achieved in some cases, particularly when the interviews where done in the house of the participant and other family members where present.

Depending on the number of children of the participant, the survey took between 20 minutes to more than one hour. Although there were no open-ended questions, many participants opted to elaborate on some points (particularly when talking about deceased children) and notes were taken about these conversations.

2.2.10.2 Questionnaire contents

Questionnaires were divided into two sub-questionnaires, one corresponding to data about the reproductive history and dietary behaviour of the participant, and a section for each one of his/her children's early dietary habits. A slightly modified version of these two questionnaires was employed for the main fieldwork in 2012, after identifying some questions that were difficult to answer during the pilot study. However the main contents were the same in both versions, and can be summarised in the following list of topics.

- **Residency:** Length of residency at the village and residency at birth.
- **Recent ancestry:** Place of origin of parents and grandparents.
- **Wealth:** Ownership of assets, livestock, communal rights, and access to services.

[7] An imaginary plane crossing the skull above the upper margin of the external auditory meatus and below the inferior margin of the orbit.

- **Milk:** Milk consumption measured in several ways, and symptoms related to it. Participants preferred to refer to their milk consumption in cups, and therefore were used as measuring units. A cup is approximately 250 ml.
- **Family and reproductive history:** This is the main part of the questionnaires, and includes the questions of the sub-questionnaire for each child: Birth place, sex, birth order, weaning, milk drinking behaviour, and date of death and number of grandchildren when applicable. It also includes questions on number of reproductive partners and their places of origin.

An English translation of the questionnaire can be found in Appendix C.

2.2.10.3 Data entry

The paper–based survey was entered into a database using EpiData Entry (Lauritsen & Bruus, 2008). This software provides special facilities designed for entry of data from paper questionnaires, such as a simple language to write tailored consistency checks, automatic backups, random double–entry verification, and encryption.

2.2.11 Assistants

The fieldwork was conducted with the assistance of final–year students of anthropology from the University of Chile. This included collection of samples, survey, anthropometry and lactose tolerance tests. Sixteen assistants were trained in these techniques as part of their degrees, but we also organised training seasons in Santiago prior to each field trip, on which each of us acted as researcher and participant in simulated interviews including measuring, testing, and collecting samples from ourselves. This allowed each assistant to go through all the procedures from both the researcher's and the participants' perspectives.

2.2.12 Published datasets

External databases from various sources were used for some of the analyses in this thesis. General demographic data at the community level used in section 2.1.1 were obtained from a special bulletin of the Chilean National Institute of Statistics (INE) devoted to the Agricultural Communities of the Coquimbo region based on the information of the Chilean National Census of 2002 (Vergara et al., 2005).

Data used for the surnames analysis in section 4.1.1 were obtained from the *Oficina Técnica de Comunidades Agrícolas* OTCA (Chilean Ministry of National Assets. Retrieved: 5 November 2010., 2010) which collates the data about inscription of commoners rights into the Chilean Land Registry authority.

A database of allele frequencies in other populations (Somali and Jaali) for the same STR markers used in this thesis was used to compare estimations of inbreeding obtained from from our data (section 4.2.1). This database was available through the tabulated dataset of the STR analyses by Ingram et al. (2009b).

Geographic information files of the Agricultural Communities were provided by the Chilean National Institute of Agricultural Research (INIA), and databases of cartographic Shape Files for other geographic landmarks were obtained from the Map Library of the University of La Frontera (Albers, 2014). Cartography and GIS used in this thesis was carried out using QGis 2.0 (Quantum GIS Development Team, 2011) and GDAL 1.10 (GDAL Development Team., 2011).

2.3 Data analysis methods [8]

2.3.1 Genotype – Phenotype association

Genotypes of *−13,910 C/T* obtained from sequencing were grouped in two categories, predicted digester and predicted non–digester, according to the *−13,910*T* allele dominance model (Enattah et al., 2002; Sahi, 1974; Sahi & Launiala, 1977), while phenotypes were determined using breath hydrogen–based lactose tolerance test (see section 2.2.8) and recorded as digester, non-digester and indeterminate. Excluding indeterminate phenotypes, these data were analysed as a 2*x*2 contingency table using Fisher's exact test under the null hypothesis of no association between genotype and phenotype, to evaluate the alternative hypothesis of *−13,910*T* as responsible for lactase persistence in this population (results in section 3.1). This allowed us to use the presence of *−13,910*T* as surrogate of lactase persistence for further analyses.

2.3.2 General Analyses

2.3.2.1 Homogeneity of allelic and genotypic frequencies

The hypothesis of homogeneity is a null hypothesis of no differences in the allelic or genotypic frequencies among different groups, which is used in this thesis to identify grouping variables that are important to take into account in further analyses. Sex, geographic origin, birth cohort, and number ancestors born locally in two generations were evaluated as relevant grouping criteria and treated as nominal variables for these analyses. The hypotheses of homogeneity were tested from contingency tables using Pearson's χ^2 test. Originally this was done to use the grouping factor as an additional confounding variable in further analyses, however the null hypothesis was not rejected for any of them (see section 3.2.2 for details).

2.3.2.2 Hardy–Weinberg Equilibrium

The idea of a tendency to the eventual extinction of recessive traits and fixation of dominant traits was a widespread misconception in the early 20th century, addressed by Hardy (1908) and Weinberg (1909) who showed mathematically that, in absence of a specific set of conditions, allelic and genotypic frequencies in a population remain constant regardless any dominance model in phenotypic expression.

This set of conditions was originally mentioned by Hardy (1908) as complete random mating (i.e. panmictic and nearly infinite population size), traits and sexes evenly distributed (i.e. no stratification), and alleles equally fertile (i.e. no selection). The list of conditions has been expanded to

[8] Unless otherwise stated, all statistical procedures were performed in GNU R (R Core Development Team, 2013).

include other evolutionary factors, such as migration, drift and mutation, and the mathematical proof presented by Hardy and Weinberg is extensively used as starting point of different analyses.

One of the uses of Hardy–Weinberg Equilibrium (HWE) is to compare the proportion of genotypes observed in a given population with the expected values they should have if the population is under HWE, usually employing a null hypothesis of homogeneity as described above (section 2.3.2.1). In diploid organisms, the HWE expected frequencies of each possible genotype for a genetic variant of n alleles are given from the multinomial expansion of the square of the n-terms polynomial, i.e. $(p_1 + \cdots + p_n)^2$. Therefore, for a biallelic variant where allelic frequencies are designed p and q, the expected genotypic frequencies are given by the expansion of the binomial square,

$$1 = (p + q)^2 = p^2 + 2pq + q^2$$

The ideal assumptions of HWE of infinite population size and no evolutionary factors acting over the population are impossible to find in real populations. However, the genotypic frequencies of most variants in most populations are in agreement with the expected values under HWE. This is because genotypic frequencies would behave as in HWE under partial fulfilment of this assumptions: Population size does not need to be infinite but large enough for mating to be regarded as random (Hardy, 1908), and other evolutionary forces do not need to be non-existent, but small enough to be regarded as absent. The corollary of this observation is that departures from HWE are rare and are treated as indicative of errors in sampling or genotyping, or as extremely strong effects of population structure, nonrandom mating, selection, migration or drift, thus violating HWE assumptions at a detectable magnitude. Additionally, genotypic frequencies are restored to HWE values in one generation if HWE assumptions are met, and thus HWE can detect effects of evolutionary factors only if they are acting over the current generation (Jobling et al., 2013).

Genotypes of SNPs used for haplotype inference, estimation of ancestry proportions, and STRs used for estimation of relatedness were all tested for deviations of HWE to rule out technical issues as is routinely used. Genotypes of *−13,910 C>T* were also tested according to the different grouping factors mentioned in section 2.3.2.1, to detect whether biological disruptions of HWE assumptions occurred at specific clusters, as has been done using birth cohorts in the study of the resistance to kuru by Mead et al. (2008, 2009).

Departures from HWE were tested using Genepop 4.2.2 (Raymond & Rousset, 1995; Rousset, 2008) which uses different approaches for loci with few variants and for loci with several variants as a way to deal with intensive computational calculations. An approach based on Fisher's exact test (termed 'complete enumeration') is used for the biallelic SNPs, and estimations of p-values from

the Markov chain algorithm are used for the multiallelic STRs (Guo & Thompson, 1992).

2.3.3 Methods for assessment of population structure

2.3.3.1 Ancestry Informative Markers

A panel of 30 SNPs, developed by the laboratory of Professor Andrés Ruiz Linares at UCL GEE, was used to estimate ancestry proportions in Latin American admixed populations, assuming a tri−hybrid admixture model of African, European and Native American parents. The development of this panel was based on the selection of unlinked markers with pronounced differences in frequencies among the parent populations.

Ancestry proportions were calculated with Admixture 1.23 (Alexander et al., 2009), which uses a maximum likelihood method adapted from the algorithm used in STRUCTURE (Pritchard et al., 2000). A dataset of the genotypes of these 30 markers in 876 individuals from parental populations (169 Africans, 299 Europeans and 408 Native Americans) was kindly provided by Andrés Ruiz Linares group. These individuals of known ancestry were used as parental reference by Admixture and all their alleles are considered to come from the same population.

The estimated ancestry proportion $q_k^{(i)}$, is the proportion of alleles in individual i inherited from an ancestor from population k. The estimation procedure considers independently a population of origin for each allele, so the population of origin of the allele a in individual i at locus l is $z_l^{(i,a)}$. Therefore, individual i has a proportion of ancestry from population k equals to the probability of each allele to be drawn independently from k (Falush et al., 2003), or formally,

$$\Pr[z_l^{(i,a)} = k] = q_k^{(i)}$$

All the individuals with less than 80% of genotyping success were excluded from this analysis. Results, expressed as a proportion of ancestry for each parental population (i.e. African, European and Native American), were used to control for ancestry as a confounding variable in relevant regression models (see section 2.3.4).

2.3.3.2 Highly variable autosomal STR identification markers

The inclusion of closely related individuals in our sample presents the problem of potential distortion introduced in all associations. The small population size of these villages (see section 2.1.1) led us to expect high degrees of inbreeding and relatedness. This assumption was also suggested by close family connections observed ethnographically.

To address this problem, a forensic identification kit (Promega PowerPlex 16 HS) was used to gen-

otype 15 microsatellite markers. Using STRUCTURE 2.3.4 (Falush et al., 2003, 2007; Hubisz et al., 2009; Pritchard et al., 2000), *i* individuals were assigned with different levels of membership into *k* different clusters, generating a matrix *Q*, of *i* rows and *k* columns. The value of *k* was determined using the method developed by Evanno et al. (2005).

Several studies (Cardoso et al., 2012; Yu et al., 2006; Zhao et al., 2007) have pointed out that this *Q* matrix cannot capture deep levels of cryptic relatedness and have proposed the use of different relatedness matrices as interacting random effects in regression models. Following Zhao et al. (2007) and Cardoso et al. (2012) we employed the STR genotypes to build a matrix of pairwise proportion of shared alleles (PSA) (Chakraborty & Jin, 1993), and mixed models were formulated using this PSA matrix as random effects in conjunction with the *Q* matrix generated by STRUCTURE as fixed effects.

Additionally, estimations of inbreeding were computed from the genotyped STR markers. Inbreeding was evaluated using the inbreeding coefficient F_{is}, a measure based in the comparison of means of expected and observed values of heterozygosity, in this case for all 15 loci, with a paired t-test after testing for homogeneity of variances using Barlett test[9]. F_{is} were estimated using the GNU R package "`adegenet`" (Jombart, 2008; Jombart & Ahmed, 2011), which employs the same principle originally formulated by Wright (1922). An excess of homozygosity across several loci in a single individual can be the result of a shared ancestry of their parents, F_{is} is the likelihood of being homozygous due to relatedness of the parents: if p_i is the frequency of allele *i*, the probability of an individual of being homozygous at one locus in random mating is,

$$\Pr(\textit{homozygote}) = \sum_i p_i^2$$

Then, the probability in presence of inbreeding is,

$$\Pr(\textit{homozygote}) = F_{is} + (1 - F_{is}) \sum_i p_i^2$$

The estimation of inbreeding is obtained by drawing a random sample from the log-likelihood function of homozygosity summed across all 15 loci (Jombart, 2008; Wright, 1922). A value F_j, defined as the chances of individual *j* of being homozygous through inheritance from a common ancestor of their parents, was computed for each individual to estimate inbreeding (Curik, 2003).

[9] Several alternatives to the methods used here (comparing simple heterozygosity) have been developed to take advantage of stepwise mutation models using STRs, such as R_{st} and d^2 (Coltman, 1998; Curik, 2003). While popular in studies of inbreeding depression in animal breeds, *"simple heterozygosity outperforms them detecting differences over short time periods"*, as pointed out by Neff (2004), making F_{is} more suited for our purposes.

2.3.3.3 Haplotype inference

Gametic phase of 27 SNPs surrounding *LCT* enhancer region was elucidated to test the hypothesis of homogeneous European-like haplotypic backgrounds in segments carrying *−13,910*T* (due to their European origin) and heterogeneous haplotypic backgrounds in segments carrying *−13,910*C* (of both European and Native American origin).

Haplotypes were inferred using PHASE 2.1.1 (Crawford et al., 2004; Li & Stephens, 2003; Stephens & Donnelly, 2003; Stephens & Scheet, 2005; Stephens et al., 2001), which uses data on linked genotypes to estimate haplotypes of each individual taking into account recombination and decay in linkage desequilibrium. This is described by the authors as *"A Bayesian method with approximate 'coalescent with recombination' prior, capturing the fact that each sampled haplotype tends to be similar to another haplotype or to a mosaic of other haplotypes"* (Stephens & Scheet, 2005). PHASE starts assuming unknown haplotypes (Z) to be similar to unambiguous haplotypes (G) or to a mosaic of recombinations of them. It randomly choose an unresolved haplotype from individual i (Z_i) and calculates its probability to be like any of the already known haplotypes from their frequencies estimated so far, then the newly estimated haplotype is moved to the group of known haplotypes and used to estimate the next unresolved haplotype, assuming all the estimations to be correct until going through all the unknown haplotypes. This process is repeated iteratively creating a set of all the estimations in each iteration: During the $(k+1)^{th}$ iteration STRUCTURE samples an haplotype of the individual i, corresponding to that iteration ($Z_i^{(k+1)}$). The haplotype is drawn from a set of all the previously phased genotypes excluding those for the individual i itself, i.e. $P(z_i \mid G, Z^k_{-i})$ (Niu, 2004; Stephens & Scheet, 2005).

PHASE was run five times with 150 iterations using different random seeds on each run to check for consistency between estimations. A dataset of 190 individuals from different Old World populations genotyped for the same variants, kindly provided by Anke Liebert, was included in the PHASE run to compare their haplotypic backgrounds with those estimated for our samples. This was done comparing the extension of the European haplotypes carrying *−13,910*T* with the extension of the *−13,910*T* haplotypes in our sample, using extended haplotype homozygosity (*EHH*) implemented in Sweep 1.1 (Sabeti et al., 2002).

2.3.3.4 Differentiation and distances by groups

Traditionally, F_{st} has been used as a measure of population differentiation between a subpopulation s and the metapopulation t. It was defined for biallelic loci by Wright (1949) as,

$$F_{st} = \frac{Vp}{p(1-p)}$$

Were p is the mean frequency of one of the alleles and Vp is its variance over all subpopulations. It ranges from 0 (absence of differentiation) to 1 (complete differentiation). This measure is used in this thesis fo evaluate differentiation between groups (as defined in section 2.3.2.1), for *−13,910 C/T*. However, there are difficulties to interpret this measure for multiallelic markers, such as the forensic identification STRs.

An alternative measure aims to apply F_{st} to multiallelic loci. Firstly proposed by (Nei, 1973), G_{st} is based on mean heterozygosity calculated for any number of alleles across all subpopulations ($H_t = 1 - \sum_i \bar{p}_i^2$, see section 2.3.3.2). This idea was later adapted by Hedrick (2005) to force the result to fall between 0 and 1 regardless the number of alleles, making it easily comparable with biallelic F_{st}. Hedrick's G'_{st} is defined as,

$$G'_{st} = \frac{G_{st}(1 + H_s)}{(1 - H_s)}$$

It has been demonstrated that G'_{st} fails to identify differentiation when there are various exclusive alleles within subpopulations (Jost, 2008). This is a likely scenario using markers as polymorphic as the STRs employed in this thesis. Jost (2008) proposed an approach that takes into account the proportion of alleles that are common to n subpopulations, and the proportion of alleles in a given subpopulation that only occur on it (Jost, 2008, 2009),

$$D = \left[\frac{(H_t - H_s)}{(1 - H_s)}\right] \cdot \left[\frac{n}{(n - 1)}\right]$$

In this thesis, Jost's D is employed as the preferred method to measure of differentiation between groups when using highly variable STR markers.

Measures of population differentiation are regularly used to evaluate genetic distances between groups through pairwise comparison, and to compare them with the differences of individuals within each group. However, several other methods are available to consider differences across all groups rather than the pair population–subpopulation. The methods proposed by Cavalli-Sforza & Edwards (1967), Nei (1972), Reynolds et al. (1983), and Chakraborty & Jin (1993), are among the most popular. Each of this methods has its own assumptions about the influence of genetic drift and mutation in the change of frequencies, and most are developed under the ideal scenario of no selection, no inbreeding, and no migration between groups. A geometric approximation without biological assumptions uses Euclidean distances,

$$D_{eucl} = \sqrt{\sum_{i=1}^{L}(p_{ji} - p_{ki})^2}$$

Where p_{ji} and p_{ki} are the frequencies of allele *i* in populations *j* and *k* respectively, summed across all loci *L*.

Choosing the right distance measure based on incomplete knowledge of the population history is difficult. These different measures were compared using visualisation techniques based on Principal Components, Neighbour Joining Trees, and Hierarchical clustering to illustrate their differences in the resulting distance matrices in Chapter 4. These analyses were performed using GNU R packages "`gstudio`" (Dyer, 2014), "`ape`" (Paradis et al., 2004), and "`adegenet`" (Jombart, 2008; Jombart & Ahmed, 2011).

2.3.4 Associations between lactase persistence and other traits

2.3.4.1 Estimation of socio-economic status and milk consumption

Some variables to be used in statistical analyses were created as part of data management processes to convert questionnaire responses into formats meaningful for some analyses. Most of them are standard calculations, such as Body Mass Index (BMI) or age at first birth, or common transformations such as converting continuous variables to categorical for some analyses. However variables of socio–economic status and milk consumption are less obvious and require further explanation.

Measurements of socio–economic status are highly dependent on the particularities of a given group. While income or expenditure are straightforward and desirable continuous measures in many cases, these methods cannot account for situations where the assumption of a completely monetary economy is not met, and thus are less useful in context where non–monetary means of subsistence (such as autonomous food production) are prevalent, or where wage labour is sporadic and informal (Falkingham & Namazie, 2002). A common approach for these cases is the use of information about access to certain household goods or conveniences, generally coded as binary variables (presence/absence), to build scores termed *asset indexes* (Booysen et al., 2008)[10].

To construct asset indexes, all variables are processed using multidimensional scaling techniques, to get a factorial weight based on the contribution to the inertia at the first principal axis of each assessed household good. Although most studies use Principal Components Analysis, the categorical nature of the original data dictates Multiple Correspondence Analysis as more appropriate, as has been demonstrated comparing differences in the results according to the selected technique (Asselin & Anh, 2008; Booysen et al., 2008).

In this thesis, 10 binary variables[11] were used to construct a wealth score based on standardised

[10] Terms such as *"score"*, *"index"*, *"composite score"*, or ''*score index"* are used interchangeably in the mentioned literature.
[11] The variables were whether the participant or his/her household have/doesn't have or have access to: tap water,

factorial weights obtained from the contribution of each to the total inertia on the first axis (83.14% of the variance) in a Multiple Correspondence Analysis.

Estimation of levels of consumption of milk and milk products presents similar difficulties regarding particular lifestyles in different societies and difficult techniques for data collection. Methods frequently used according to the literature can be classified in three broad categories: dietary recalls, food frequency questionnaires, and diet journals (Thompson & Subar, 2008).

In recalling methods, the participant is asked to mention what he/she ate in a given period of time, or asked to recall amounts of consumption of particular foods in that period. As long as the period is recent and short, this methods has the advantage of collecting a more accurate picture of actual consumption rather than self–perception. However, it fails to capture seasonality in levels of consumption, and results will be distorted if the person had an exceptional dietary behaviour during the recalled period (Dehghan et al., 2005).

Food frequency questionnaires are based on rating scale responses to questions formulated as how regularly and how much a person consumes a given food, either in general terms or referring to specific long and distant periods of time (e.g. last months, a given year, when unemployed, during drought, etc.). Unlike recall methods, questionnaires are even more heavily influenced by the image a participant wants to project, self–perception of diet, and other psychological factors (Kumanyika, 2003; Marks et al., 2006), but they have the advantage of being able to refer to specific periods of interest and are less influenced by exceptional behaviour at the time of the survey.

Journals are based on periodical records (at least once a day) of every meal a subject has had, indicating quantities in detail (sometimes weighting the food). As long as they are completed for enough time, journals can capture variability due to different situational factors, and account for exceptions while recording actual food consumption. They can be done either by researchers or by the participants themselves, with implications in both invasiveness and accuracy respectively. The obvious drawback is the intensive workload demanded by this method (Kumanyika, 2003).[12]

The questionnaires used during the pilot study included four questions about consumption of fresh milk and milk products; frequencies and quantities of consumption were assessed using both recall and questionnaire frequency methods, by asking for a specific dietary recall of last week's consumption, and general consumption through the year. The question of general dietary behaviour was posed first, asking for quantities and general frequency of milk and dairy products. Then, we asked for a dietary recall for the last seven days, asking how many days a week milk and milk products were consumed. Most participants expressed their concern about the effect of seasonality

electricity, ceiling, floor, water heater, washing machine, fridge, television, computer and vehicle.

[12] Due to both disposition and literacy rates among participants, we could not convince any subject to complete a journal.

in their responses, arguing that milk consumption was higher between September and February, when goats were milking.

The experience during the pilot suggested some confusion in the participants caused by the way these questions were asked, and the results, though there was correlation between employed methods, were suspiciously unexpected. These problems led us to change the way data on milk consumption was collected during the main fieldwork.

One of the problematic aspects of the first set of questions was to ask separately for both frequencies and quantities of milk consumption, which was confusing for the participants and difficult to analyse afterwards. To solve this problem, both questions on frequencies and quantities were fused into a single scale adapted from the Adults Food Frequency Questionnaire developed by the Department of Nutrition at the University of Harvard (Harvard School of Public Health, 2007). Questions regarding seasonality of consumption were added as suggested by participants.

To improve dietary recalls, we added a question focused on quantities of consumption the day before and kept the questions based on last week for comparative purposes. Surprisingly, the results were in agreement with those obtained on the pilot study and no significant differences were found on reported seasonal consumption (see section 3.3.2).

For the pilot phase, estimation of cups/portions a day were calculated multiplying responses on frequency and quantity, and therefore introducing a large amount of error to the final estimation. This was changed during the main fieldwork, and participants chose one option from a scale of different portions/cups of milk/milk products on a specific frequency (i.e. never, two cups a week, one cup a day, etc.), allowing us to estimate number of cups/portion a day directly from their response.

2.3.4.2 Statistical models

Several models were used to find associations between variables related with lactase persistence and variables related with selective advantages, tested using analysis of variance (ANOVA), analysis of covariance (ANCOVA), linear regression, and generalised linear models.

In general, they were built to evaluate the effect of inferred lactase persistence status on height, weight, fertility or child mortality, as response variables. Other variables, such as sex, age, age at first birth, birth cohort, socio-economic status, ancestry, inbreeding, and matrices of population structure were included in models as confounding variables. A summary of the general scheme of these models is shown in Table 2.6.

Even if this scheme is used for most of the tested models, other arrangements are explored (see

Input/explanatory	Output/response	Confounding variable/covariant/fixed or mixed effects
Lactase persistence status	Height	Sex
	Weight	Age
	BMI	Age at first birth
	Total No. of children	Socio-economic status
	Surviving children	Ancestry
	Child mortality	Inbreeding
		Matrices PSA/STRUCTURE (see 2.3.3.2)

Table 2.6. Usage of variables according to the general scheme of most models tested in this thesis.

Chapter 5) using the same statistical methods.

Chapter 3

General Results

A descriptive overview of the results of this study is presented in this Chapter, including the first report to our knowledge of the frequencies of lactase persistence in a pastoralist population of the New World, and the study of the association between the European variant of lactase persistence and lactose digestion phenotype in one of these populations. Additionally, the chapter provides an overview of the demographic profile, milk consumption, and height and weight in the sample.

3.1 Association of *−13,910*T* and phenotyped lactase persistence

As previously mentioned, *−13,910*T* is highly associated with and causative of lactase persistence in European and Indian populations (Enattah et al., 2002; Gallego Romero et al., 2012), but not in some other populations, where other genetic variants have been found (Enattah et al., 2008; Gallego Romero et al., 2012; Imtiaz et al., 2007; Ingram et al., 2007; Jones et al., 2013; Ranciaro et al., 2014; Tishkoff et al., 2007). Because of the historical background of these Coquimbo populations, introduction of the *−13,910*T* through Spanish ancestry is likely, but introduction of variants originating from other continents is less likely but also possible.

To confirm this, and to test for the presence of any previously reported or *de novo* enhancer mutations, a total of 41 volunteers were phenotyped by lactose tolerance test (see Section 2.2.8). Of these, lactose digestion status could not be determined for 3 participants due to fluctuating rise in breath hydrogen but without a substantial rise after 3 hours[1]. In one participant, starting breath hydrogen was above 20 ppm, suggesting that fasting was not accomplished, and in another participant starting breath hydrogen was 0 ppm and remained so all through the 3 hours of testing, suggesting no production of hydrogen.

LTT phenotype	CC	CT	TT	Total
Non–digesters	17	0	0	**17**
Digesters	1	2	16	**19**
Indeterminate	3	1	0	**4**
H_2 non-producer	1	0	0	**1**
Total	**22**	**3**	**16**	**41**

Table 3.1. Association of lactose tolerance test phenotypes and *LCT −13,910* genotype in 99.64% of the subjects (Fisher's exact test *p-value* = 4.187×10^{-9}). No other lactase persistence variants were found.

Of the remaining 36 acceptable lactose tolerance tests, 19 participants were classified as lactose digesters, and 17 as non–digesters. Complete sequencing of the enhancer region showed that 18 digesters carried *−13,910*T*, while none of the non–digesters did. No other lactase persistence variants were found. Results of this lactose tolerance test were highly associated with predicted status according to *−13,910* genotype (Fisher's exact test *p-value* = 4.187×10^{-9} see table 3.1). The one participant genotyped as *−13,910 C/C* and as a digester according to lactose tolerance test, showed no additional variants in the enhancer sequence[2].

[1] Time consumption of the test was an important limitation in the recruitment of volunteers and the achieved sample size.

[2] Some speculative ideas to explain the results of this C/C digester could be a very slow digestion, insufficient air flow when blowing during the lactose tolerance test, or specific microbiomal conditions.

These results are in agreement with published studies for South America (Bulhões et al., 2007; Morales et al., 2011) and allow us to suppose that the principal allelic variant causative of lactase persistence in this population is *−13,910*T*, and thus to consider carriers of this variant as lactase–persistent and non–carriers likely non–persistent.

3.2 Frequencies of lactase persistence

3.2.1 Allelic and genotypic frequencies *−13,910*T* in all samples

A total of 437 of the 451 samples collected were successfully sequenced, of which 13 correspond to samples collected in the Elqui valley, 223 in the Limari valley and 206 in the Choapa valley. From the examination of chromatograms it was possible to determine absence of unreported polymorphic regions and the presence of the European *−13,910*T* in all three groups as the only lactase–persistence associated variant found.

The frequencies of *−13,910* genotypes of these samples are shown in Table 3.2. From these, predicted frequencies of lactase persistence were calculated.

Valley	CC	CT	TT	Total (n)	*T freq.	Predicted Digest. (%)
Elqui	10	2	1	13	0.15	**23**
Limarí	135	73	15	223	0.23	**39**
Choapa	128	65	8	201	0.2	**36**
Total	**273**	**140**	**24**	**437**	**0.22**	**38**

Table 3.2. Frequencies of *−13,910* genotypes in Norte Chico's Agricultural Communities.

These results are in agreement with HWE expectations, and show sufficient predicted digesters (38%) to make comparisons between persistent and non–persistent groups.

Assuming that no other evolutionary factors are in place, any selection coefficient higher than 0.094 acting over the current generation would have resulted in significant ($p > 0.05$) deviation from HWE given this sample size (n = 437)[3]. If there is any selection acting in this population in the current generation, selection coefficient is not likely to be between 0.09 – 0.19 as estimated by Bersaglieri et al. (2004) for neolithic northern Europe based in current frequencies of LP and the allotted time since animal domestication.

3.2.2 Distributions of frequencies by groups

3.2.2.1 Geography

To examine geographic distribution, genotype and allele frequencies were calculated for the different locations (see section 2.2 for details). and are presented in Table 3.3.

Geographic distribution of these frequencies was evaluated by the method of kernel density estimation (Figure 3.1), as has been done in previous studies to estimate worldwide frequencies of

[3] HWE power to detect selection in the current generation depends on a high selection coefficient (s) and a large sample size (n). In a model of complete dominance, the null hypothesis is rejected only if $s \geq \sqrt{\chi^2/n}$, where χ^2 is the value of the χ^2–test statistic for the desired significance level with 1 degree of freedom. See Appendix E for a proof of this rule.

Location	CC	CT	TT	Total (n)	*T freq.	Predicted Digest. (%)
Puclaro	10	2	1	13	0.15	**23**
Canelilla	16	6	1	23	0.17	**30**
Monte Patria	24	10	5	39	0.26	**38**
Barraza	22	16	3	41	0.27	**46**
Espinal	7	5	2	14	0.32	**50**
Calera	30	14	0	44	0.16	**32**
Mal Paso	36	22	4	62	0.24	**42**
Canela	106	48	6	160	0.19	**34**
Huentelauquén	22	17	2	41	0.26	**46**
Total	**273**	**140**	**24**	**437**	**0.22**	**38**

Table 3.3. Frequencies of *−13,910* genotypes by Village.

lactase persistence variants (Gallego Romero et al., 2012; Ingram et al., 2009a; Itan et al., 2010, 2009). Interpolation was done using the R package "`spatstat`" (Baddeley & Turner, 2005).

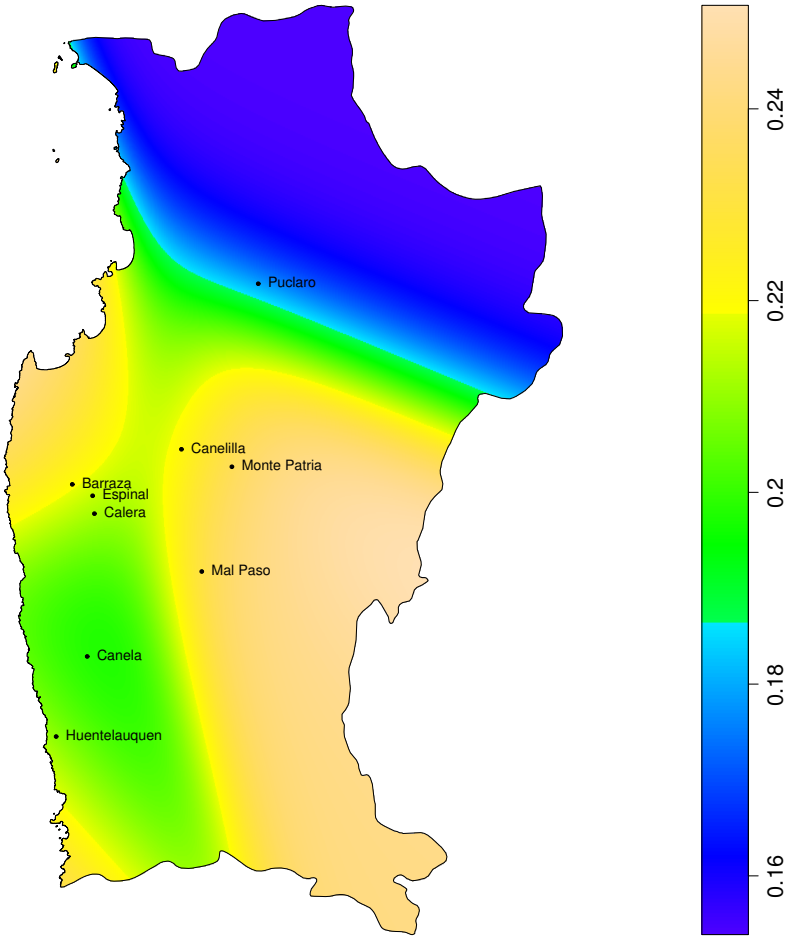

Figure 3.1. Interpolation map showing distribution of *−13,910*T* based on the method of kernel density estimation calculated from the frequencies of this allele in each of the nine villages from which data were collected.

There are higher frequencies of *−13,910*T* towards the southern part of the region, with a second

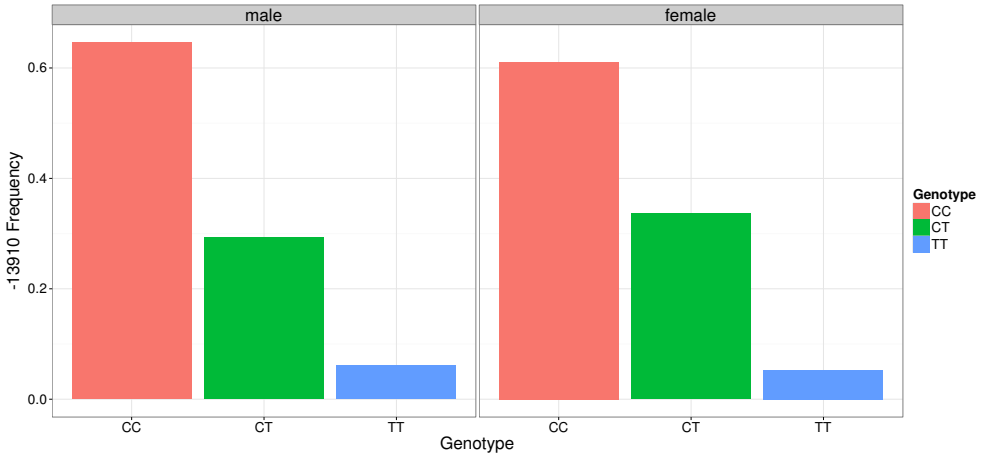

Figure 3.2. Frequencies of *−13,910* genotypes by sex.

cline corresponding to a increment of frequencies with longitude (see Figure 1.4). However, these correlations are not significant (*p-value* = 0.1) tested as a linear multiple regression model combining both latitude and longitude [4].

The frequencies differed from location to location, but did not show differences in HWE. Only Monte Patria shows small deviation (χ^2 = 4.19, *p-value* = 0.04), with fewer heterozygotes than expected. Population differentiation (Fisher's exact test) was not significant for all populations (genotypic *p-value* = 0.4215, allelic *p-value* = 0.366), nor for any pair of populations. Accordingly, F_{st} was small (= 0.0007), and pairwise F_{st} not significant for any pair of populations. No significant correlation was found between pairwise F_{st} and geographic distances between villages (Mantel test *p-value* = 0.942).

3.2.2.2 Sex, age, and birth place of grandparents

Genotypic and allelic frequencies of *−13,910* were compared next between sexes, age groups, and number of grandparents born outside the Coquimbo Region.

The null hypothesis of homogeneity of the groups cannot be rejected for any of these factors. Grouped by sex (Figure 3.2), there is no differentiation (Fisher's method) in either genotypic (*p-value* = 0.734) or allelic (*p-value* = 0.727) frequencies.

Following Mead et al. (2008), comparison of age groups was used to explore historical trends acting as selective pressures in particular generations. No significant differences were found using four age groups (Figure 3.3, genotypic *p-value* = 0.226 , allelic *p-value* = 0.193, F_{st} = 0.003), and all groups are in agreement with HWE expectations.

At a pairwise level, no significant differences were found for any pair of age groups using both

[4] $digest_i = \beta_0 + lat_i\beta_1 + lon_i\beta_2 + \epsilon_i$

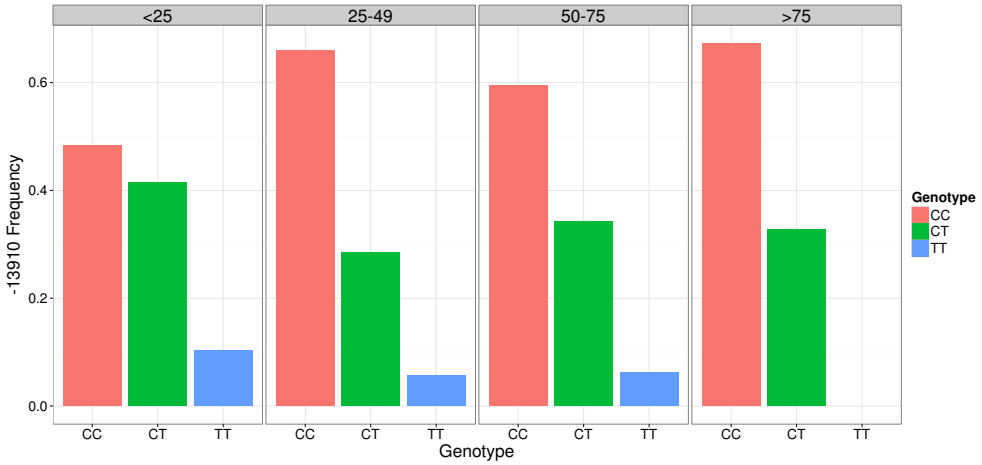

Figure 3.3. Frequencies of $-13{,}910$ genotypes by age group.

Fisher's method and F_{st}, yet differences were slightly higher between the youngest and oldest groups (Pairwise Fisher's genotypic *p-value* = 0.076, Pairwise Fisher's allelic *p-value* = 0.088, Pairwise F_{st} = 0.029). However, this difference is not shown when the analysis is defined for two groups only (born before and after 1960).

Birth places of the grandparents of each participant were recorded on questionnaires, and missing data (20%) were handled using multiple imputation as implemented by the R package "`Amelia`" (Honaker et al., 2011). No significant differences or significant deviations from HWE were found according to the number of grandparents born outside the Coquimbo Region (Figure 3.4, genotypic *p-value* = 0.735, allelic *p-value* = 0.731). As most people in the dataset have no grandparents born outside the region (316 out of 437), differentiation tests were also used to compare this group and all the others together (i.e. all grandparents born in the region versus at least one grandparent born outside), yet no significant differences were found.

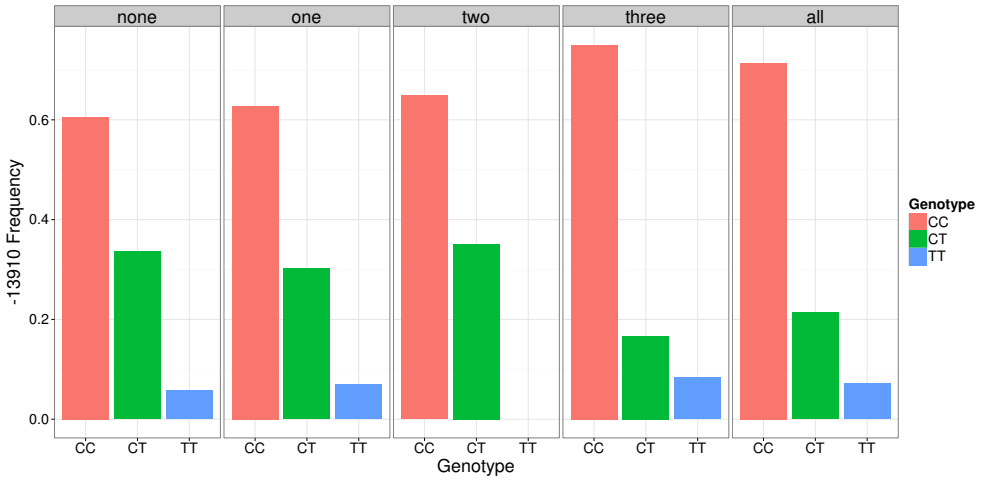

Figure 3.4. Frequencies of $-13{,}910$ genotypes by number of grandparents born outside the Coquimbo Region.

3.3 Summary statistics

3.3.1 Demographic profile

Table 3.4 shows a general socio–demographic profile of the sample. Missing data were excluded, and percentages were rounded to the nearest integer[5]. The table shows a sample composition that is biased towards females and older age groups. This sex bias does not correspond to the sex ratio of 1.05 reported from census data (Vergara et al., 2005), and is probably a consequence of more women staying at home in working hours, when most of the survey took place. The report by Vergara et al. (2005) does not have census data of age structure at community level, but other sources describe a similar composition and high levels of outmigration among young age groups (Instituto Nacional De Investigaciones Agropecuarias, 2005; Morales & Parada, 2005). In this sample, ages range from 18 to 92 years, with an average of 52.44 years (males $\overline{x}$ = 55.19, s = 18.27, females $\overline{x}$ = 51.03, s = 17.72).

Parity (both sexes) has a mode of 2 and a mean of 2.92 children per person. Dead children per person are 0.19 on average (s = 0.52, divided by sex: sons $\overline{x}$ = 0.8, s = 0.38, daughters $\overline{x}$ = 0.12, s = 0.29). Additional statistics related to fertility and child mortality are presented in Table 3.5. The effects of milk consumption and lactase persistence on fertility and child mortality are examined in Chapter 6.

Category	N	%
Sex		
Males	152	34
Females	297	66
Grandparents born out		
None	327	73
1	43	10
2	39	9
3	12	3
All	29	6
Age groups		
18–29	48	11
30–39	70	16
40–49	82	18
50–59	85	19
60–69	65	15
70+	95	21
Parity		
0	66	15
1	51	11
2	111	25
3	80	18
4	53	12
5	30	7
6	23	5
7+	34	8
European Ancestry (proportion)		
0–0.2	3	1
0.2–0.39	110	25
0.4–0.59	247	56
0.6–0.79	77	18
0.8–1	1	0
Access to goods/services (wealth)		
Tap water	413	92
Electricity	436	97
Ceiling	439	98
Floor	429	96
Water heater	283	63
Washing machine	386	86
Fridge	413	92
Television	429	96
Computer	173	39
Vehicle	169	38

Table 3.4. Socio-demographic profile

According to estimations of ancestry proportions on individuals[6], the sample contains a wide range of Native American/European admixture, with marginal African contribution. The proportion of Native American ancestry shows a mean of 0.47 (s = 0.12, min–max = 0–0.8) and is normally distributed (Shapiro–Wilk normality test p = 0.4). Proportion

[5] Therefore, population size can differ for different variables, and percentages may not add up to 100%.
[6] Methods used to estimate ancestry are described in section 2.3.3.1.

			Children ever born			Total living children		
Stats.	Age	At 1^{st} *birth*	Sons	Daughters	Total	Sons	Daughters	Total
$\overline{x}$	52.44	24	1.56	1.36	2.92	1.45	1.28	2.73
s	18	6.8	1.56	1.41	2.4	1.47	1.3	2.21
$min - max$	18–92	11–52	0–10	0–8	0–16	0–9	0–7	0–15

Table 3.5. General statistics: Age at 1^{st} birth, Births, and Child survival

of European ancestry has a mean of 0.48 (s = 0.13, min–max = 0.09–1) and is also normally distributed (Shapiro–Wilk normality test p = 0.25). In contrast, proportion of African ancestry is only 0.05 on average (s = 0.06, min–max = 0–0.29) with a heavily skewed distribution (Shapiro–Wilk normality test $p = 2.2 \times 10^{-16}$)[7].

People in the sample with access to a list of goods and services are listed at the end of Table 3.4. Most people have access to basic services (such as water or electricity), while goods of high value are not very rare (38% have a vehicle). From these data it is clear that the selected goods and services are good estimators for socio-economic status only for the lowest end of the distribution (i.e. It would classify as poor only those without basic services, while most of the other people would be classified as wealthy). These data were used to build a scale (see section 2.3.4.1), which varies from 0 (poorest) to 1 (wealthiest), and has an average of 0.85 in the sample (s = 0.16).

Category	**N**	**%**
Milk Drinkers	372	83
Drank milk yesterday	130	37
Consume other milk products	425	95
Feels unwell with milk	179	40
Daily cups of milk		
0 (never)	86	19
> 2	336	75
Between 2 and 3	23	5
More than 3	5	1
Weekly milk consumption (days)		
0 (no milk last week)	162	38
1	51	12
2	43	10
3	40	9
4	14	3
5	7	2
6	3	1
7	108	25
Weekly dairy consumption (days)		
0 (no dairy last week)	78	18
1	66	15
2	74	17
3	60	14
4	26	6
5	15	3
6	7	2
7	115	26

Table 3.6. Consumption of milk and milk products

3.3.2 Milk consumption

A summary of variables related to milk consumption is presented in Table 3.6[8]. Only 16.9% (76) participants said that they never drink milk. To extract the most information, we asked about consumption in general, consumption the day before the survey, and consumption in the last seven days[9]. According to these results, the three variables are highly associated (log linear test of independence for three–ways contingency table $p = 7.4 \times 10^{-62}$). These data were used to estimate

[7] Interestingly, and as anecdotal test of the method, a Spanish missionary priest living in the communities shows 100% of European ancestry.

[8] As before, missing data were excluded, and percentages were rounded to the nearest integer.

[9] The question of consumption the day before was included only in the questionnaires of 2012 and the reported percentage corresponds to the total of that year.

	Males			Females		
Persistence status	Height (cm)	Weight (kg)	BMI	Height (cm)	Weight (kg)	BMI
All						
$\bar{x}$	167.79	77.3	27.41	154.3	69.22	29.07
s	7.7	11.92	33.46	6.2	13.11	5.11
$min - max$	139.8–195	48.7–113.6	17.84–35.99	130.2–180	37.4–146.4	18.32–50.54
Persistent						
$\bar{x}$	169.67	82.4	28.58	153.98	69.1	29.15
s	7.52	11.92	3.5	6.3	12.76	5.04
$min - max$	145.5–195	54.1–113.6	21.17–35.99	140.2–180	40.7–108.1	18.32–44.64
Non–persistent						
$\bar{x}$	166.86	74.97	26.89	154.37	69.37	29.08
s	7.56	11.06	3.26	5.99	13.29	5.09
$min - max$	139.8–184.6	52.8–99.5	20.19–35.67	130.2–146.4	37.4–146.4	18.58–50.54
Student's t-test						
$t - value$	2.1842	3.713	2.838	−0.534	−0.399	−0.09
$p - value$	**0.03***	**3** $\times$**10**$^{-4}$***	**0.005****	0.594	0.69	0.928

Table 3.7. Summary statistics of height, weight and BMI by sex and lactase persistence status

daily milk consumption (see section 2.3.4.1). Consumption of other milk products is higher, with only 5% (23) participants reporting no consumption, and 56% (198) consuming a milk product the day before the survey.

As expected, participants who reported feeling unwell when they drink milk consume less (ANOVA p = 0.006). But surprisingly, milk consumption is not associated with digestion status predicted by lactose tolerance test (ANOVA p = 0.82), nor with predicted lactase persistence status according to $-13{,}910$ genotype (ANOVA p = 0.94).

Association of milk consumption with variables of interest such as socio-economic status, height and weight, and lactase persistence will be examined in detail in Chapter 5.

3.3.3 Height and weight

Table 3.7 shows summary statistics of height, weight and BMI, divided by sex and by lactase persistence status according to $-13{,}910$ genotype. Average BMI is 28.5 kg/m^2, and while BMI $>$ 25 kg/m^2 is considered overweight in adults, BMI is known to be higher in older age groups (Heiat et al., 2001), and these results are likely to be influenced by the age composition of the sample.

Height, Weight, and BMI were all significantly higher in lactase–persistent individuals, but only in males (see Figure 3.5). Differences in some or all of these variables in relation with lactase persistence have been previously reported in European populations (Corella et al., 2011; Lamri et al., 2013; Smith et al., 2008), but not as specific to males. This could be caused due to other variables playing a role more important than lactase in female stature, such as reproductive schedule or gender bias in nutrition at early ages. These hypothesis will be commented in relation to our results in Chapter 6.

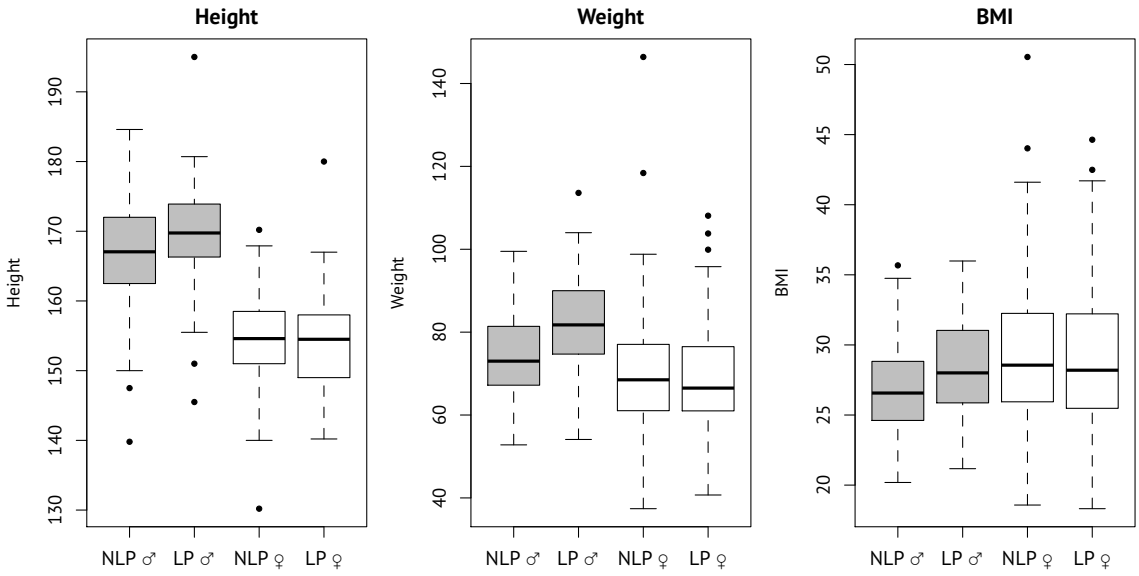

Figure 3.5. Differences in height, weight and BMI between lactase–persistent (LP) and lactase non–persistent (NLP), in males (♂) and females (♀). Significant differences between persistent and non–persistent were found for the three variables only in males.

3.4 Chapter conclusions and discussion

The results described throughout this chapter show the presence of lactase persistence in this population, and sufficient variability in traits of interest to explore possible selective advantages of lactase persistence. Additionally, this is the first report showing genotype frequencies of lactase persistence in a pastoralist group from the New World.

As European gene flow is the most likely origin of the lactase persistence variant found in this population, ancestry is an important confounding variable in associations. A similar north-south cline reported here for $-13{,}910^{*}T$ has been found by Acuña et al. (2000) in estimations of ancestry proportions in the same geographic area, showing higher indigenous ancestry in the North or the region. Similarly, several studies (Malina & Reyes, 2007; Story et al., 1999, 2003) have reported high prevalence of obesity associated with Native American ancestry in the Americas. In this sample, proportion of native ancestry is strongly correlated with both BMI and weight but only in females (Pearson's correlation test females: BMI $p = 8 \times 10^{-4}$, weight $p = 0.003$, males; BMI $p = 0.9$, weight $p = 0.6$). An inclusion of an outgroup to compare this results would be an interesting possibility for future works.

These conclusions highlight the possible impact of the confounding effects of geography and ancestry in further analyses. These and other aspects of population structure are examined in detail in the next chapter.

Chapter 4

Population Structure in the Agricultural Communities

To examine any possible relationship between lactase persistence, milk drinking, and phenotypic outcomes such as height, weight, and fertility, a full consideration should be made of population structure. This chapter concerns relatedness and ancestry, identified as likely sources of stratification in this population. Preliminary, relatedness was explored using surnames as an aid to deciding on a sampling strategy and further genetic analyses. Next, forensic identification markers were used to estimate inbreeding, proportion of shared alleles between individuals, and proportions of allocation to different presumed clusters using STRUCTURE 2.3.4 (Falush et al., 2003, 2007; Hubisz et al., 2009; Pritchard et al., 2000), and Ancestry Informative Markers (AIMs) were used to estimate ancestry proportions using Admixture 1.23 (Alexander et al., 2009) and STRUCTURE. At the end of the chapter, these estimations are examined in association with predicted lactase persistence status to evaluate the importance of the confounding effect of population structure.

4.1 Approaching population structure using surnames as inherited markers

4.1.1 Surname studies: Overview

Planning a sampling strategy before data collection presented the problem of how to deal with population structure. Since most previous studies are sociological, there was little information about genetic structure in the literature (see section 1.3). However, a registry of names of people living in these communities was available from the Chilean Ministry of National Assets (see section 4.1.2), suggesting the idea of examining population structure using surnames.

Different societies have different naming systems defining how personal names are composed. Numerous systems have existed, such as those based on occupation or trades; patrons, partners, or parents given names; place of birth, geographic landmarks, physical attributes, or ethnic and religious affiliations (Jobling, 2001; Longley et al., 2007).

In some European languages (as in English), surnames are inherited traits passed down in a patrilineal fashion from fathers to children. This system of transmission is analogous to the inheritance of DNA, and particularly, to the inheritance of the Y chromosome (King & Jobling, 2009). This fortunate coincidence makes surnames a very good tool to examine population structure, in the absence of other data.

Surnames have been used to study population structure for many years. Reviews (Colantonio et al., 2003; Jobling, 2001) attribute the origin of this field to the work of George Darwin (son of Charles Darwin), who used surnames to estimate frequencies of marriages between first cousins, and to compare the frequency of first–cousin marriages between parents of asylum inmates and the general population (Darwin, [1875] 2009)[1].

Originally, the method known as isonymy, from the Greek words ἴσος: *'same'* or *'equal'*, and νυμία: *'name'*, was developed to be used with data on marriages, counting the frequency of marriages between couples with the same surname. Isonymy was defined in genetic terms and formalised by Crow & Mange (1965), who coined the word 'isonymy' and described methods to calculate F-statistics (Wright, 1932, 1965) and to estimate inbreeding from this kind of data.

Several methods expanded this work to further examine inbreeding, isolation by distance, and migration from data on surnames. These methods included improvements in the estimations of

[1] He found similar frequencies of first–cousin marriages among parents of mental patients and the general population, and hence discarded an association between inbreeding and mental disease. Interestingly, George's parents, Charles and Emma Darwin were themselves first cousins.

F-statistics (Jorde & Morgan, 1987; Yasuda & Furusho, 1971), the use of lists of surnames instead of list of marriages (Weiss, 1980), methods to build similarity and distance matrices from surnames (Lasker & Kaplan, 1985; Relethford, 1988; Weiss, 1980), development of standard parameters to compare inbreeding across populations (Rodríguez-Larralde et al., 1993), and methods to identify geographic barriers to gene flow (Rodriguez Diaz et al., 2010).

These methods draw criticism because of their assumption of monophyletic origin of surnames (reviewed in Rogers 1991), and because high degrees of relatedness (estimated from the Y chromosome) have been found only for rare surnames (Jobling, 2001; King & Jobling, 2009). Regardless of the flaws of the methods based in isonymy, they have been widely used in many countries (see Colantonio et al. 2003 for a review), showing, in some cases, strong associations with estimations based on Y chromosome (Sykes & Irven, 2000).

4.1.2 Methods and Analysis

Analyses of surnames were used in two phases during the course of this thesis. Firstly, we used this approach using data from the 20 most populated villages, to identify the three locations with the highest variability of surnames and highest estimated distances, and thus avoiding samples from closely related individuals as much as possible[2].

During the main fieldwork, for practical reasons such as difficulty of access, proximity to other villages, and interest of the villagers in our research, surname information was not prioritised in deciding locations. Indeed, two locations included in our final sample were not among the 20 most populated villages used for the first analysis. We therefore repeated the analysis, but this time using data of surnames from the villages included in the actual sample. This second analysis is reported in this Chapter, and the data used are presented in Table 4.1. A report with the results of the first analysis, used to decide a sampling strategy, can be found in Appendix F.

The data on surnames were readily available in a convenient form for these purposes as a result of a public policy, in place for the last 20 years: due to the particular system of land tenure of these communities, governments have been encouraging members to register their individual rights as commoners in a dedicated section of the Land Registry. Since these lists are recorded at the community level they are particularly useful for our purposes. However, many commoners have not yet registered their rights and some commoners who have migrated out of the villages might still have their rights registered. Additionally, each list contains only the name of the person who is registering the rights on behalf of the whole household, thus does not include partners, children, or any other people living with them. Therefore, these lists are not accurately representative of

[2] These are the three villages visited during the pilot study: Puclaro, Monte Patria, and Huenteluquen.

the real populations, and the ratio of people registered varies widely between communities and is not proportional to the total population size.

A second source of population data comes from a special report on Agricultural Communities published by INE (Chilean National Institute of Statistics) based on data obtained from the National Census of 2002 (Vergara et al., 2005), described in section 2.2.12. This report contains data on population size at community level, but it is out of date, and the figures may have been affected by patterns of seasonal migration.

Community	Total names	*S*
Puclaro	195	44
Canelilla	128	36
Monte Patria	525	84
Barraza	137	37
Espinal	108	24
Calera	335	64
Mal Paso	236	46
Canela	1319	140
Huentelauquén	391	87
Total	**3275**	**307**

Table 4.1. Number of total names and number of unique surnames (*S*) obtained from the Land Registry records for each community.

The data on the Census available to the public do not include names of individuals, and therefore, our surnames analysis is done based on the lists from the Land Registry authority. Land registry data were used to compare the size of the list of each community with the population size reported in the census, assuming that the latter includes all the people living under a single land registration as well as non-registered commoners. We found a weak correlation between registered commoners in the Land Registry list and the number of houses in each village (r = 0.5299), and between registered commoners and overall population size (r = 0.4937).

These communities use the Spanish system of surname inheritance; people have two surnames, a first surname from the father, and a second surname from the mother, and normally women do not change their surnames upon marriage. Since a person only has two surnames, only the paternal surnames (one from each parent) are passed down. Therefore maternal surnames are lost from one generation to the next, so that people do not share any surnames with their grandmothers. While surnames are usually described as analogous to the Y chromosome, this system would be better described as male-inherited mitochondrial DNA, since a woman does have a surname (inherited from her father instead of her husband), but cannot pass it down to her grandchildren. This feature allows an extended method of isonymy using four surnames instead of two, resulting in more precise estimations of inbreeding (Lasker & Kaplan, 1985; Pinto-Cisternas et al., 1985; Shaw, 1960), and has been exploited to study surnames in several Spanish speaking populations (Bronberg et al., 2009; Dipierri et al., 2005; Pettener et al., 1998; Rodriguez Diaz et al., 2010), including Chile (Barrai et al., 2012). However, we could not take advantage of this trait since the maternal surnames were not always recorded in the data available.

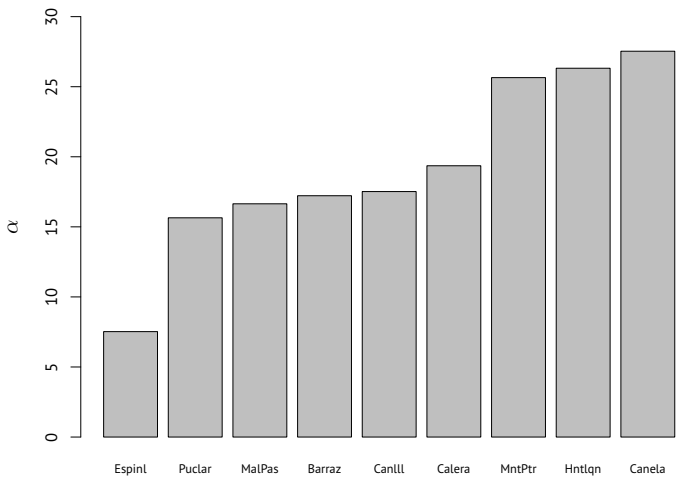

Figure 4.1. Fisher's α for each Village based on frequencies of surnames. Values of α are equivalent to the effective number of surnames in each population.

These lists of surnames have some interesting traits: Firstly, many communities are named after family names (e.g. Barraza, Castillo); and many names that are rare in other parts of Chile (and associated with rich families) are very common here (e.g. Tagle, Cox, Cotapo, Ossandon). In addition, some surnames are clear mutations of others (e.g. Velasco and Velazco, Rojas and Roja, Cox and Coz, Sapiaín and Zapiaín). However, these features have been ignored in the analysis since we do not have the means to test them empirically.

The extent of inbreeding was estimated form the diversity of surnames, following a method conceived to study species abundance (Fisher et al., 1943) and adapted to the study of inbreeding using surnames from data on marriage (Rodríguez-Larralde et al., 1993) or lists (Barrai et al., 1996). The procedure starts by defining Unbiased Isonymy (I) as $I = \sum_i (p_i)^2 - 1/N$, where p_i is the relative frequency of the i^{th} surname and N is the total number of people in the list[3]. I is normally converted into Fisher's alpha ($\alpha = 1/I$), a value homologous to the effective number of alleles, as it results in a number of surnames which, all at the same frequency, would have the observed I (Rodriguez-Larralde et al., 2011).

Values of α for each village are presented in Figure 4.1. These values are very low compared to other studies (Bronberg et al., 2009), and are in agreement with previous findings showing the Coquimbo (and particularly the Limari Valley) as the region with the smallest α in the whole country[4] (Barrai et al., 2012). The results of α as an estimate of inbreeding will be compared with the estimations of inbreeding using STR markers in section 4.2.1.

To examine isolation by distance, the surname lists were use to compute a similarity matrix (Fig-

[3] This is homologous to F_{is} as shown in section 2.3.3.2

[4] The results by Barrai et al. (2012) at the province level are: Choapa's α = 113.3, Limari's α = 76.9, Elqui's α = 140.2, Coquimbo Region α = 124.

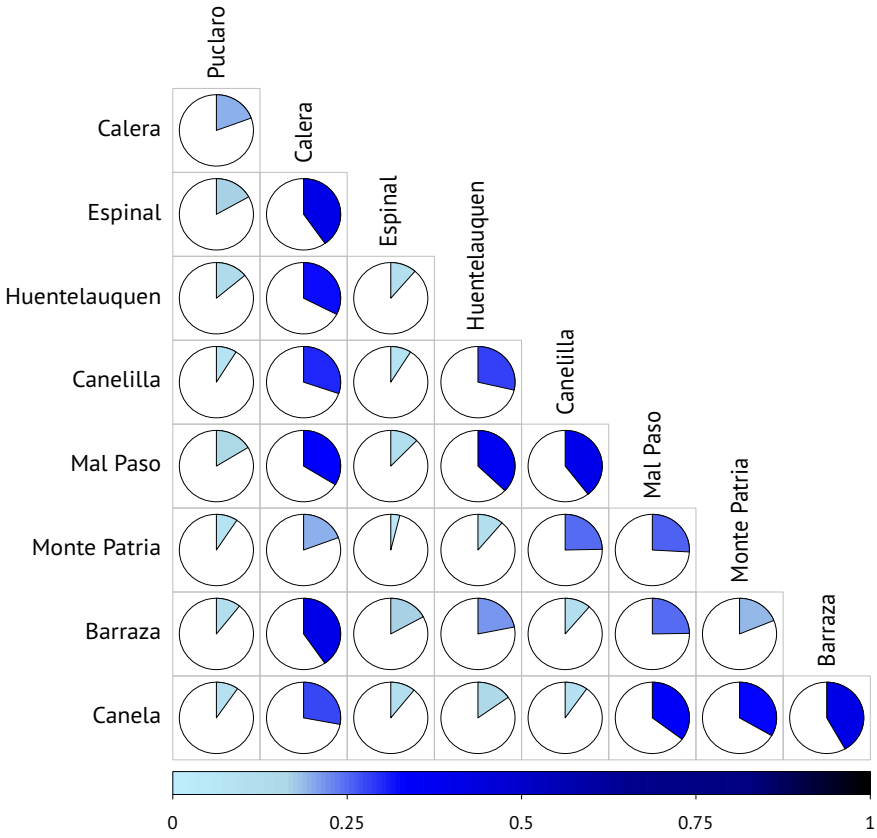

Figure 4.2. Similarity matrix of surnames frequencies (Hedrick's method). The grid shows a pie chart at each pair of villages, representing the similarity in the distribution of surnames between two populations calculated using the method described by Hedrick (1971) and adapted for surnames by Weiss (1980). Pairwise similarity represented by pie chart can range from 1 (same surnames at the same frequencies in both villages) to 0 (no surnames in common between both villages). Colours are used to highlight the same differences shown in the pie charts according to the colour key below the grid.

ure 4.2) based on the method described by Hedrick (1971), adapted for surnames by Weiss (1980), using the R package "`Biodem`" (Boattini & Calboli, 2012).

This matrix shows low similarity between Puclaro (the northernmost village) and the other villages, and high similarity between the contiguous villages of Calera, Espinal and Barraza. Furthermore, a correspondence analysis of surnames (see Figure 4.3) divides communities geographically (from south to north) on the first dimension (17.64% of variance).

Correspondence analysis shows weak geographic structure. The first axis captures Puclaro and Huentelauquen at the northern and southern extremes, as expected, but Mal Paso and Canela should be closer to Huentelauquen, and Canelilla and Monte Patria closer to the cluster of Espinal, Barraza and Calera (see the map in Figure 2.1 to identify these locations). The second axis (16.31% of variance) cannot be interpreted geographically in the same fashion, though it roughly suggests distribution from east to west, with the exceptions of Canela, which should be in the middle, and

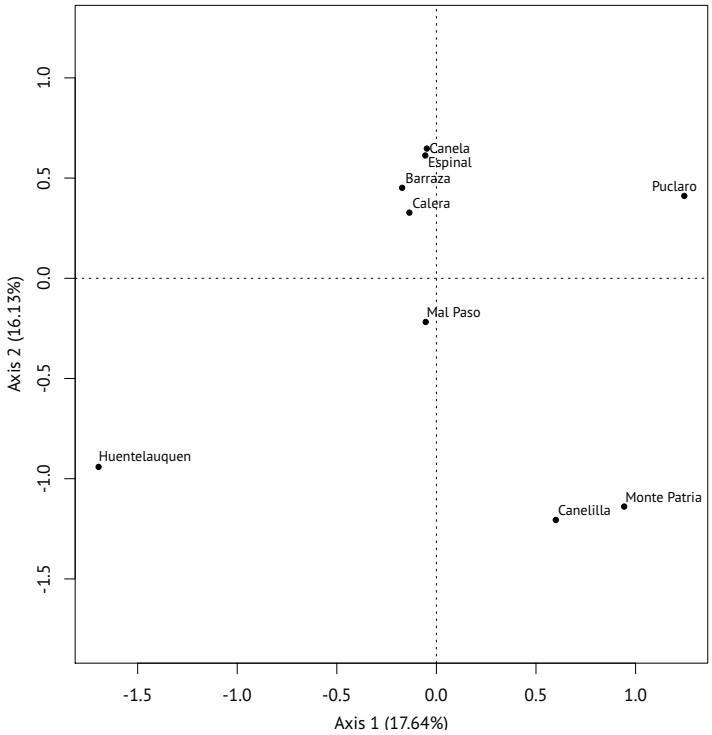

Figure 4.3. Correspondence analysis of surnames frequencies order villages from South to North at the first axis (17.64% of variance). The second axis (16.31% of variance) cannot be interpreted geographically in the same fashion and does not show geographic structure.

Huentelauquen, which is the westernmost location and should be at the top.

To measure this trend, pairwise kinship coefficients were calculated using Lasker's method (Lasker, 1977), and transformed into a distance matrix according to $L_{dist} = -\log_e(2R)$, where R is Lasker's kinship coefficient (Rodriguez-Larralde et al., 1998). A Mantel test was used to assess correlation between this matrix and a matrix of geographic distances, confirming a small but significant correlation (r = 0.4, $p-value$ = 0.03, based on 100,000 replicates).

To conclude, the analysis of surnames suggests some levels of inbreeding and small, but detectable, effect of isolation by distance. A comparison between geographic and surname distances can be visualised using neighbour–joined trees, as is shown in Figure 4.4. These suggestions will be tested using highly variable STR markers in the next section.

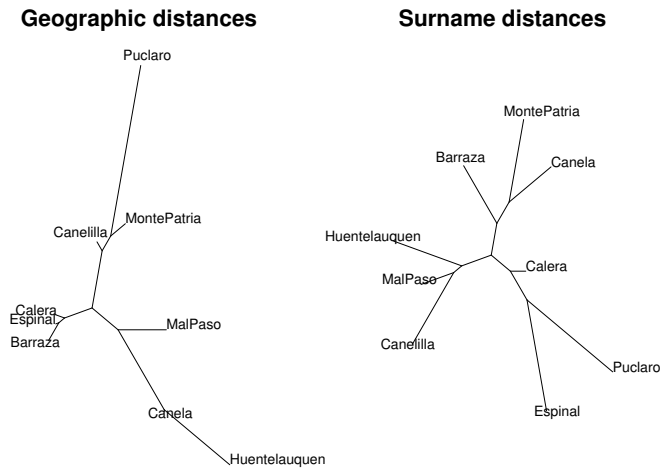

Figure 4.4. Comparison of neighbour-joined trees of geographic and surnames distances: Unrooted neighbour-joined trees of Geographic distances (left) and surnames distances estimated from Hedrick's matrix (Matrix in Figure 4.2).

4.2 Relatedness

4.2.1 Distribution of forensic STR identification markers at individual level

The use of STR markers for paternity testing, forensic identification, and in this case for estimation of inbreeding and relatedness is based in its high variability, an effect of their high mutation rate. Due to this variability, a single locus usually would have multiple alleles, and therefore, sharing several alleles with an unrelated individual is not likely. Based on the frequencies of the alleles in the population, STR markers allow estimation of probabilities of homozygosity due to identity by descent. However, the high mutation rate came with the disadvantage of the increased chance of a new mutation that would lead to mistakenly consider two related individuals as unrelated. Weighing advantages and disadvantages STRs are a cost-effective way to estimate relatedness and inbreeding, and the inclusion of several loci would reduce the effect of hypothetical mutations on the estimations of relatedness.

The 15 forensic identification markers described in Table 2.5 were used to estimate relatedness. Allele frequencies per locus are presented in Table 4.2. All markers showed high variability, comparable to what has been found in other Latin American mixed populations (Bravo et al., 2001; Hill et al., 2013; Rubi-Castellanos et al., 2009a). There were no significant differences between expected and observed heterozygosity over all loci (t–test $p–value$ = 0.8264). Individually, only two loci (*D21S11* and *CSF1PO*) showed lower heterozygosity than expected under Hardy–Weinberg Equilibrium. Considering all loci, the variance in heterozygosity is similar between expected and observed values (Bartlett test of homogeneity of variances K^2 = 0.1742, $p–value$ = 0.6764)

For further analyses, cases with incomplete genotyping were removed, resulting in a final sample

size of 351 individuals.

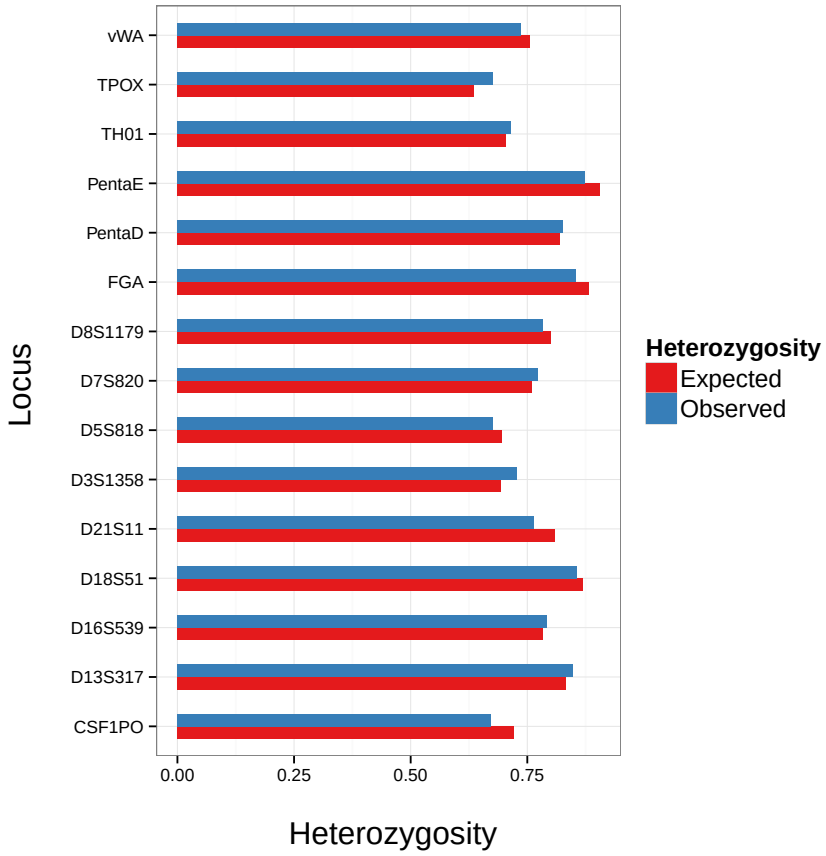

Figure 4.5. Expected vs Observed Heterozygosity of 15 STR markers. Differences are not significant considering all loci together (t–test $p-value$ = 0.8264)

This dataset was used to estimate inbreeding coefficients (F_{is}) from a sample of the likelihood distribution of homozygosity due to a common ancestor of both parents, as described in section 2.3.3.2. An average value of F_{is} ($\overline{F}$) was calculared for each individual based on 1,000 values sampled from the likelihood distribution of homozygosity, $\overline{F}$ ranges from 0.071 to 0.611, with a mean of 0.149 and a standard deviation of 0.07. Figure 4.6 shows a histogram based on the mean estimation of inbreeding per individual with the kernel density curve of the likelihood distribution superimposed. To interpret these values, the same procedures were used with a dataset of 159 unrelated individuals from East Africa, genotyped at the same loci by Ingram et al. 2009b. In that dataset $\overline{F}$ ranges from 0.072 to 0.471, with a mean of 0.151 and a standard deviation of 0.072. Both datasets show similar positively skewed distributions, with a longer tail in the Chilean dataset due to outliers with high F_{is}, but differences in these distributions are not significant (two samples Kolmogorov–Smirnov test $p-value$ = 0.913). These results suggests that inbreeding in our sample is not significantly higher than inbreeding in this sample of African individuals who were reported as unrelated at the grandparental level[5].

[5]Although a more closely related population would have been better suited for this comparison, the Jaali–Somali

Allele	D3S1358	TH01	D21S11	D18S51	Penta E	D5S818	D13S317	D7S820	D16S539	CSF1PO	Penta D	vWA	D8S1179	TPOX	FGA
2.2	–	–	–	–	–	–	–	–	–	–	0.0014	–	–	–	–
5	–	–	–	–	0.0222	–	–	–	–	–	–	–	–	–	–
6	–	0.2681	–	–	0.0014	–	–	–	–	–	–	–	–	–	–
7	–	0.3139	–	–	0.0956	0.0637	–	0.0152	–	0.0028	0.0264	–	–	0.0028	–
8	–	0.0625	–	–	0.0139	0.0042	0.0708	0.0734	0.0056	0.0014	0.0083	–	0.0014	0.5194	–
9	–	0.1222	–	0.0041	0.0055	0.0485	0.1931	0.0623	0.1931	0.0279	0.1681	–	0.0028	0.0776	–
9.3	–	0.2250	–	–	–	–	–	–	–	–	–	–	–	–	–
10	–	0.0083	–	0.0041	0.0637	0.0388	0.0972	0.2687	0.1222	0.2493	0.2639	–	0.0499	0.0235	–
11	–	–	–	0.0097	0.0651	0.4612	0.1875	0.3324	0.2486	0.3134	0.1583	–	0.0720	0.2825	–
12	0.0014	–	–	0.1105	0.1953	0.2756	0.2431	0.2161	0.2917	0.3384	0.1861	–	0.1662	0.0942	–
13	0.0028	–	–	0.1271	0.0817	0.1025	0.1208	0.0291	0.1306	0.0599	0.1431	0.0014	0.3199	–	–
14	0.0801	–	–	0.2099	0.0845	0.0055	0.0806	0.0028	0.0083	0.0056	0.0375	0.0651	0.1814	–	–
15	0.4530	–	–	0.1533	0.1094	–	0.0069	–	–	0.0014	0.0069	0.0693	0.1676	–	–
16	0.2749	–	–	0.0967	0.0457	–	–	–	–	–	–	0.3490	0.0374	–	–
17	0.0856	–	–	0.1367	0.0609	–	–	–	–	–	–	0.2922	0.0014	–	0.0028
18	0.0967	–	–	0.0732	0.0499	–	–	–	–	–	–	0.1510	–	–	0.0042
18.2	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0.0028
19	0.0055	–	–	0.0276	0.0485	–	–	–	–	–	–	0.0693	–	–	0.0900
19.2	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0.0014
20	–	–	–	0.0249	0.0235	–	–	–	–	–	–	0.0028	–	–	0.0623
20.2	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0.0014
21	–	–	–	0.0166	0.0194	–	–	–	–	–	–	–	–	–	0.1316
22	–	–	–	0.0041	0.0069	–	–	–	–	–	–	–	–	–	0.1288
22.2	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0.0166
23	–	–	–	–	0.0055	–	–	–	–	–	–	–	–	–	0.0942
23.2	–	–	0.0014	–	–	–	–	–	–	–	–	–	–	–	0.0042
24	–	–	–	0.0014	0.0014	–	–	–	–	–	–	–	–	–	0.1288
25	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0.1662
26	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0.0983
27	–	–	0.0264	–	–	–	–	–	–	–	–	–	–	–	0.0526
28	–	–	0.0902	–	–	–	–	–	–	–	–	–	–	–	0.0069
29	–	–	0.1748	–	–	–	–	–	–	–	–	–	–	–	0.0069
30	–	–	0.2885	–	–	–	–	–	–	–	–	–	–	–	–
30.2	–	–	0.0139	–	–	–	–	–	–	–	–	–	–	–	–
31	–	–	0.0832	–	–	–	–	–	–	–	–	–	–	–	–
31.2	–	–	0.111	–	–	–	–	–	–	–	–	–	–	–	–
32	–	–	0.0111	–	–	–	–	–	–	–	–	–	–	–	–
32.2	–	–	0.1248	–	–	–	–	–	–	–	–	–	–	–	–
33	–	–	0.0028	–	–	–	–	–	–	–	–	–	–	–	–
33.2	–	–	0.0610	–	–	–	–	–	–	–	–	–	–	–	–
34.2	–	–	0.0097	–	–	–	–	–	–	–	–	–	–	–	–
35	–	–	0.0014	–	–	–	–	–	–	–	–	–	–	–	–
H_{obs}	0.7283	0.7155	0.7640	0.8571	0.8736	0.6770	0.8479	0.7725	0.7915	0.6723	0.8254	0.7360	0.7837	0.6770	0.8539
H_{exp}	0.6936	0.7051	0.8085	0.8686	0.9055	0.6951	0.8330	0.7596	0.7838	0.7203	0.8191	0.7549	0.7999	0.6360	0.8824
HWE	0.894	1	**0.0005****	0.4565	0.8955	0.5995	0.668	0.9995	1	**0.0015***	1	0.964	0.8925	0.944	1

Table 4.2. Allele frequencies of 15 autosomal STRs in the sample. H_{obs}: Observed Heterozygosity, H_{exp}: Expected Heterozygosity, HWE: *p–value* of Hardy-Weinberg Equilibrium from exact test (10,000 permutations). Frequencies were rounded to the fourth decimal place.

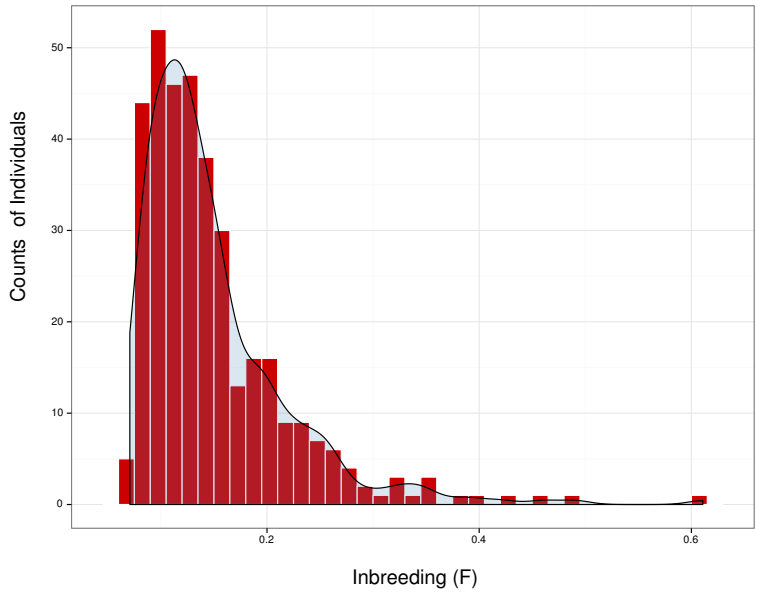

Figure 4.6. Histogram of frequencies of mean estimation of inbreeding by number of individuals: $\overline{x}$ = 0.149, *sd* = 0.07, min–max = 0.071–0.611. Superimposed: Kernel density curve of the likelihood distribution of homozygosity due to a common ancestor of both parents.

In this study we questioned participants about their relatedness to other volunteers. Since not all samples were collected in the same day, we had to ask each day for relatedness with the participants interviewed the day before. As expected, this method was very unreliable in the largest villages, and several inconsistencies in the records were found after the fieldwork. Because of this, collected data on reported relatedness were not used in the analyses on the previous paragraph. Nevertheless, reported relatedness can be affected by unknown kinship relationships such as paternal uncertainty or consanguinity between ancestors more than two generations away. To acknowledge this problem, termed *"cryptic relatedness"* in the literature (Astle & Balding, 2009; Voight & Pritchard, 2005), population structure was explored further using a matrix of proportion of shared alleles (PSA matrix, as described by Chakraborty & Jin, 1993; Zhao et al., 2007 and Cardoso et al., 2012), and a matrix of membership to defined clusters using STRUCTURE (*Q* matrix, as described by Falush et al., 2003, 2007; Hubisz et al., 2009; Pritchard et al., 2000 and Evanno et al., 2005). These methods are described in detail in section 2.3.3.2.

The matrix of pairwise PSA (63,903 pairs) has a mean of 0.33, with an standard deviation of 0.079. The minimum proportion of shared alleles was 0.067 (in 17 pairs of individuals), while 7 pairs of individuals shared the same genotype at the 15 loci (PSA = 1). Figure 4.7 shows that extreme outliers are rare, and this distribution is a signal of the possible usefulness of this matrix as random effects to take into account the confounding effect of relatedness in further regression models.

To build the *Q* matrix, STRUCTURE was configured with a length of 100,000 burn-in periods and

dataset was already genotyped in our lab using the same forensic kit.

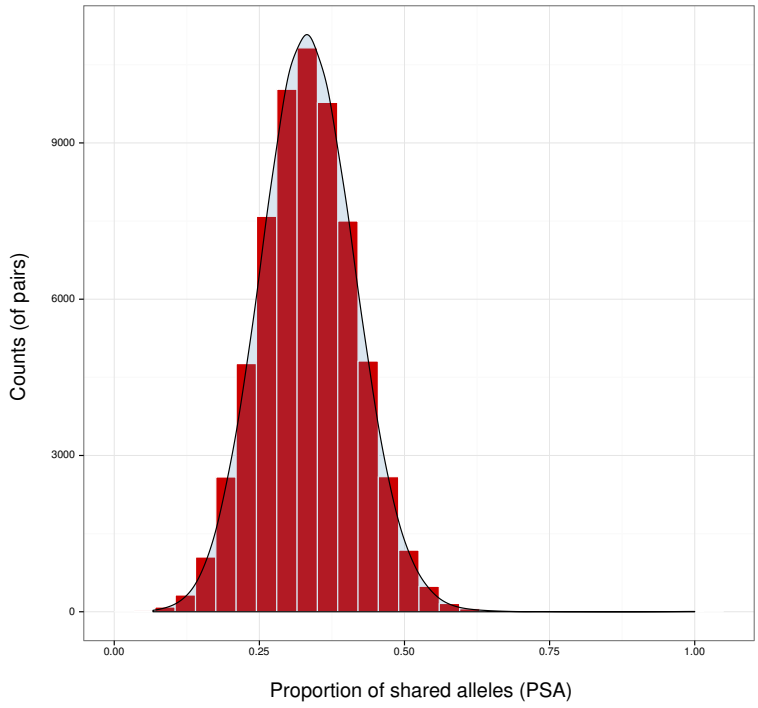

Figure 4.7. Histogram of frequency of proportion of shared alleles (PSA) between all pairs of individuals in the sample. PSA per pair of individuals $\overline{x}$ = 0.33, *sd* = 0.079, min–max = 0.067–1. Superimposed: Kernel density curve of the likelihood distribution of proportion of shared alleles.

100,000 MCMC repeats, using Admixture LOCPRIOR model and different values of α for each population as suggested for detection of weak structure signals in closely related populations (Porras-Hurtado et al., 2013). All other parameters were set to default. These settings were used in 10 runs of STRUCTURE for each value of *k* between 1 and 10 (a total of 100 runs). The output was organised using Structure Harvester 0.6.93 (Earl & VonHoldt, 2011), which employs the method described by Evanno et al., 2005, which resulted in a recommendation of 2 as the optimal value for *k*[6]. Afterwards, all runs of the identified $k = 2$ were processed using CLUMPP 1.1.2 (Jakobsson & Rosenberg, 2007), and a final bar plot (Figure 4.8) was made using Distruct 1.1 (Rosenberg, 2003).

These results can be interpreted as suggesting weak population structure in the distributions of these markers (i.e. variation within villages cannot be accounted for from the allocated proportions of one of the cluster). The optimal value of *k* implies that assuming two clusters account for sufficient variance, which is not significantly increased at higher values of *k* quantified using the values of ΔK (Evanno et al., 2005). All individuals are admixed with respect to these two clusters, but with differences in contribution of each (as will be examined further in the next section). Another interesting alternative hypothesis to test is the correspondence between these two clusters with European and Indigenous American ancestries. This will be explored in section 4.3.

[6]The Evanno method recommends a value of *k* based on ΔK, defined as the rate of change of the log probability of data for consecutive values of *k*, and thus indicating if increasing the value of *k* would result in a significant change in the allocation of individuals. For details refer to Evanno et al. (2005).

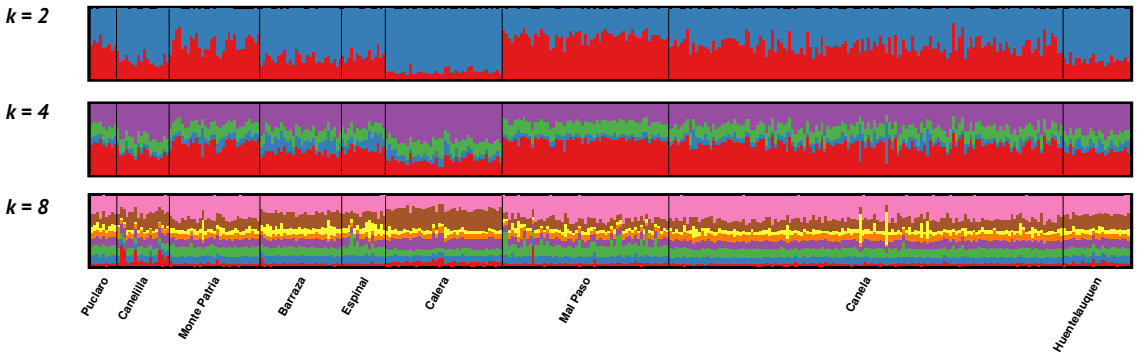

Figure 4.8. Results of STRUCTURE using 15 STR markers for k = 2, 4 and 8. According to the method by (Evanno et al., 2005), no more information is obtained beyond the model of two clusters (k = 2). Population structure is very weak, and all individuals are admixed regarding these two clusters, but with some differences in contribution of each in different populations as can be seen in Calera and Huentelauquen.

Following the method described by (Cardoso et al., 2012), the Q matrix (for k = 2) will be used in combination with the previously mentioned PSA matrix to control for relatedness in further regression models.

4.2.2 Distribution of STR identification markers by groups

The analysis at the village level shows frequencies that are in agreement with HWE in all markers (with the exception of *vWA* in Barraza, $p-value$ = 0.001). As expected for highly variable markers, population differentiation (Fisher's method) is highly significant at both allelic (Overall loci $p-value = 2.159 \times 10^{-8}$) and genotypic levels (Overall $p-value = 7.44 \times 10^{-7}$). This is likely to be a result of the inability of the Fisher's method to deal with the absence of some alleles in most populations (Jost, 2008) (which could be the case here due to sample sizes, but also due to differences in census sizes for each population). When evaluated using Jost's D (see section 2.3.3.4) these differences are not significant (D = 0.0088, multilocus).

Inbreeding (F_{is}) within villages was on average 0.139[7], without significant differences between villages (ANOVA $p-value$ = 0.6 see Figure 4.9 for a visual comparison). Despite a few outliers, inbreeding (as estimated using this markers) is not very high, even within villages. The estimation of F based on STR markers is very similar to the the estimation of F based on surnames[8] (see Table 4.3), with one exception (*Espinal*), with a much smaller list of surnames as shown in Table 4.1). Removing this outlier correlation between both measurements of F_{is} changes from a $p-value$ of 0.07 to 0.008, becoming highly significant. In conclusion, inbreeding within villages is not as high as we originally though, and is not higher than inbreeding considering all villages as a single population, rejecting the hypothesis of strong population differentiation based on this

[7] The reported F_{is} = 0.139, is slightly higher than most world populations and other Latin American admixed populations, but much lower than small native populations form the Americas and Oceania as reported by Pemberton & Rosenberg (2014). According to them, F increases with distance from Africa, showing the highest levels in Native Americans and Natives from Oceania.

[8] Usually approximated to $2.5/\alpha$ (Lasker, 1977)

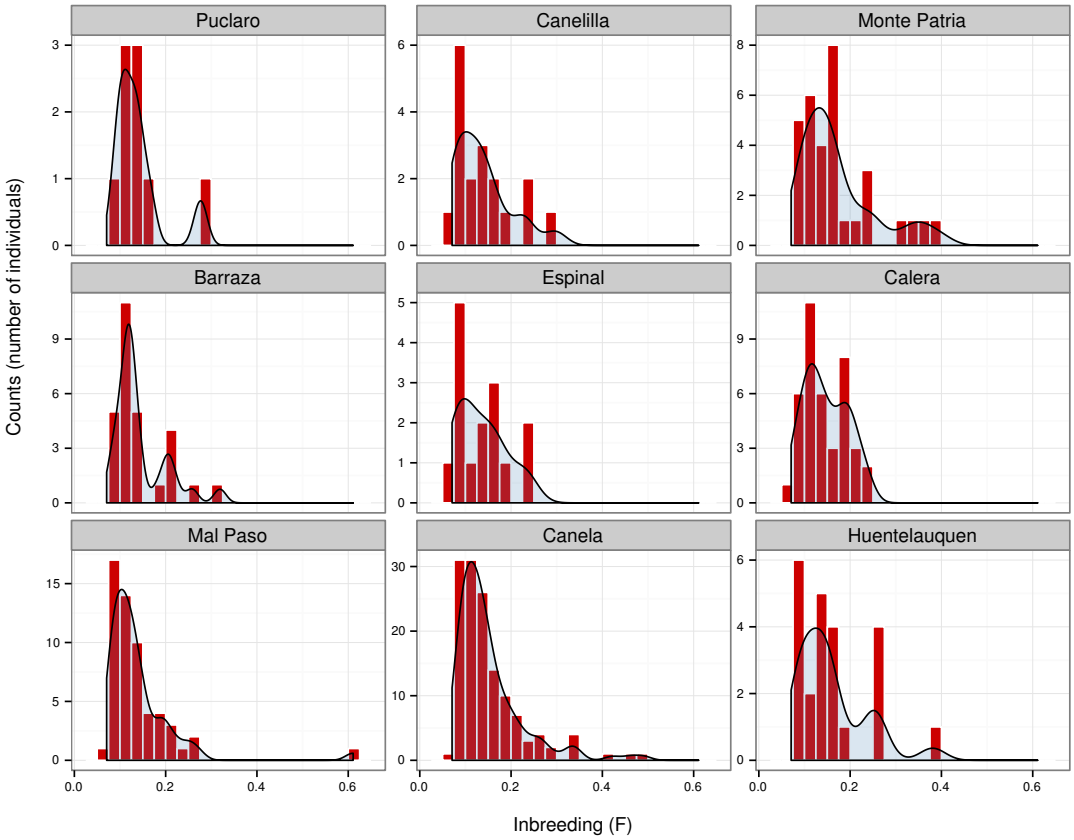

Figure 4.9. Histograms for each village based on the frequency of mean estimation of inbreeding: $\overline{x}$ = 0.139, average *sd* = 0.06. Superimposed: Kernel density curve of the likelihood distribution of homozygosity due to a common ancestor of both parents.

measure.

Analysis using proportion of shared alleles does not show further population structure at the village level. A principal component analysis of PSA does not cluster individuals from the same village (see 4.10), and although a few groups of clustered outliers correspond to reported relatives in the questionnaires, there is not systematic clustering, rejecting the hypothesis of a higher proportion of shared alleles within villages.

Village	*F* surnames	*F* STRs
Puclaro	0.169	0.139
Canelilla	0.143	0.137
Monte Patria	0.097	0.171
Barraza	0.145	0.143
Espinal	0.332	0.136
Calera	0.129	0.148
Mal Paso	0.15	0.139
Canela	0.091	0.152
Huentelauquén	0.095	0.159

Table 4.3. Comparison of estimations of Inbreeding based on surnames and based on 15 autosomal STR markers.

Through this section, different methods to examine forensic identification STR markers were used to detect population structure due to relatedness. While at individual level, a combination of proportion of shared alleles and STRUCTURE *Q* matrix has been showed as a successful method to account for cryptic relatedness (Cardoso et al., 2012), at the location level neither distance/structure methods (F_{st}, Joost's *D*), inbreeding, or PSA show evidence of differences, and therefore villages

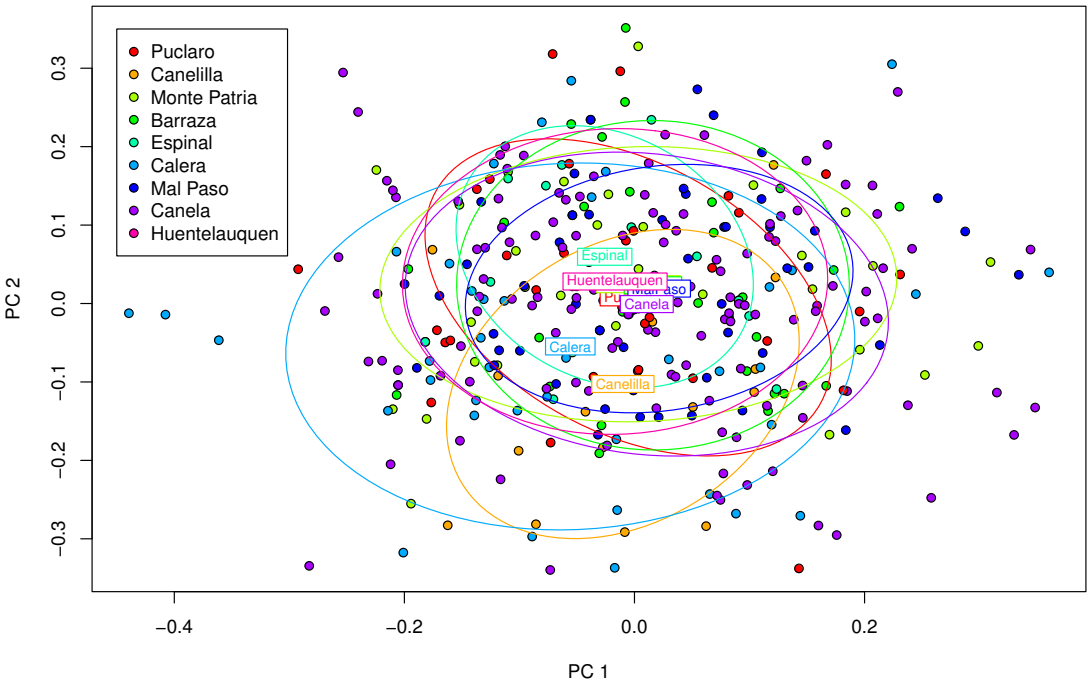

Figure 4.10. Principal component analysis of proportion of shared STR alleles between individuals. Individuals from the same village do not cluster according to shared alleles.

behave as a single population.

The only suggestion of population structure using these markers, although not particularly strong, is presented in the results from the clustering algorithm of STRUCTURE. As mentioned above (see Figure 4.8), the optimal model, based on two clusters $k = 2$, classifies all individuals as admixed (with respect to these STR markers), yet some differences across villages can be identified in the proportion of the contribution of this two clusters. Distruct 1.1 (Rosenberg, 2003) was used to plot the average composition of clusters per village in Figure 4.11.

This analysis shows some weak population structure under the model of two clusters $k = 2$, which

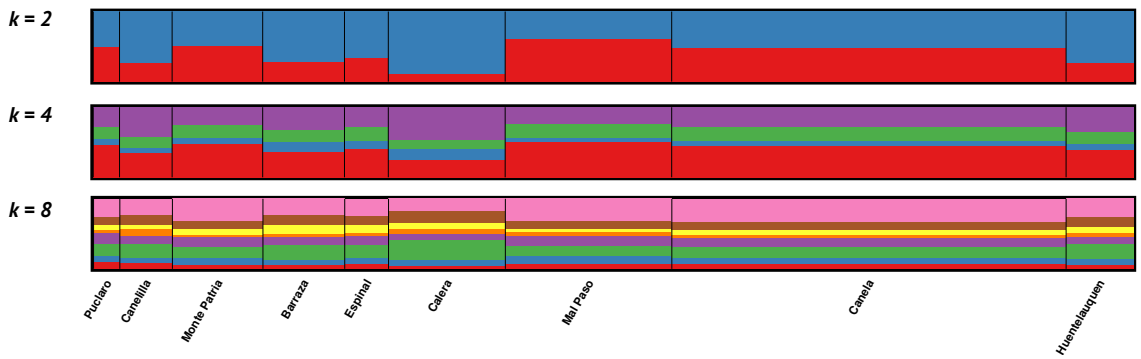

Figure 4.11. Per village average of the results of STRUCTURE using 15 STR markers for $k = 2$, 4 and 8. According to the method by (Evanno et al., 2005), no more information is obtained beyond the model of two clusters ($k = 2$). At $k = 2$, most villages show similar proportions of assignment to both groups, but a few villages (particularly Calera) show an average with very high contribution of only one cluster.

diminish as more clusters are added. Most villages show similar proportions of assignment to both groups, but a few villages (particularly Calera) show an average with very high contribution of only one cluster. Since these are averages of values obtained at individual level, it is expected that controlling for the individual Q matrix for $k = 2$ will also take into account any of the structure at village level. In any case, population differentiation at these markers is very week. Examined as village averages, these two clusters do not seem to correspond to European vs American ancestry, as will be examined further in the next section.

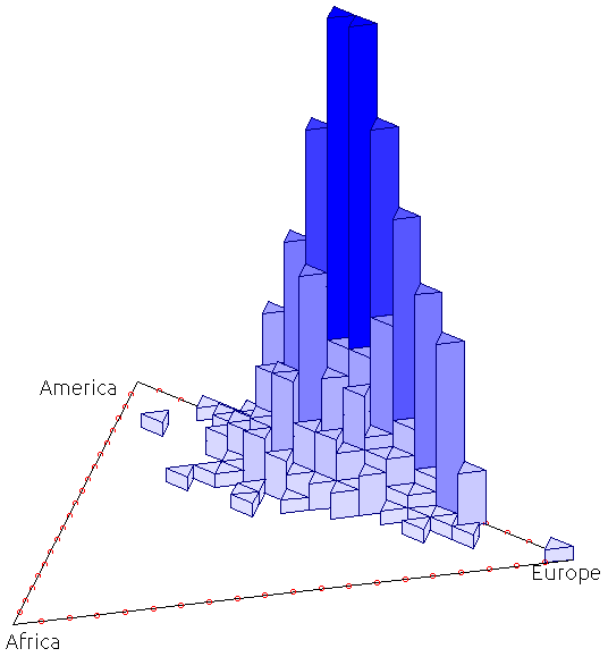

Figure 4.12. 3D histogram estimated ancestry proportion from three parental populations: European, African, and Native American. Distribution by number of individuals (vertical axis) is concentrated between European and Native American ancestry, with little contribution from Africa. (Plot courtesy of Kaustubh Adhikari, Andrés Ruiz Linares Lab, UCL GEE.)

4.3 Ancestry

4.3.1 Distribution of ancestry informative markers at individual level

Population structure can be the result of differential genetic contribution of distinct parental populations, as is likely to be the case for Latin America. Ancestry was analysed using the 30 ancestry informative markers (AIMs) described in Table 2.3. Genotypes were obtained for 437 individuals, yet some with too many missing genotypes (more then 20% failure rate) were removed from some of the analyses. All loci are in HWE equilibrium, with the exception of rs4145160, which shows an excess of homozygotes in one village ($p-value$ = 0.01, Calera).

These markers were used to estimate proportions of European, African, and Native American[9] ancestry for each individual using Admixture 1.23 (Alexander et al., 2009). Individuals were allocated into different ancestry proportions according to their similarity to a reference dataset of 876 individuals from the parental populations (299 Europeans, 169 Africans, and 408 Native Americans) provided by Andrés Ruiz-Linares Lab at UCL GEE (see section 2.3.3.1 for details). The resulting estimated ancestry proportions are shown in Figure 4.12.

This result shows similar proportions (~50%) of European and Native American ancestries in most

[9] Several terms are used in the literature as collective names for the pre–Columbian populations of the Americas, such as *"Amerindian"*, *"Native American"*, *"Indigenous"*, *"Aboriginal"*, etc. Here, the terms *"Native American"* is used in reference to this ancestry, and not as an exclusive term referring to Native indigenous peoples of the United States, nor to citizens of the United States, as is the general usage in English.

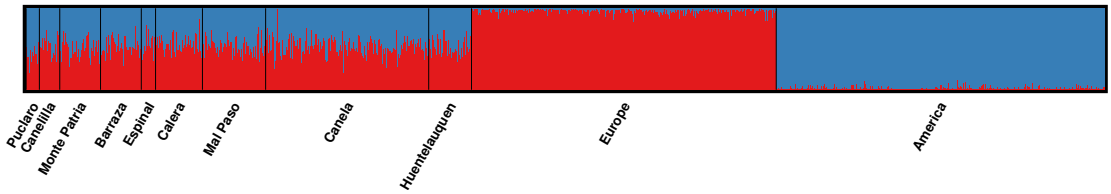

Figure 4.13. Individual STRUCTURE clustering using 30 Ancestry Informative Markers for k = 2 and including a set of European and Native American samples. Both European and Native American samples are successfully identified as distinct clusters, while most individuals from the Chilean communities have ancestry from both populations, with variable proportions of assignment to each group.

individuals, covering a wide range from mostly American to mostly European. By contrast, the contribution of African ancestry is minimal (< 5%). A summary of these admixture proportions was introduced in section 3.3.1. These estimated proportions of ancestry will be used as covariates in regression models introduced in the next chapters.

Ancestry estimations computed with the method of Admixture 1.23 were compared with the clustering algorithm of STRUCTURE using the same set of SNPs, but including only European and Native American parental populations and without prior information of ancestry (i.e. without specifying which samples belong to parental populations). As for the STR analysis, STRUCTURE was run 10 times with a length of 100,000 burn-in periods and 100,000 MCMC repeats, using the Admixture LOCPRIOR model and different values of α for each population, but this time with a pre-defined number of clusters (k = 2) in order to evaluate the assignment of admixture proportions. All other parameters were set to default. Results were processed using CLUMPP 1.1.2 (Jakobsson & Rosenberg, 2007) and Distruct 1.1 (Rosenberg, 2003) and can be visualised in Figure 4.13.

STRUCTURE successfully identified the parental populations as separated clusters, while grouping the samples from the Agricultural Communities as admixture of these two. The estimated ancestry proportions were similar using STRUCTURE and Admixture (R^2 = 0.588, Residuals mean = 2.9×10^{-18}, Residuals standard deviation = 0.08), although results based on STRUCTURE show a slightly higher degree of Native American ancestry than results based on Admixture. Despite this, the general similarity allows us to consider only one of these measures as a control for ancestry and thus reducing the the number of variables to account for in further analyses, since adding a variable compromises power and adds error. Admixture results are summarised in Figure 4.12 and in section 3.3.1.

4.3.2 Distribution of ancestry informative markers by groups

Genotype frequencies for the ancestry markers are in HWE in all villages, with the exception of rs4145160 in Calera ($p-value$ = 0.01). Evaluated using the Fisher's exact method, population differentiation is significant (Allelic $p-value$ = 7.08×10^{-5}, Genotypic $p-value$ = 0.0001), and par-

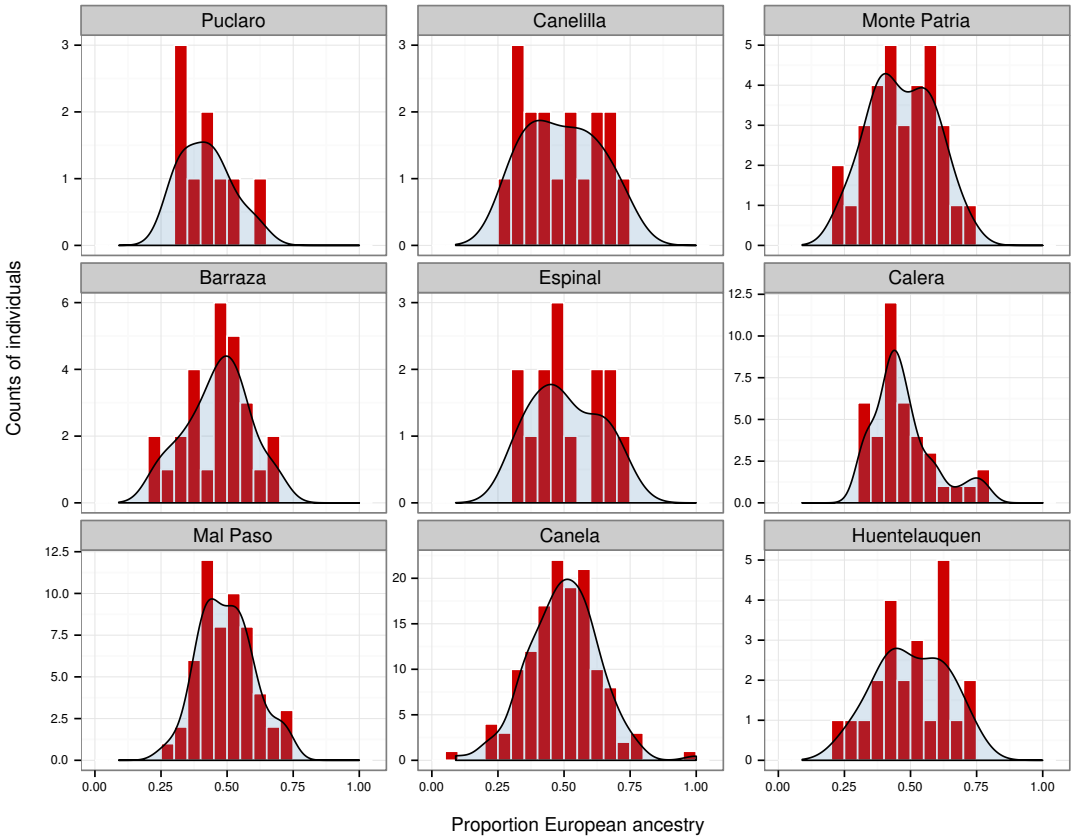

Figure 4.14. Histograms for each village based on the number of individuals at each value of proportion of European ancestry: $\bar{x}$ = 0.488, average *sd* = 0.125. Superimposed: Kernel density curve of the likelihood distribution of proportion of European ancestry.

ticularly high for some loci and for some pairs of populations (e.g. for Puclaro and Calera/Espinal, which show frequencies ~0.9 for some Native American and European ancestry informative markers respectively). However, genetic distances between all pairs of populations are not significant overall (F_{st} = 0.0049, Joost's D = 0).

Table 4.4 and Figure 4.14 shows the proportion of European ancestry per village as estimated by Admixture 1.23. All locations show similar proportions of European and Native American admixture, and no statistically significant differences in mean European ancestry (ANOVA $p-value$ = 0.5), but a small increment towards the south ($p-value$ = 0.009). Although weak, this north-south cline of ancestry proportions has been previously reported for the Coquimbo region (Acuña et al., 2000), and might be related to the differences in frequencies of $-13{,}910^{*}T$ examined in Chapter 3.

Location	$\bar{x}$	sd	min	max
Puclaro	0.419	0.098	0.303	0.601
Canelilla	0.492	0.141	0.276	0.734
Monte Patria	0.468	0.124	0.238	0.714
Barraza	0.464	0.122	0.222	0.691
Espinal	0.505	0.132	0.321	0.724
Calera	0.473	0.118	0.306	0.788
Mal Paso	0.497	0.104	0.261	0.726
Canela	0.499	0.133	0.09	1
Huentelauquen	0.504	0.135	0.241	0.735
All	**0.488**	**0.125**	**0.09**	**1**

Table 4.4. General statistics of estimated proportion of European ancestry per village.

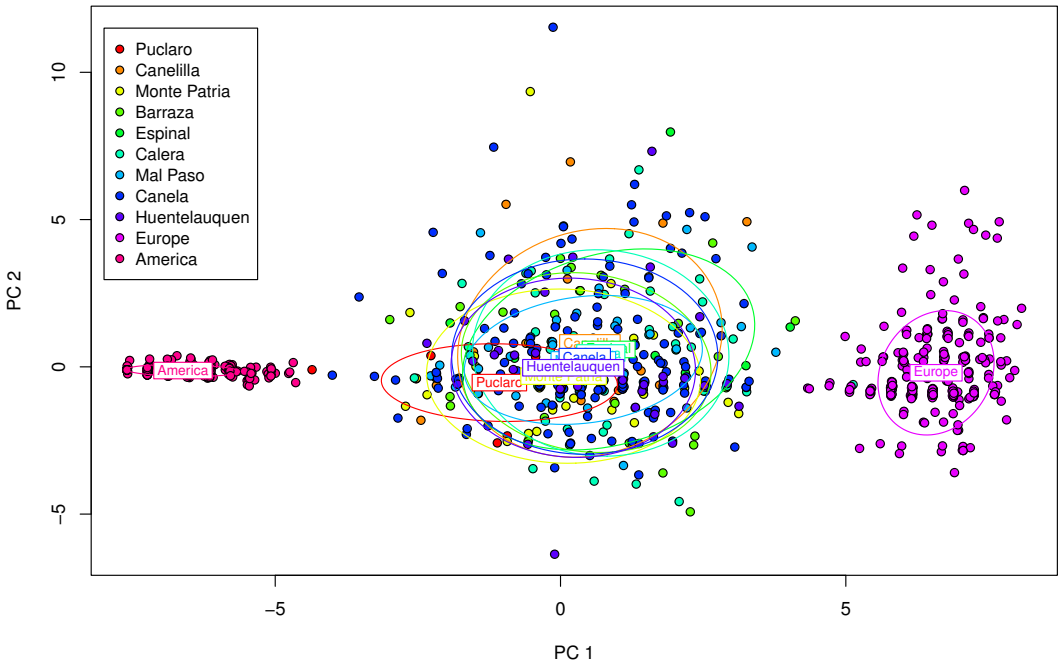

Figure 4.15. Principal component analysis of ancestry informative markers using 299 European samples and 408 Native American samples for comparison. All locations show similar proportions of European and American ancestry, with slightly more Native American markers in Puclaro. The first component accounts for 44.78% of the variance and distinguishes between the two continental group and places the admixed group in the middle. The second component accounts for only 4.02%.

Figure 4.15 further explores differences in ancestry informative markers between locations using Principal Components Analysis. The first component accounts for 44.78% of the variance, while the second component accounts for only 4.02%. Parental populations are clustered at opposite ends of the first component, as expected, while all Chilean locations are clustered together in the middle, showing similar proportions of European and American ancestry, with only Puclaro grouping closer to the Native American cluster.

To summarise, this set of ancestry informative markers is effective in identifying Native American and European individuals and assigning ancestry proportion to admixed individuals and admixed subpopulations regardless of differences in methods, as can be asserted from the similar results provided by both Admixture and STRUCTURE. The population of Andean goat herders does not show significant proportions of African ancestry, and can be better described as admixed Native American and European. Analysed using this bi−parental model, estimated ancestry proportions are similar from both parental groups, with some variability at both individual (Figure 4.13) and subpopulation (Figure 4.16) levels.

The analysis of ancestry does not show significant structure between villages, yet an important variability at the individual level. This is very likely to constitute a confounding factor for further

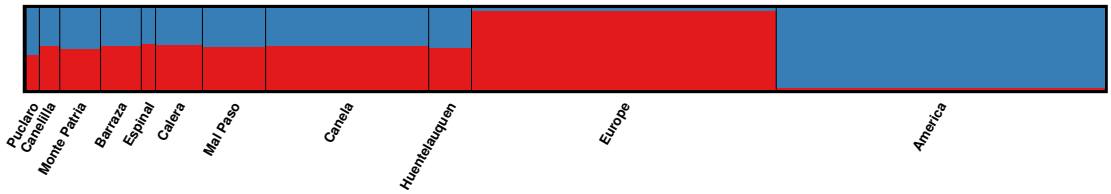

Figure 4.16. Per village STRUCTURE clustering using 30 Ancestry Informative Markers for k = 2 and including a set of European and Native American samples. Both European and Native American samples are successfully identified as distinct clusters. All villages show similar proportions of both European and Native American ancestries. While there is variability at individual level in all villages (see 4.13), Native American ancestry in Puclaro is systematically higher, while differences between the other locations are small.

analysis, particularly under the very likely assumption of a European origin of the allelic variant causative of lactase persistence in this population. By contrast, the analysis of relatedness does show some variability at village level according to the STRUCTURE algorithm, particularly in Calera, which need to be examined in relation to lactase persistence. These assumptions will be explored in the next sections of this chapter.

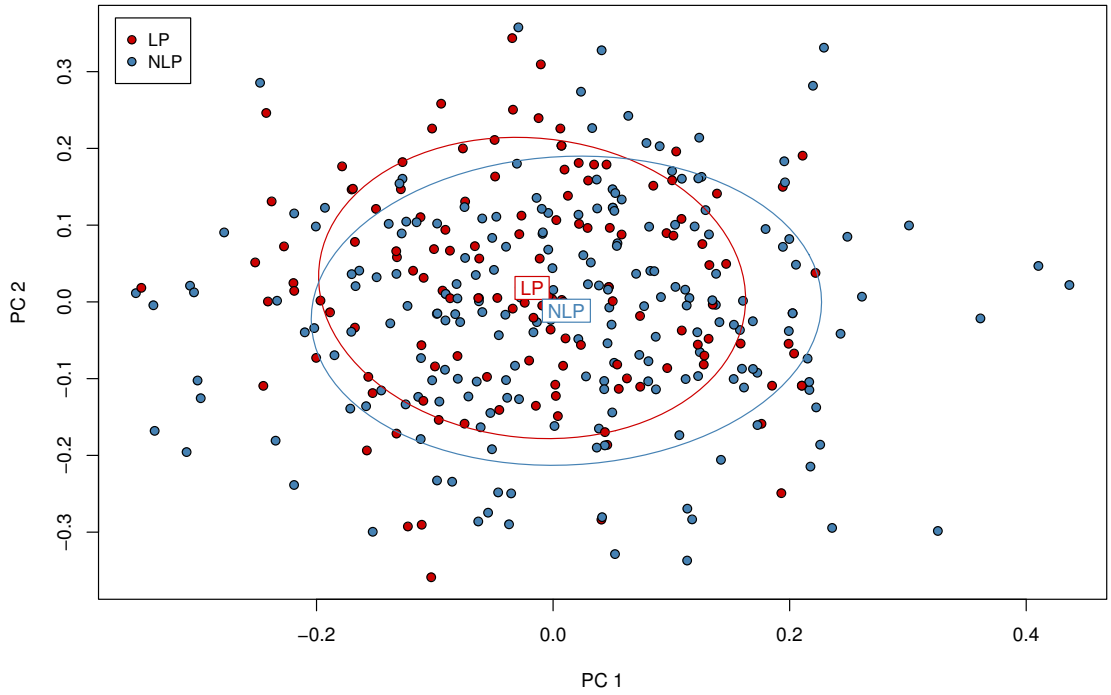

Figure 4.17. Principal coordinates analysis of proportion of shared alleles in 15 multiallelic loci, divided by predicted lactase persistence (LP: Lactase−persistent, NLP: Lactase non−persistent). There is no clustering of individuals according to their predicted digestion status.

4.4 Population structure and lactase persistence

4.4.1 Predicted lactase persistence status and relatedness

Predicted lactase persistence according to $-13{,}910$ genotype is not associated with estimated inbreeding or distribution of STR markers in the principal components analysis. Estimated values of $\overline{F}$ (average inbreeding) are not significantly different (t−test $p-value$ = 0.25) between lactase non−persistent ($\overline{x}$ = 0.14) and lactase−persistent ($\overline{x}$ = 0.15) groups.

Figure 4.17 shows the result of a principal components analysis based on proportion of shared STR alleles grouped by predicted lactase persistence after removing individuals with unknown genotype. Cumulative projected inertia is very low (Axis 1 = 2.3%, Axes 1 and 2 = 4.5%), and individuals are not clustered according to their predicted lactase persistence status.

Figure 4.18 shows STRUCTURE clustering of the STR markers using predicted lactase persistence status as if they were two different populations. Both clusters are homogeneously assigned to lactase−persistent and non−persistent groups. This result suggest that even if there is some undetected systematic underlying structure of these markers between groups (as shown in Figures 4.8 and 4.11), this is not associated with lactase persistence status, and thus not likely to cause spurious associations when evaluating the effect of lactase persistence in other traits.

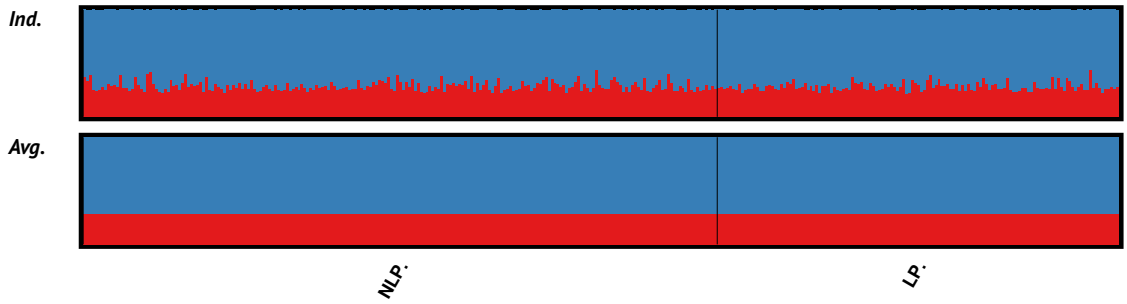

Figure 4.18. Results of STRUCTURE using 15 STR markers for k = 2 according to predicted lactase persistence status (LP: Lactase−persistent, NLP: Lactase non−persistent). Above: Individuals. Below: Averages. No differences were found between lactase persistence and non-persistence in the clustering of STR markers

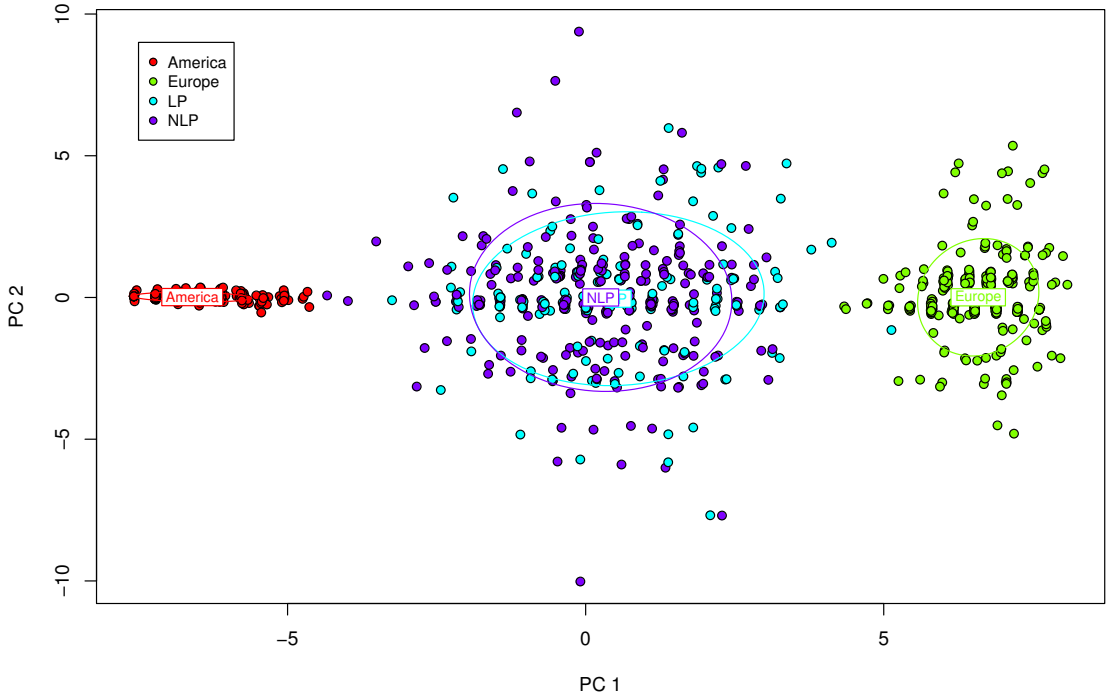

Figure 4.19. Principal components analysis of ancestry informative markers by predicted lactase persistence (LP: Lactase−persistent, NLP: Lactase non−persistent). There is no clustering of individuals according to their predicted digestion status

4.4.2 Predicted lactase persistence status and ancestry

Predicted lactase persistence according to *−13,910* genotype is not associated with estimated proportion of European ancestry, which is, on average, not significantly higher in predicted lactase−persistent than in non−persistent (LP = 0.5, NLP = 0.48, t−test $p-value$ = 0.062). Similarly, principal component analysis of these AIMs (Figure 4.19) shows no evidence of clustering between persistent and non−persistent and no clustering is observed for *−13,910* genotypes (CC, CT, and TT), suggesting a weak association between predicted lactase persistence and genome-wide ancestry.

To check this result, the STRUCTURE clustering algorithm was run using the ancestry informative markers and grouping Chilean samples according to predicted lactase persistence status as if they

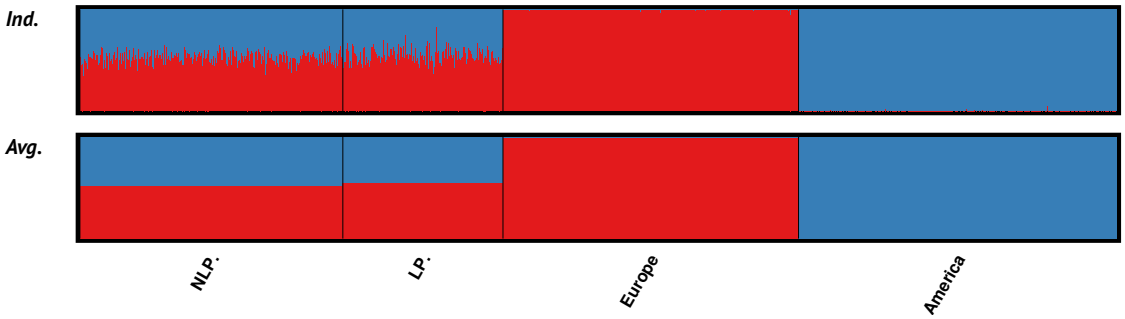

Figure 4.20. Results of STRUCTURE using 30 AIM for k = 2 according to predicted lactase persistence status (LP: Lactase–persistent, NLP: Lactase non–persistent). Above: Individuals. Below: Averages. No statistically significant differences were found between lactase–persistent and non–persistent in the clustering of ancestry informative markers

were two different populations. The output (Figure 4.20) shows very little differences in average European ancestry, though ancestry proportions are very variable at individual level in both groups.

These results might suggest that there has been enough hybridisation to evenly introduce *−13,910* across the population, regardless of genome-wide ancestry, as a product of enough chromosomal segregation and recombination since the initial admixture event. Nevertheless, over the likely 15–20 generations since contact, the European origin of *−13,910*T* should be traceable using phased genotypes surrounding the LCT enhancer region, which should show a European background in haplotypes carrying *−13,910*T*. This possibility will be explored in the following section.

4.4.3 Haplotypic background of LCT enhancer region

A remarkable feature of *−13,910*T* is its strong association with an extended undisrupted haplotype first described to span ~70 kb termed core haplotype A (Hollox et al., 2001), and then shown to be extended to 0.5 Mb in many people (Poulter et al., 2003). Though this is a common haplotype in many populations, it is in particularly high in frequency (0.86) in Northern Europe. The haplotypic background of *LCT* in the Chilean sample was analysed and compared with samples from the Old World, as a method of confirming the European origin of *−13,910*T* in the population of the Agricultural Communities.

Gametic phase of 27 SNPs (in Table 2.4 plus *−13,910*) in 621 individuals (1242 chromosomes: 862 from our Chilean sample and 380 from Old World populations) was obtained using PHASE 2.1.1 (Crawford et al., 2004; Li & Stephens, 2003; Stephens & Donnelly, 2003; Stephens & Scheet, 2005; Stephens et al., 2001) as described in section 2.3.3.3. These SNPs span from ~7.9 Mb distal to ~9.7 Mb proximal from *−13,910*, covering a region of ~17.7 Mb:

A total of 627 haplotypes were identified and classified into core haplotypes. Hollox et al. 2001 defined 42 haplotypes using a set of 11 SNPs, of which only four were common worldwide: A, B, C and U. Only two of the core SNPs, $666G < A$ (rs3754689) and $5579T < C$ (rs2278544), were

Haplotype	Africa *n* = 76	Europe *n* = 110	Middle East *n* = 118	Other Asia *n* = 76	South America *n* = 862	**All n = 1242**
A	0.16	0.73	0.36	0.39	0.5	**0.48**
B	0.26	0.15	0.41	0.47	0.23	**0.25**
C	0.46	0.12	0.2	0.13	0.26	**0.24**
Other	0.12	0.01	0.03	0	0.02	**0.02**

Table 4.5. Frequencies of core haplotypes A, B and C in Africa, Europe, Middle East, Other Asian populations, and Agricultural Communities from Chile (labelled as South America). Haplotype A is the most common in Chile, and is at frequencies between Europe and Asia. Haplotype C is at higher frequencies than both Europe and Asia.

included in the set of 27 SNPs genotyped for our sample (see Table 2.4), and were used to classify the inferred haplotypes into groups A, B, C, or Other. With only two haplotype–defining markers, this classification is not completely unambiguous with respect to haplotypes carrying the same combination of alleles at these two markers, though all other relevant haplotypes are uncommon outside Africa (frequency < 0.02)[10].

Table 4.5 shows the frequencies of the core haplotypes in the analysed populations. Haplotype A is the most common in Chile, and is at frequencies between those of Europe and of Asia. The other two haplotypes are difficult to interpret without knowledge of haplotypic background of the *LCT* region in Native Americans, in particular, Haplotype C is at unexpectedly high frequencies, higher than both Europe and Asia.

Haplotypes were compared between the Chilean samples and the samples of the Old World dividing the dataset between those with $-13{,}910^*T$, those in A haplotypes with $-13{,}910^*C$, and all other haplotypes. All but one of the $-13{,}910^*T$ alleles in the Chilean sample are on A core haplotypes[11].

A visual comparison of the extension of the different haplotypes is shown using plots of Extended Haplotype Homozygosity decay (Figure 4.21). Extended Haplotype Homozygosity (EHH) was measured using Sweep 1.1 (Sabeti et al., 2002) over a region of ~1.77 Mb, defining $-13{,}910$ as the core for each haplotype, and then plotted using GNU R (R Core Development Team, 2013).

Haplotypes A$-13{,}910^*T$ are in an extended block of homozygosity in both Chilean and Old World samples. However, there is some decay in the Chilean sample absent in the Old World. By contrast, A$-13{,}910^*C$ and all other haplotypes show similar EHH decay in both Chilean communities and the Old World.

The analysis of haplotypes shows a high diversity of haplotypes in the samples from the Agricultural Communities, an unexpected result under the hypothesis of a small founder population. Particularly surprising is the diversity of A$-13{,}910^*T$ haplotypes, which was expected to be a small

[10] With the exception of haplotype U, which is common outside Europe and indistinguishable from haplotype B using only these two markers.

[11] An erroneous phasing cannot be ruled out for this exceptional haplotype, who is heterozygous and could have resulted in an A haplotype with other phasing.

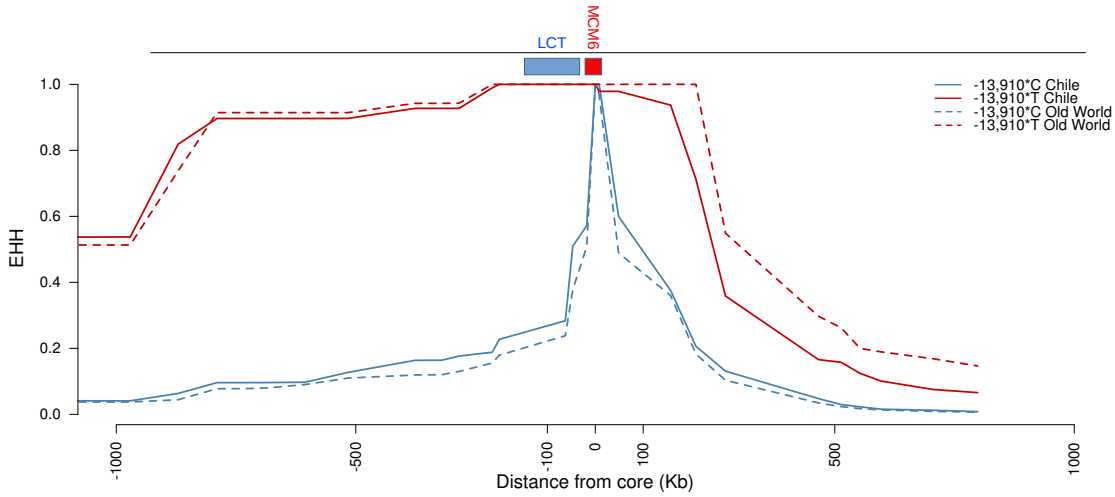

Figure 4.21. Extended Haplotype Homozygosity (EHH) decay in Chile (solid line) and the Old World (dashed line) for *−13,910*T* (red), and *−13,910*C* (blue). Haplotypes A*−13,910*T* are in a extended block of homozygosity in both Chilean and Old World samples. Approximate positions of *LCT* (blue) and *MCM6* (red) are also shown.

subset of those present in Europe given that relatively little recombination is expected over 15–20 generations. This can be interpreted as either a diverse and numerous founding European group, or as continuous gene flow of European haplotypes into the population through migration.

Another interesting finding is the relative high frequency of the C haplotype, which has been previously reported at the most common haplotype in Brazilian Native Americans (Friedrich et al., 2012a). In Brazil, C haplotypes might have been attributable to African gene flow, but this result in a Chilean admixed population with very low African ancestry constitutes an additional support for high frequency of C haplotypes in Native Americans. However, more studies are needed to test this hypothesis. Future work analysing these data adding haplotypes from the 1000 Genomes project (The 1000 Genomes Project Consortium, 2012) would contribute to examine this hypothesis.

In relation to *−13,910*T* in the Chilean sample, a European–like haplotypic background associated with haplotype A in an extended block of homozygosity was found, supporting the idea of a European origin of lactase persistence in this population and the usefulness of estimations of ancestry proportions were used as confounding variables in further models.

Chapter 5

Examining the possible evolutionary advantages of Lactase Persistence

Effects of lactase persistence on milk consumption, height and weight, and mortality and fertility are explored in this chapter. Two possibilities are considered to assess the importance of milk consumption in the possible evolutionary advantages conferred by lactase persistence: Diminished relative fitness of lactase non–persistence by avoidance of milk, or differential nutritional benefits of milk in persistent and non–persistent individuals regardless of consumption. Afterwards, the effects of height and weight are examined taking into account our analyses of population structure to account for as many confounding variables as possible, acknowledging the complexity of these traits. Finally, the effects of lactase persistence on fertility and mortality are analysed.

5.1 Milk consumption and Lactase Persistence

5.1.1 Introduction: Variability in milk consumption behaviour

As has been mentioned in Chapter 1, strong positive natural selection favouring lactase persistence is known to have occurred in the past in some populations, although the specific mechanisms are unknown. Possibilities to take into account include either improved chances of survival or improved fertility for lactase–persistent milk–drinkers (analysed in section 5.3), but not for lactase non–persistent individuals. Diminished relative fitness of lactase non–persistence could be caused by avoidance of milk to prevent digestive effects, or by milk not providing the same benefits to non–persistent individuals even if they drink it. Thus to evaluate the evolutionary effects of lactase persistence, variability in milk consumption needs to be considered as well as variability in the genetics of lactase persistence: if drinking milk is advantageous in terms of natural selection some degree of variability in consumption is needed to make meaningful comparisons between drinkers and non–drinkers. This section will therefore analyse variability in milk consumption behaviour that will be later associated with other variables in the following sections.

A summary of milk consumption in the population is shown in Table 3.6. In this section we explore how milk consumption is distributed in the groups of interest. Different methods employed to estimate milk consumption were described in section 2.3.3.4. The results in this chapter are based on the estimations of daily cups of milk and daily portions of milk products for the reasons outlined in section 2.3.3.4[1].

People who reported feeling unwell with milk consume significantly less milk (ANOVA p = 0.006), even controlling for sex and age (average number of cups per day of 0.45, compared to 0.64 in people who does not feel unwell with milk). Some possible explanations to milk drinking behaviour in people feeling unwell with milk are its availability and the spread social perception of milk as healthy (Kingfisher & Millard, 1998). With respect to these other variables, sex is not related to milk consumption (p = 0.5), but females do more often report feeling unwell after drinking milk than males ($\chi^2 p-value$ = 0.002). On the contrary, milk consumption significantly increases with age (regression β = 0.007 cups per year, $p-value = 7 \times 10^{-4}$, $F-statistic$ = 11.56, df = 348) without statistically significant differences on average age between those not reporting and reporting milk digestion problems. No statistically significant association was found between milk consumption and livestock owning ($t-test\ p-value$ = 0.3), wealth (Pearson's correlation $p-value$ = 0.2), or village (ANOVA p = 0.5).

[1] Participants preferred to refer to their milk consumption in cups, and therefore were used as measuring units. A cup is approximately 250 ml.

Consumption of processed milk products is negatively associated with feeling unwell with milk (Fisher's Exact Test $p-value$ = 0.004), and is not statistically different by sex groups or by owning livestock. Neither average age nor wealth show statistically significant differences between consumers and non consumers of milk products.

The most significant aspects of distribution of milk consumption variability are the increased consumption with age and the avoidance of milk and milk products by those who report having digestive problems with milk. This association with age seems very surprising, and although Chilean public health service provides free powdered milk to people over 70 years old, the trend of increased milk consumption with age remains significant excluding people over 70. Maybe they are from a generation when milk consumption and goat milking was a more important part of they livelihood than it is today. In the following section these and other aspects of milk consumption behaviour are analysed in association with predicted lactase persistence status.

5.1.2 Milk consumption behaviour, lactose tolerance, and $-13{,}910$ genotype

Avoidance of milk and milk products motivated by digestive symptoms in the lactase non–persistent could imply an improved fitness in those who can drink milk and get the nutritional benefits from it, in comparison to those who cannot. However, neither self–reported digestive problems with milk nor milk consumption are associated with predicted lactase persistence in this population.

Analyses were done for feeling unwell with milk, cups of fresh milk per day, and consumption of milk products, which were tested for association with results of breath hydrogen lactose tolerance test, $-13{,}910$ genotype, and lactase persistence status predicted from $-13{,}910$.

In relation to self-reported digestive problems, feeling unwell with milk is not statistically associated with any genotype ($\chi^2 p-value$ = 0.6812), lactase persistence status predicted from genotype (Fisher's Exact Test $p-value$ = 0.92), nor lactose digester status according to results of lactose tolerance test (Fisher's Exact Test $p-value$ = 0.32).

With respect to consumption of fresh milk, no statistically significant association was found between estimated number of cups of fresh milk per day and genotype (mean *CC* = 0.55, mean *CT* = 0.52, mean *TT* = 0.72, ANOVA p = 0.47). And though $-13{,}910$*T homozygous have higher consumption, this is not statistically significant ($t-test\ p-value$ = 0.4). Milk consumption is also not statistically associated with predicted lactase persistence status (mean digester = 0.55, mean non-digester = 0.55, $t-test\ p-value$ = 0.94), nor with lactose tolerance tests (mean tolerant = 0.54, mean intolerant = 0.54, $t-test\ p-value$ = 0.99). Similar results were found in the analysis of con-

sumption of milk products, with no association between avoidance of dairy and genotype ($\chi^2 p-value$ = 0.4), predicted digester status (Fisher's Exact Test $p-value$ = 0.92), or lactose tolerance (Fisher's Exact Test $p-value$ = 1).

These results show no avoidance of dairy products and milk in lactase non–persistent individuals, and no association between feeling unwell with milk and being lactase non–persistent. One possible interpretation would be secondary lactose intolerance caused by digestive problems other than lactase persistence, and while that would explain milk maldigestion in $-13{,}910^*T$ carriers (only 15% of the subjects), it would not account for predicted non-digesters reporting no symptoms (37% of subjects), who would have appeared as predicted digesters and intolerant in the lactose tolerance test (no cases in our study). On the other hand, there are several ways to reduce digestive symptoms and drink milk (as discussed in section 1.2), and symptoms can be ignored if they are mild or the subject is accustomed to them.

Non-persistent individuals consume as much lactose as persistent individuals in this population[2]. Therefore, if milk consumption mediates in a selective advantage favouring $-13{,}910^*T$, milk avoidance of individuals homozygous for $-13{,}910^*C$ does not seems to be the cause according on this study. However, the hypothesis of different nutritional effects of milk consumption in lactase–persistent and non–persistent individuals offers an alternative way for selection to operate, as will be examined in section 5.2.

[2]Although this cannot be extrapolated as the prevalent behaviour where and when lactase persistence evolved by selection, as will be discussed in Chapter 6.

5.2 Lactase persistence, height and weight

5.2.1 Height and weight are complex traits

Height and weight are two of the most complex phenotypic traits in humans. The continuous nature of height and weight presented an early challenge to the notion of discrete Mendelian inheritance because of the evident similarities of height and weight between relatives (Fisher, 1919). Mendelian and biometric views on heritability were reconciled with the idea of multiple genetic and environmental factors contributing together to the phenotypic expression of a continuous trait, founding the field of quantitative genetics (Fisher, 1919; Orr, 2005).

The important role of genetics in height and weight is strongly suggested by several studies with large datasets of relatives and twins, with estimations of heritability between 0.87–0.93 for height, and 0.5–0.9 for BMI (Dubois et al., 2012; Schousboe et al., 2003; Silventoinen et al., 2003).

While a polygenic determination of height and weight has been acknowledged since the modern synthesis (Fisher, 1919), the emergence of genome–wide association studies provided evidence of a particularly large number of loci associated with both traits: more than 180 and 165 loci across all the genome are likely to have an important effect on height and weight respectively (Bell et al., 2005; Lango Allen et al., 2010).

A sustained increment of height has been widely documented, and has been commonly attributed to the improvement of living conditions since the beginning of the 20^{th} century (Cole, 2000; Gustafsson et al., 2007; Silventoinen et al., 1999, 2003). Comparative studies of height between developed and developing countries suggest that both improvement in nutrition and general health conditions during childhood are responsible of this trend (Silventoinen, 2003; Steckel, 1995). Similarly, the availability of food and sedentary lifestyle has lead to an increment of BMI in developed countries (Bleich et al., 2008) and, more recently, also in the developing world as industrialised lifestyle and diet are adopted (Bell et al., 2005; Hoffman, 2001; Hossain et al., 2007). These observations lead to the conclusion that, in addition to its genetic complexity, environmental factors can dramatically affect height and weight.

Both height and weight are very likely to have been affected by recent natural selection. For instance, the exceptional height of short populations or “pygmies” has been explained as resulting from an earlier onset of reproduction in the context of high mortality rates in youth and early adulthood (Migliano et al., 2007; Stock & Migliano, 2009). In a similar fashion, the increasing prevalence of obesity and related diseases has been often interpreted as resulting from a metabolic syndrome (Luca et al., 2010; Weiss et al., 1984) and a thrifty genotype (Neel, 1962), under the

assumption that genetic traits associated with an increment in weight have been advantageous in a context of low food availability and constant threats of famine in the past, but became maladaptive in the context of modern food availability and lifestyle. This hypothesis has been criticised due to the assumption it makes about the extent of mortality in periods of famine, and the lack of evidence of signatures of selection for variants with already well–established associations with weight (Gibson, 2007; Speakman, 2006).

On average, men are taller than women in all populations, while BMI tends to be greater in females. Additionally, there is evidence of contribution of Y–chromosome genes to stature (Kirsch et al., 2002; Salo et al., 1995), and and effect of sex hormones in height and weight (Albertsson-Wikland et al., 1994; McTiernan et al., 2006; Power & Schulkin, 2008). This pattern of sexual dimorphism in both traits has made them subject to several hypothesis based on sexual selection. There is some evidence of positive correlation between stature and reproductive success in males but not in females (Nettle, 2002a,b), as well as between BMI and reproductive success in both sexes (Allal et al., 2004; Power & Schulkin, 2008; Sear, 2006; Weng et al., 2004). Suggested explanations for the dimorphism are the hypotheses of a higher susceptibility to environmental determinants of height in males (Kuh et al. 1991; Stinson 1985, but no evidence was found by Gustafsson & Lindenfors 2004), sex–bias in children's nutrition according to women's contribution to domestic economy (Holden & Mace, 1999), positive correlation between stature and attractiveness in men but not in women (Nettle, 2002a), and male–male intrasexual competition favouring increment in both height and weight (Kanazawa, 2005).

If lactase digestion has an effect on height or weight, as suggested in other studies (Almon et al., 2010; Corella et al., 2011; Lamri et al., 2013), It is also possible that height and weight might mediate in the contribution of lactase persistence to reproductive success. However, many genetic and non–genetic factors not accounted for in this study are very likely to have confounding effects in both height and BMI. Acknowledging this limitation, an attempt to model stature and BMI as responses to lactase persistence is introduced in the next section.

5.2.2 Modelling height and weight as response to lactase persistence

Multiple regression analysis was used to examine the association between predicted lactase persistence and body size, to further explore the hypothesis of different effects of nutrition in lactase–persistent and non–persistent individuals. Models were developed for height and BMI as response variables[3], separately for males and females since this allowed model selection methods to identify different explanatory variables for each sex. For example, if an earlier age at first birth is indicative

[3] Using BMI as a measure of weight controlling for height.

of an earlier menarche, the age at first birth could have an effect in height and BMI in females, but not in males.

The following list describes the variables considered relevant to build a full model. The description also contains abbreviations used in tables through this section.

- **Lactose digestion status** predicted by *−13,910* genotype. Abbreviated as "`digester`".
- **Age**.
- **Daily consumption of milk** estimated from the questionnaire. As described in section 2.3.4.1, this variable was estimated from a scale and therefore is not a proper count of cups as the name of the variable might suggest. Data was log–transformed and abbreviated as "`cups`".
- **Proportion of European ancestry** estimated using ADMIXTURE as described in section 2.3.3.1. Abbreviated as "`eur.anc`".
- **Wealth** estimated using a composite score based on access to goods as described in section 2.3.4.1. Abbreviated as "`wealth`".
- **Inbreeding coefficients** (F_{is}) estimated from distribution of homozygosity of STR markers as described in sections 4.2.1 and 2.3.3.2. Abbreviated as "`F`".
- **Microsatellites structure.** Proportion of assignment to one of two parental clusters inferred from the *Q* matrix of the STRUCTURE algorithm using STR markers (see Figure 4.8), as described in section 2.3.3.2. Abbreviated as "`str.stru`".
- **Age at 1st birth.** This variable was considered relevant only for females. Abbreviated as "`age.1`st`.birth`"

Logit transformation was applied to variables expressing 0 to 1 ratios or proportions (i.e. Proportion of European ancestry, Inbreeding (*F*), wealth, and STRUCTURE assignments). The model was tested for fulfilment of regression assumptions, assessing distribution of studentised residuals and Q–Q plots to test for normality, Durbin-Watson test for autocorrelation, and non-constant Variance score test for homoscedasticity. Tests were done in R using the "`car`" package (Fox & Weisberg, 2011)

Akaike information criterion (AIC) and All Subsets Regression (ASR) were used to narrow down the full models specifically for each sex and response variable, using R packages "`MASS`" (Venables & Ripley, 2002) and "`leaps`" (Lumley & Miller, 2009) respectively, but explicitly retaining predicted digestion status.

Backwards stepwise AIC compared the AIC of the full model with the AIC of all the possible resulting models after removing one variable. Afterwards, the model with the best AIC is selected

Models	Variables (formula)	F−stat	D.F.	$R^2_{adj.}$	p−value	Resid. S.E.
Full						
Males						
BMI	$\sim$ *digester + age + cups + eur.anc + wealth + F + str.stru*	1.548	106	0.033	0.159	3.307
Height	$\sim$ *digester + age + cups + eur.anc + wealth + F + str.stru*	1.305	106	0.019	0.255	7.294
Females						
BMI	$\sim$ *digester+age+cups+eur.anc+wealth+F+str.stru+age.*1^{st}*.birth*	1.52	180	0.022	0.153	4.683
Height	$\sim$ *digester+age+cups+eur.anc+wealth+F+str.stru+age.*1^{st}*.birth*	5.116	181	0.148	$\mathbf{9\times10^{-6}}$	5.497
Best						
Males						
BMI	$\sim$ *digester + age + eur.anc + wealth*	2.675	109	0.056	**0.036**	3.268
Height	$\sim$ *digester + age + cups*	2.339	111	0.034	0.077	7.283
Females						
BMI	$\sim$ *digester + age + eur.anc + str.stru + age.*1^{st}*.birth*	2.399	183	0.036	**0.039**	4.649
Height	$\sim$ *digester + age + cups + wealth + str.stru*	9.214	218	0.156	$\mathbf{6\times10^{-8}}$	5.792

Table 5.1. Comparison of full and best models (according to backward stepwise AIC and All subsets regression) tested for BMI and height in males and females. The best models yield better $R^2_{adj.}$ and p−values than the full models, and improve p−values of effects while keeping the same variables as significant effects.

and the process continues iteratively until the point where variables cannot be removed without resulting in a worse AIC. Since our dataset has missing data and this method requires the same degrees of freedom for all models, multiple imputation was used to obtain 5 datasets for each model, which were pooled to apply backwards stepwise AIC as described in Wood et al. (2008). Multiple imputations were computed using the R package "`mice`" (van Buuren & Groothuis-Oudshoorn, 2011)

ASR computes every possible combination of explanatory variables and ranks them according to their $R^2_{adj.}$ values. This value is based on the more common measure of goodness of fit of a regression model R^2, but is adjusted to the number of explanatory variables used, and thus increases only if new variables improve the model (Kabacoff, 2011). This method can deal with missing data and thus was applied to both the imputed and the original dataset, with similar results to those obtained by backward stepwise AIC. The resulting best models for height in males, BMI in males, height in females, and BMI in females were re−tested for regression assumptions using the non−imputed dataset.

Table 5.1 shows a comparison of all the models with summary statistics based on the non-imputed dataset. The best models yield better $R^2_{adj.}$ and better *p−values* than the full models, and improve *p−values* of effects while keeping the same variables as significant effects.

Table 5.2 shows the effects of each variable in the reduced models. Predicted lactose digestion has a significant effect increasing BMI in males by 2.04 kg/m^2 $\pm$ 0.66 S.E. (t = 3.12, $d.f$ = 109, p = 0.002),

	BMI					Height				
Variables	B	S.E B	β	t−value	p−value	B	S.E B	β	t−value	p−value
Males										
Digester (yes)	2.042	0.655	0.293	3.119	**0.002**	1.576	1.446	0.102	1.09	0.278
Age	0.018	0.017	0.096	1.031	0.305	-0.081	0.039	-0.199	-2.077	**0.04**
Prop. Europ. anc.	-0.251	0.279	0.084	-0.899	0.371	–	–	–	–	–
Cups of milk	–	–	–	–	–	1.955	1.784	0.105	1.096	0.275
Wealth	0.164	0.230	0.066	0.714	0.477	–	–	–	–	–
Females										
Digester (yes)	0.514	0.706	0.053	0.728	0.468	-0.196	0.806	-0.015	-0.243	0.808
Age	0.010	0.021	0.035	0.473	0.637	-0.127	0.024	-0.361	-5.303	$\mathbf{3{\times}10^{-7}}$
Prop. Europ. anc.	-1.153	0.616	-0.177	-2.456	**0.015**	–	–	–	–	–
Cups of milk	–	–	–	–	–	-1.173	1.1	-0.066	-1.035	0.302
Wealth	–	–	–	–	–	0.421	0.398	0.069	1.058	0.291
STRU strs. group	-0.446	0.398	-0.081	-1.121	0.264	0.234	0.459	0.032	0.509	0.611
Age at 1^{st} birth	-0.134	0.066	-0.149	-2.038	**0.043**	–	–	–	–	–

Table 5.2. Effects of variables of interest in the best models according to backward stepwise AIC and All subsets regression. Predicted lactose digestion has a significant effect increasing BMI in males by 2.04 kg/m^2 $\pm$ 0.66 S.E. (t = 3.12, $d.f$ = 109, p = 0.002), but has not significant effect on height or any of the female models.

but has no significant effect on height or any of the female models. Age is the only confounder variable with significant effect on height in both sexes, and BMI in females is significantly affected by proportion of European ancestry and age at 1^{st} birth, decreasing BMI by 1.15 kg/m^2 $\pm$ 0.02 S.E. and 0.13 kg/m^2 $\pm$ 0.06 S.E. respectively.

Noticeably, values of $R^2_{adj.}$ are very low. One of the reasons is the reduced amount of variance we are trying to explain in these models after separating them by sex. The full model with both sexes combined yield values of R^2 = 0.5 and $R^2_{adj.}$ = 0.49 for height, and R^2 = 0.08 and $R^2_{adj.}$ = 0.05 for BMI. However, this also means that other variables not included in the model are affecting what remains of the variance in height, and most of the variance in BMI.

After the analysis based on fixed−effects, these models were compared with mixed−effects models adding the PSA matrix[4] as random−effects as a way to add controls for relatedness. The use of mixed−models using a relatedness matrix to account for cryptic relatedness is a common routine in applied research of species of agricultural importance (Cardoso et al., 2012; Kang et al., 2008; Yu et al., 2006) and model organisms (Kang et al., 2008; Zhao et al., 2007), and has been successfully used in human genetics to account for relatedness in association studies (Hoffman, 2013; Yu et al., 2006).

The PSA matrix was added as random−effect to all models using linear mixed−effects kinship model fit by maximum likelihood (Hoffman, 2013)[5] as implemented in R by the package "coxme" (Therneau, 2012). Table 5.3 shows results for all mixed-models. In all models, the contribution of

[4] A pairwise matrix of the Proportion of shared alleles based on 15 STR markers. For details see section 4.2.1

[5] Different alternatives of relatedness matrices are present in the literature. A Kinship matrix is used in Hoffman 2013, but we used a PSA instead following the rationale presented in Chakraborty & Jin, 1993; Zhao et al., 2007 and Cardoso et al., 2012.

Models	Log likelihood	n	Var. random-effects	Mixed-model residual S.E.	$\Delta_{resids.}$ (*fixed* − *mixed*)	Significant fixed-effects ($p < 0.05$)
Full						
Males						
BMI	-293.976	114	8×10^{-4}	3.181	0.126	`digester` (p = 0.002)
Height	-384.132	114	0.004	7.033	0.261	none
Females						
BMI	-555.394	189	0.002	4.57	0.113	`eur.anc` (p = 0.016) `age.1st.birth` (p = 0.032)
Height	-588.785	190	0.002	5.365	0.132	`age` ($p = 1\times10^{-6}$)
Best						
Males						
BMI	-294.187	114	8×10^{-4}	3.195	0.073	`digester` (p = 0.001)
Height	-389.482	115	0.004	7.155	0.128	`age` (p = 0.034)
Females						
BMI	-555.571	189	0.002	4.575	0.074	`eur.anc` (p = 0.003) `age.1st.birth` (p = 0.038)
Height	-708.262	224	0.003	5.714	0.078	`age` ($p = 8\times10^{-8}$)

Table 5.3. Mixed-models of BMI and height with a matrix of proportion of shared STR alleles (PSA matrix) added as random−effects as further controls for relatedness. In all models, the contribution of the random−effects to the variance of the model is very small, as is the reduction of residual S.E. achieved by adding the PSA matrix. Mixed-models identify the same fixed-effects as significant, and the p−value of each variable is always lower than in the model without random−effects

the random−effects to the variance of the model is very small, as is the reduction of residual S.E. achieved by adding the PSA matrix. Mixed−models identify the same fixed−effects as significant, but the *p−value* of each variable was slightly lower than in the model without random−effects shown in Table 5.2.

These results confirm the trend already reported in Chapter 3 of an increased BMI in lactase−persistent compared to non−persistent males. This trend is significant and remains so even controlling for age, milk consumption, ancestry, wealth, inbreeding and relatedness. However, the detected trend for height reported in Chapter 3 is not significant, controlling for other variables. This seems to be caused by differences on average age of lactase−persistent and non−persistent males (LP $\overline{x}$ = 50 years, NLP $\overline{x}$ = 57 years). This difference is small, but significant (t−test *p−value* = 0.046), and is possibly the cause of the differences on height in Chapter 3 (LP $\overline{x}$ = 169.67 cm, NLP $\overline{x}$ = 161.86 cm, t−test *p−value* = 0.03). Controlling for age, the effect seems to disappear, and thus the models suggest that variables not included here have a significant contribution to stature in males.

Predicted lactase persistence does not have a significant effect on size in females. Ancestry and age at first birth are better predictors of BMI in women, and age alone is the only significant effect in women's height. A significant effect of lactase persistence in BMI has been already reported

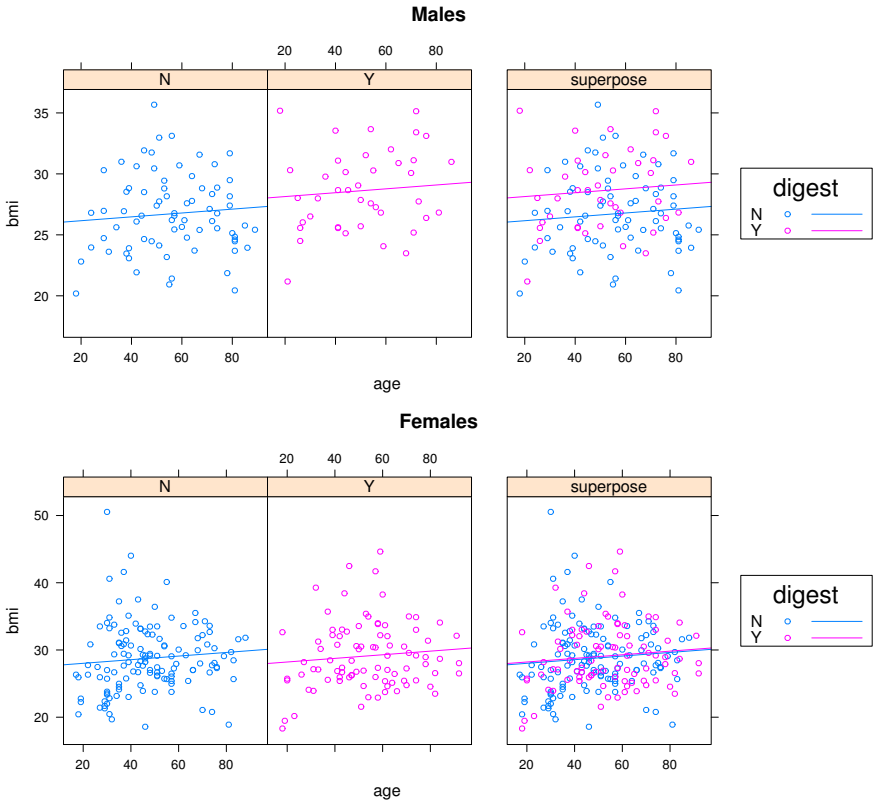

Figure 5.1. Comparison of increment of BMI with age between lactose digesters and non-digesters in males and females. There is a significant difference in BMI between lactase–persistent and non–persistent males. This trend remains significant controlling for age, milk consumption, ancestry, wealth, inbreeding and relatedness. However, there are not significant differences in BMI between lactase–persistent and non–persistent females.

in European populations (Corella et al., 2011; Lamri et al., 2013; Smith et al., 2008), but in these studies the trend is significant for both sexes. This difference in the effect of lactose digestion on BMI for each sex (see Figure 5.1) is difficult to interpret, but some possible hypotheses are explored in Chapter 6.

An increment in BMI could be considered by itself an evolutionary advantage in our recent evolutionary past whenever episodes of famine were common threats. If sex differences detected in this study are not a product of cultural sex–bias in feeding practices, a hypothesis of sexual selection of BMI as an advantageous trait in male–male competition is also plausible[6]. Although this study does not offer ways to test these "*stories*", it is possible to test whether either lactase persistence, BMI, or both have a direct effect in total number of children or child mortality in this dataset, as will be examined in the next section.

[6] And the existence of sex–bias in feeding practices in the past is a sensible possibility.

5.3 Lactase persistence and fitness

5.3.1 Demographic trends

One of the most interesting trends in recent human history has been the demographic transition from high mortality and high fertility towards low mortality and low fertility. This process appears to occur in an specific order: a decline in mortality comes first, with a particularly pronounced decrease in child mortality, which is followed by a decline in fertility afterwards (Wilson, 2011). In between, the period when mortality is low but fertility is high, population growth is particularly high, but becomes stable once fertility starts to decline. Demographic transition in the world has been an asynchronous process. Signals of gradual decline in mortality in Europe appeared as early as the 18th century and were caused by the improvement of general living conditions, while the decline in mortality in developing countries started in the middle of the 20th century, was very sharp and was a result of direct public health intervention policies (Reher, 2012). Accordingly, the decline of fertility in developed countries started not too long after transition to low mortality, with a short period of increased population growth. In contrast, increased population growth has been much larger in developing countries undergoing demographic transition (Reher, 2012, 2004).

The field of evolutionary demography (reviewed in Mace 2000, and Mace 2014) attempts an explanation of the demographic transition in terms of evolutionary theory. Since a trend towards reduction of fertility is, at first glance, selectively detrimental, research in this field has strongly focused on selection.

Three popular hypotheses to explain this dilemma are the quantity–quality trade–off hypothesis, the cultural selection hypothesis, and the maladaptive hypothesis (reviewed in Mulder 1998). The quantity–quality trade–off hypothesis is based on the idea that more resources can be allocated to fewer children, which will increases their fitness in the long–term (Smith & Fretwell, 1974). The cultural selection hypothesis suggests that the trait emerged by imitation of the reproductive behaviour of successful individuals, as a rule of thumb for success, which led to fertility reduction despite their detrimental effects in terms of biological fitness (Boyd, 1988). And the maladaptive hypothesis suggests that low fertility was not a result of natural selection but a by–product of contraception (Perusse, 1993). All these approaches have critical flaws: The fact that numerous families, although not successful in other aspects of life, have better reproductive success (even in the long term) than high–quality small families seems to contradict the trade–off hypothesis (Kaplan & Lancaster, 1995; Kaplan et al., 1995). The motivations of successful groups to adopt low reproduction, and how this ultimately contributes to biological fitness is not explained by the cul-

tural selection hypothesis (Mulder, 1998), and reduction of fertility occurs before the introduction of modern contraception in Europe, contradicting the maladaptive hypothesis (Mulder, 1998; Reher, 2012). Attempts to modify and combine aspects of these three point of views are likely to yield better explanations (Mace, 2000; Mulder, 1998); however, the problem remains largely unsolved.

Apart from its causes, the study of the evolutionary consequences of demographic transition has received much less attention, but the scenario of low mortality and low fertility could produce a pressure favouring traits related to fertility over traits improving child survival (Moorad, 2013). Additionally, demographic trends seem to have an effect on height and weight due to the effect of more resources allocated to fewer children (as documented for height by Lawson & Mace 2008), or the effect of changing reproductive success conferred by height and weight, as shown by evidence of transition from favouring high BMI to favouring height in populations ongoing early stages of the demographic transition (Courtiol et al., 2013; Sear, 2006).

In Latin America, the decline in mortality started around 1930, while the decline in fertility appears as recently as the 1970s and has been very gradual (Reher, 2004). The relatively long period between these declines implied accelerated population growth and the maintenance of high levels of poverty in the region (Reher, 2012). In Chile, fertility decline started in the 1960s in response to public health policies, and today the country has figures for fertility, mortality, and population growth corresponding to late phases of the demographic transition (Meza, 2003). As in other developing countries, the demographic transition in Chile was fast, and is therefore likely to be reflected in our data; 21% of the subjects were born before 1940, and thus are likely to have lived childhoods with higher mortality rates and reproductive periods in times of high fertility. Additionally, it seems safe to assume that a delayed onset of the declines occurred in the poorest rural areas, increasing the chances of having a dataset with a combination of life histories from subjects who were born and reproduced under different mortality–fertility trends.

5.3.2 Summary statistics of child mortality and fertility in association with lactase persistence

Figure 5.2 shows an histogram for the full range of the number of children ever born to each participant and their frequencies. On average, participants have 2.92 children, with a survival rate of 93%. Some other general statistics on fertility and mortality are presented in section 3.5. Transition to low mortality makes deaths of children rare events, and thus these have occurred almost exclusively in the older age groups. This makes it very difficult to analyse the number of child deaths in a meaningful way as they were only experienced by 14% of the participants.

Nevertheless, child mortality rate was calculated based on the total number of children ever born

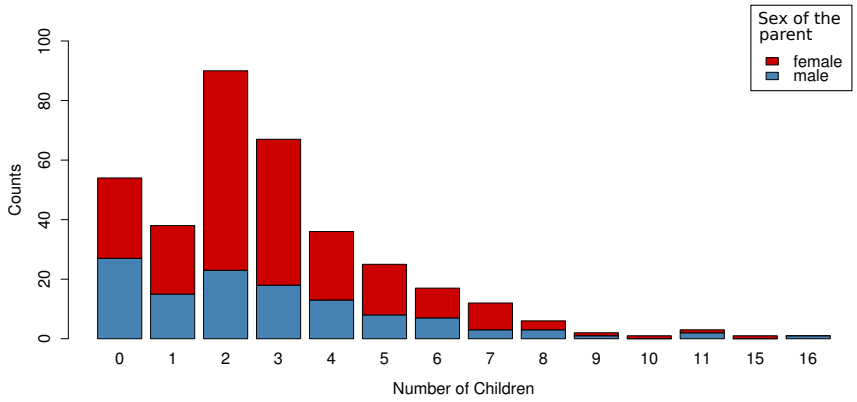

Figure 5.2. Histogram showing the frequency of number of children ever born by sex of the reporting parent.

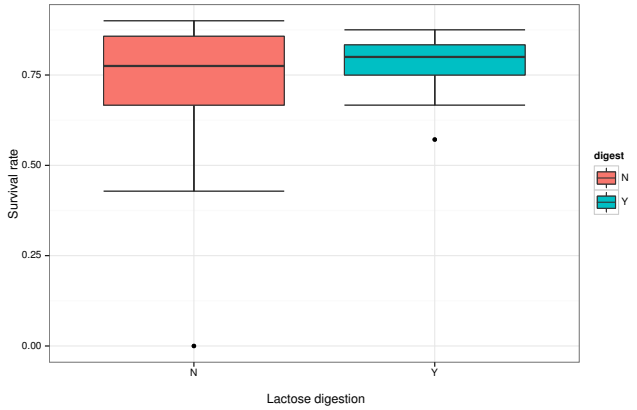

Figure 5.3. Rate of surviving children in predicted lactose–digesters and non–digesters. Predicted digesters have an average children surviving rate of 0.77, while non–digesters have an average of 0.7. Differences are not statistically significant (n = 67, t–test: $d.f$ = 37.7, $p–value$ = 0.12)

(*CEB*) and the number of surviving children (*SV*) (i.e. $S = CEB/SC$), and values of $S = 1$ were excluded of the analysis to be able to detect differences, reducing our sample size to $n = 67$. Figure 5.3 shows differences in rate of surviving children between predicted lactose–digesters and non–digesters. Predicted digesters have an average rate of children surviving of 0.77, while non-digesters have an average of 0.7, but this difference is not statistically significant (t–test: $d.f$ = 37.7, $p–value$ = 0.12). Although age is an important confounding variant to take into account, the frequency of deaths is too small to allow further division of the dataset. Given this sample size, a minimum effect size of 0.55 (i.e a difference of 0.55 in survival rate) is needed to get significant differences using t–test with a power of 0.9. Therefore other possibilities will need to be examined based only on crude number of children ever born (CEB).

The same analysis was applied to total number of children ever born in Figure 5.4. Predicted digesters have 3.03 children on average and non–digesters have an average of 2.78. However, this difference is not statistically significant (t–test: $d.f$ = 263.367, $p–value$ = 0.36, to be detectable

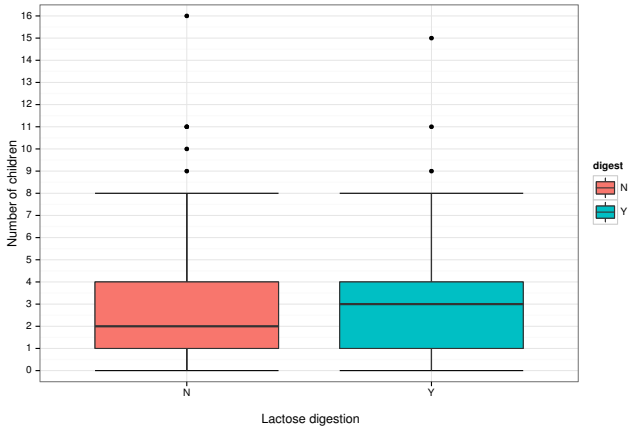

Figure 5.4. Number of children ever born to predicted lactose–digesters and non-digesters. Predicted digesters have 3.03 children on average and non-digesters have an average of 2.78. Differences are not statistically significant (t–test: $d.f$ = 263.367, $p-value$ = 0.36

by this sample size with a power of 0.9, difference in average number of children would have need greater than 0.4). Demographic changes mean that transition to low fertility and delayed reproduction makes older age group more fertile and the older age groups have already completed their reproductive span. These features demand careful consideration of age when comparing fertility between groups. Additionally, the trend towards low fertility distorts the influence of extreme outliers with very high fertility in any comparison. Figure 5.5 shows boxplots of number of children in predicted lactose digesters and non–digesters in different age groups and removing extreme outliers with more than 10 children (n = 5), and although some apparent increased fertility of lactose digesters starts to appear in the older groups, differences are not significant.

Assuming 45 years old to be the end of the reproductive span of most of the population offers

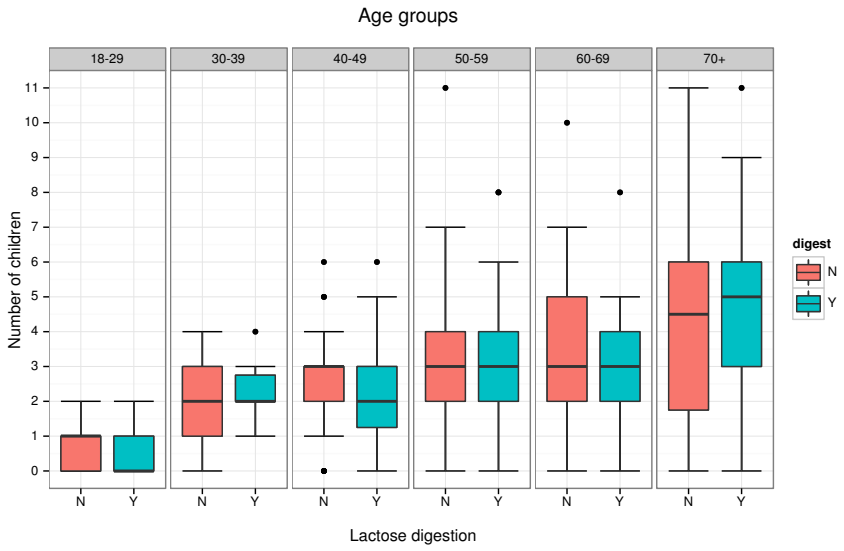

Figure 5.5. Number of children in predicted lactose digesters and non-digesters in different age groups and removing extreme outliers with more than 10 children (n = 5), Differences are not significant in any age group.

another way to examine differences in fertility only for those who have completed their reproductive span. Figure 5.6 shows differences in fertility analysed only for people over 45. Differences are small and not statistically significant (t−test: $\overline{x}$ digesters = 3.84, $\overline{x}$ non−digesters = 3.57, $d.f$ = 59.57, $p-value$ = 0.46).

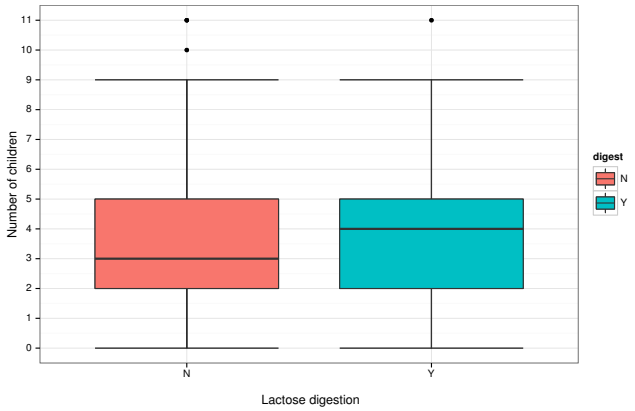

Figure 5.6. Differences in number of children between lactose digesters and non-digesters analysed only for people over 45 years old. Differences are small and not statistically significant (t−test: $\overline{x}digesters$ = 3.84, $\overline{x}non-digesters$ = 3.57, $d.f$ = 59.57, $p-value$ = 0.46)

A problem with analysing the data as age groups is the partitioning of a continuous variable into discrete categories. An analysis of covariance (ANCOVA) provides another way to summarise difference between predicted digesters and non−digesters controlling for age, but again, the difference is not statistically significant (see Figure 5.7). These results suggest that other analyses based on these two variables are unlikely to show significant differences between digesters and non−digesters, but since other variables could be hindering the differences, multiple regression models could yield different results , allowing new ways to examine this data. Moreover, since total number of children ever born is a discrete count (and therefore not a continuous variable with normal distribution), generalised linear models based on Poisson regression are more adequate and will be explored in the next section.

5.3.3 Modelling fertility as response to lactase persistence

Other studies using regression analysis to examine fertility measured as number of children often mistakenly use log transformation with count data of number of children, and has been identified as a common mistake by O'Hara & Kotze (2010) (see Bailey et al. 2014 for an example). To avoid this issue, zero−inflated regression models, a family of Poisson generalised linear models specialised in count data with a positive skew due to an excess of zeroes, was the preferred approach in this study. The zero−inflated approach is particularly relevant for models of mortality and fertility due to the considerable number of people that has not experienced child death and does not have had

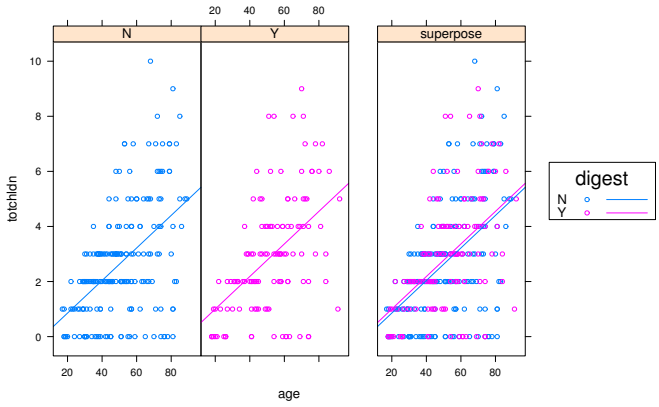

Figure 5.7. Analysis of covariance of total number of children and age by predicted lactose digestion status. Differences between lactose digesters and non–digesters are not statistically significant.

children. Therefore, the use of this type of regression is expected to decrease the confounding overall effect of age drastically. However, this dataset seems insufficient to attempt the model for total deaths of children as response leading to nonconvergence, so the models for children ever born (CEB) are discussed herein.

Models were developed following a similar procedure as for BMI and height in section 5.2, starting from a full model based on relevant variables in the dataset and narrowing down to a model with better Akaike Information Criterion using backward stepwise AIC in the R packages "MASS" (Venables & Ripley, 2002). Variables considered relevant for these models were lactose digestion status predicted by *−13,910* genotype ("digester."), sex, age, daily consumption of milk ("cups"), wealth, and, based on the results of section 5.2, BMI.

Full model		**Log-likelihood**	**D.F.**	**AIC**
ceb	∼ **digester** + **sex** + **age** + **wealth** + **cups** + **bmi**	-640.6	14	1309.1
		B	**B S.E.**	**z − value**	**p − value**
	digest (yes)	2×10^{-4}	0.069	0.004	0.999
	sex (female)	0.011	0.075	0.15	0.881
	age	0.023	0.002	10.68	$\mathbf{2\times10^{-16}}$
	wealth	-0.022	0.031	-0.708	0.479
	cups	-0.002	0.091	-0.017	0.986
	bmi	0.023	0.008	2.893	**0.004**
Best model		**Log-likelihood**	**D.F.**	**AIC**
ceb	∼ **digester** + **sex** + **age** + **wealth** + **cups** + **bmi**	-645.2	8	1306.3
		B	**B S.E.**	**z − value**	**p − value**
	digest (yes)	0.007	0.069	0.1	0.833
	age	8×10^{-4}	0.002	11.788	$\mathbf{2\times10^{-16}}$
	bmi	0.023	0.008	2.999	**0.003**

Table 5.4. Poisson regression models tested for total number of children: Comparison of full and best model (according to backward stepwise AIC) tested for total number of children (ceb). Differences between both models are very small in terms of AIC, significant effects, and *p–values*. There is no evidence of an effect of predicted lactose digestion status on total number of children ever born.

Table 5.4 shows the full model compared with the best model deducted by backwards stepwise AIC. Difference between both models are very small in terms of AIC, significant effects, and *p–values*. There is no evidence of an effect of predicted lactose digestion status in number of children, and therefore no direct evidence of improved fitness in lactase–persistent individuals from this dataset. But, interestingly, the effect of BMI in total number of children is significant, and although increased fitness through BMI is not mediated by lactase persistence in this population, there is support to the hypothesis of an effect of lactase persistence on BMI in this and other studies (see section 5.2). From these data is not possible to conclude that increased BMI to lactase persistence could have had an effect in fertility in the past, but it is an attractive hypothesis. This and other ideas are commented in the following chapter.

Chapter 6

Discussion and conclusion

At the outset of this project my main aim was to understand how these pastoralist populations of South American coped with a dietary adaptation that had occurred first no more than 500 years ago. Examination of historical data indicated that the population had resulted from admixture of native Americans and Europeans, in a similar way as occurred elsewhere in Latin America. The questions then were: Had a significant proportion of the people acquired lactase persistence alleles from their European ancestors; had this allele increased in frequency due to the affects of either genetic drift, resulting from small effective population size, or natural selection, resulting from the advantage conferred from being able to digest milk? Could differences be detected in the different villages? Was there any indication of benefit conferred by lactase persistence? Was there evidence of population structure that might confound the results? In many respects the results differed from those anticipated. This discussion considers each of the mentioned aspect.

6.1 Findings and implications

6.1.1 Lactase persistence in a South American pastoralist population

South America, a continent where pastoralism is ancient but milking and lactase persistence are likely to be recent introductions, offered an interesting scenario for the study of the adaptation to milk drinking and evolution of lactase persistence because of the known association between pastoralism and the selection of this trait in the Old World. This is the first study to examine lactase persistence in a pastoralist population of South America, combining genetics, lactose tolerance tests, and evolutionary demography.

As reported previously in Europe, we found a strong association between the presence of the *−13,910*T* allele and lactose digestion, predicting lactose digestion status in 99.64% of the subjects (see section 3.1). Further inspection of the enhancer sequences did not reveal any of the other previously reported alleles associated with lactase persistence nor any additional unreported variant. This result supports the hypothesis that *−13,910*T* is the major allele causative of lactase persistence in this population, in agreement with results found in non−pastoralist South American populations. (Bulhões et al., 2007; Morales et al., 2011).

In this population, *−13,910*T* has a frequency of 0.22, and 38% of the subjects are predicted lactase−persistent (Table 3.2), which is also in agreement with previous studies based on *−13,910*T* or lactose tolerance testing in other Latin American populations (see Tables 1.2 and 1.4). This seems consistent with expected values due to historical admixture in these communities in the ~500 years since lactase persistence was introduced and does not suggest an elevated prevalence of lactase persistence associated with their pastoralist way of life.

6.1.2 Surnames applied to the study of population structure

As an initial approach to studying population structure, we first used surnames (section 4.1.1). Although not a novel approach and with limitations, this yielded interesting results at no financial cost. With some exceptions, estimations of inbreeding based on surnames resulted in surprisingly similar estimations to those based on forensic STR markers (Table 4.3). Additionally, the use of surnames allowed us to identify useful clues of population differentiation for planning sampling strategies (Appendix F) and suggested further analyses to be done based on genetic data (Chapter 4).

The use of surnames to estimate inbreeding has obvious limitations. It is strongly based on an assumption of monophyletic origin of surnames (Rogers, 1991), and is likely to be less useful in

large populations where common surnames have overwhelming frequencies and identity by state is unlikely to represent identity by descent. However, lists of surnames are much easier to obtain than DNA data. Despite the advent of cheaper genetic testing this still can be an unjustifiable expenditure for exploratory studies. Likewise, public databases such as HapMap and 1000 Genomes, are focused on a few highly differentiated populations, and thus surnames can still be useful to study groups unrepresented in the literature, small populations, and small–scale differentiation. Surnames could also embed historical information not embedded in the genes.

6.1.3 Forensic identification STR markers applied to the study of population structure

Highly variable DNA markers, such as STRs, have proved useful in human identification and have become part of the routine of DNA techniques used in forensic sciences for paternity testing and criminal investigation. The wide use of these applications led to the development of an array of very accessible forensic identification kits, with well documented population profiles and high quality control standards. However, these kits are not as commonly used outside the scope of forensic sciences, even though they can have useful applications in the study of population structure, as documented in this thesis and other studies (Gaikwad et al., 2006; Ingram et al., 2009b; Kashyap et al., 2006; Rajkumar & Kashyap, 2004; Rubi-Castellanos et al., 2009b; Sahoo & Kashyap, 2005).

In this study, STR markers identified two parental clusters according to the method proposed by Evanno et al. (2005) to identify the number of clusters from the data (see Figure 4.8). Further analyses showed that these clusters are not related to European/Amerindian ancestry as judged by using the ancestry informative markers (Figure 4.13) or lactase persistence status (Figure 4.19). The structure detected by the STRs might possibly result from clustering of individuals sharing several unlinked alleles inherited from a recent common admixed ancestor a few generations ago (i.e. some generations after the first admixture event).

These markers were also used to estimate inbreeding and to control for cryptic relatedness using a matrix of proportion of shared alleles (section 4.2.1). Estimations of inbreeding show an average of 0.149 (range from 0.071 to 0.611), and was similar to estimations from a dataset of 159 individuals from East African ancestry reported as unrelated. The use of the PSA matrix generally improved the significance of effects detected by models without it, but the amount of added explained variance was marginal and the variables identified as significant were the same as in the models without the PSA matrix. The evaluation of how useful the PSA matrix is for the purposes of control for cryptic relatedness poses some limitations, as discussed in section 6.2.

6.1.4 Ancestry and the origin of the Agricultural Communities

Analysis of the Ancestry Informative Markers shows similar proportions of European and Native American ancestry in this population, with almost no contribution of African ancestry. The estimated proportion of European ancestry ranges from 0.419 in the northernmost study site, to 0.5 in the southernmost, confirming a north–south cline of ancestry proportions previously documented in this region (Acuña et al., 2000), but not significantly associated with a cline in proportions of lactase–persistence (see Figure 3.1). Although European ancestry estimated with Admixture 1.23 (Alexander et al., 2009) is, on average, higher in predicted lactase–persistent individuals (*LP* = 0.5, *NLP* = 0.47), the difference is small, and lactase–persistent individuals are not differentiated from non–persistent individuals using Principal Components Analysis and the STRUCTURE 2.3.4 algorithm (Falush et al., 2003, 2007; Hubisz et al., 2009; Pritchard et al., 2000).

These results show the proportion of European ancestry to be a bad predictor of lactase persistence. This is clearly not a reflect of a non–European origin of LP alleles since all were *−13,910*T* on the same the haplotypic background as that of Europeans (discussed in the next section) and there is only a very small chance of the independent recurrence of the same mutation. This illustrates a distribution of the European segments carrying this allele that is independent of global ancestry in the genome due to segregation and recombination in the 500 years following the contact and admixture process.

As introduced in section 1.3, the ethnic composition of the labour force in the Coquimbo Region colonial administration is an ongoing debate between those who allocate more or less importance to Spanish, Native, and African contribution. Our findings suggest an ancestral population composed of Spanish and Native American ancestry, without much African contribution, although our findings do not mean that these proportions were necessarily the same in the past , since they do not take into account the possibility of lineages without descendants, lineages with higher reproductive success, nor effects of migration. They are also estimated by extrapolation from relatively few markers. These genetic inferences give clues about the genealogical origin of these groups, but are not necessarily related to the origin of cultural traits such as the system of land management and social organisation.

6.1.5 Haplotypic background of *LCT* and the European origin of *−13,910*T*–carrying haplotypes

Haplotypes of the *LCT* region were inferred in this population using a total of 28 other genetic markers over a distance of 1.77 Mb, with the aim of seeing whether the *−13,910*T* allele was on

the same extended haplotypic background as is found in Europe, as a way of testing the hypothesis of a European introduction of the variant in this population. Based on the idea of a small Spanish group that introduced the allele and with only 15 to 20 generations for recombination, we expected to find rather little diversity of haplotypes carrying *−13,910*T*, all of European origin, while there would be expected to be a larger diversity of haplotypes carrying *−13,910*C*, of both European and Native American origin. We also expected haplotypes carrying *−13,910*T* to be mostly a subset of the European haplotypes we used as controls. Surprisingly, a high diversity of haplotypes carrying *−13,910*T* was found.

All haplotypes carrying *−13,910*T* were on an A core haplotypic background and in a European–like large block of extended homozygosity spanning ~900 Kb, supporting the hypothesis of *−13,910*T* as a European introduction. However decay in linkage disequilibrium upstream *−13,910* is slightly more pronounced in the Chilean than the European samples used for comparison (see Figure 4.21), which seems most likely attributable to recombination after introduction of the allele to Chile. But it is also possible that some of this diversity originated in the Old World but was not represented in the European samples used for comparison, which unfortunately did not include Iberian samples, since the samples were selected for the purposes of another study. Whether or not this is the case, the results point to a larger and more diverse founding European group and/or continuous gene flow of European haplotypes into the population, and also provides no evidence of very recent natural selection.

Another surprising finding was a high frequency (0.26) of a core haplotype that is most likely the one named C. According to Hollox et al. (2001), the C core haplotype is common in Bantu–speaking populations (0.31), but is rarer in Europe (with frequencies of 0.03 in Northern Europe and 0.12 in Italy, but without data from Spain), and is not very common in East Asia either (0.15 in Japan and 0.09 in China). Interestingly a very high frequency of this haplotype (0.41) was also reported in Native Brazilian populations by Friedrich et al. (2012a), the only study which reports *LCT* haplotypes in Native American populations. This suggests core haplotype C as common in Native Americans, but needs to be confirmed in future studies.

6.1.6 Lactase persistence and natural selection

One of the main aims of this study was to examine associations between lactase persistence, height, weight, and number of children, as a way of understanding the mechanisms of the natural selection of this trait.

Initially, we found a strong association between predicted lactase persistence, height, and weight in males (Table 3.7 and Figure 3.5). But this analysis was susceptible to the effects of popula-

tion structure. Campbell et al. (2005) found a similar association between lactase persistence and height in European–Americans, but the association was proved to be a spurious effect of stratification and dissipated after controlling for structure using ancestry informative markers. Similarly, Astle & Balding (2009) make reference to the same study as a warning on the effect of cryptic relatedness.

To prevent these issues, a considerable part of this thesis was dedicated to the study of population structure (Chapter 4), in an attempt to take into account inbreeding, relatedness, and ancestry. Those analyses (in Chapter 5) were used in multiple regression models to test these associations after controlling for them as potential confounding factors and only the effect of lactase persistence on BMI in males remained significant across all models. A similar procedure was followed to study the effect of lactase persistence on reproductive success, measured as total number of children ever born, and although the effects of lactase persistence on number of children were not significant, BMI resulted significantly correlated with number of children.

These findings should be taken with some caution because of the rather limited amount of variance explained by our model (details in section 6.2). However, the fact that the effect of lactase persistence on BMI is very significant ($p-value$ = 0.002), even taking into account relatedness and ancestry and the fact that similar effects of lactase digestion have been reported previously (Almon et al., 2010; Corella et al., 2011; Lamri et al., 2013; Smith et al., 2008) suggests that this is a real association not produced by confounding variables. In the same way, the effect of BMI in total number of children ($p-value$ = 0.003) is also supported by other studies (Courtiol et al., 2013; Sear, 2006).

Some possible explanations are suggested by these results. Lactase persistence could have been a way to gain weight in the context of threats of famine or low food availability when lactase persistence was selected, playing the same role as variants responsible for increment in weight in other geographic regions, such as those proposed by the hypothesis of the thrifty genotype (Gibson, 2007; Neel, 1962) in the Americas, putting people at risk of metabolic syndrome (Luca et al., 2010; Weiss et al., 1984). Although attractive, this "hypothesis by analogy" exceeds the scope of this study and cannot be tested without similar data from other geographical contexts to compare additive effects of alleles in different loci converging in traits related with weight, all hypothetically originated as response to scarcity of resources in different regions and now adding their effects in admixed populations.

Another possibility could be the presence of genes involved in weight or height in close proximity to the lactase enhancer region, a plausible option considering the long region of LD around *LCT* and the complexity and the large number of loci involved in height and weight, even though none of the variants reported by Bell et al. (2005) and Lango Allen et al. (2010) are close to the *LCT* enhancer region.

6.2 Limitations

6.2.1 Restrictions in genetic analyses

The genetic analyses used in this thesis were limited by time and financial constraints. The set of Ancestry Informative Markers used for ancestry estimations (see Table 2.3) contained only 30 SNPs, and they were selected taking into account their extreme frequencies in parental populations. Other studies are based on panels greater by orders of magnitude (such as 446 SNPs in Galanter et al. 2012 or as many as 2,100 SNPs in Mao et al. 2007), a difference that could raise scepticism about our estimations of ancestry used as confounding variables in other analyses . The use of panels based on insertion–deletions (INDELs) has been claimed to offer better ancestry estimations with fewer markers (Friedrich et al., 2012b; Pereira et al., 2012), and it is possible that combining methods would improve estimations.

Similarly, the use of non–recombinant human DNA (Y–chromosome and mitochondrial DNA) would improve our estimations of ancestry, relatedness and inbreeding. It would allow estimations of paternal and maternal lineages and complement the data obtained from autosomal STRs. The comparison between Y–chromosome and the analysis of surnames would also have been of interest.

The use of DNA microarrays of thousand of genomewide SNPs could have been used to analyse both ancestry and relatedness together, but was not possible considering the original budget for this project. However, microarrays would be the best choice to expand this research in the future.

The set of SNPs genotyped used in *LCT* haplotypes inference was also limited. Complete characterisation of *LCT* haplotypes diversity was not an aim of this analysis, and the set of SNPs used to infer haplotypes (listed in Table 2.4) does not include defining SNPs for core haplotypes other than A, and can only distinguish unambiguously between the commonest core haplotypes (A, B, and C). This limits the scope of the implications of the high frequency of C haplotypes found, which could also be a combination of various minor haplotypes identical to C in the context of the SNPs included in the analysis. The study of the extended A haplotypes was also constrained by the European samples used for comparison, which did not include Iberian samples. Future work will consider the use of the data available from the 1000 Genomes project (The 1000 Genomes Project Consortium, 2012) in phasing and haplotypes analysis.

6.2.2 Survey, sampling and power issues

Recruitment of participants during the pilot and fieldwork stages proved to be extremely difficult due to initial lack of trust in our objectives and the size and sparsity of the settlements. Further-

more, these communities have high levels of seasonal migration and years of temporary migration for work in early adulthood. By the end of the pilot study it was clear that a strict control of sampling to select only non–relatives with four grandparents born in the region was not feasible. Thus control for birth–place of grandparents and reported relatedness was sacrificed to obtain the largest possible sample size, resulting in 451 samples collected in nine locations during the 10 months of fieldwork.

Birth–place of parents and reported relatedness were collected during the fieldwork, but the large amount of missing data would have severely reduced the sample size if incomplete cases had been deleted. With the exception of a few descriptive and exploratory analyses mentioned in the thesis, the use of multiple imputation techniques was avoided, but could be used in future research to continue the exploration of this dataset.

As a way of assessing the extent of relatedness, estimations of inbreeding using STR markers were compared with data obtained with the same kit (details in sections 2.2.7 and 4.2.1) from a sample of East African individuals reported as unrelated, finding similar inbreeding estimations. However, a similar procedure with a more closely related population (Latin American or European) would have been preferable had the data been available.

Despite the efforts to achieve a large sample size, our regression models have very low $R^2_{adj.}$ values, meaning that even the best models explained very little of the variation. We expected an important part of the variation to be explained in our reduced models, after trying with all the variables that seemed pertinent in our dataset. Some possible reasons for this are the reduction of the variance in height and BMI after dividing the dataset by sex, the omission of variables with an important effect in the model, or problems with the sample size. As mentioned in Chapter 5, a single model controlling for sex (instead of evaluating each sex separately) improved R^2 values because of the larger amount of total variance in height and BMI in the dataset explained by sex. Another possibility is that variables not included in the models have large effects on what remains of the variance in height or BMI, for instance, the estimation of ancestry for particular chromosomal segments (local ancestry) may house key ancestry–specific genes affecting height or BMI in the groups under comparison, a possibility that is harder to adjust for, and would have required genomewide markers. It could also be that the effect of lactase persistence in the models is too small to be detected with this sample size and power issues could been impeding the achievement of higher $R^2_{adj.}$ values.

Sample size, missing data, and time constraints restricted the analyses of a large amount of data collected in the questionnaires (Appendix C), and data on weaning, birth intervals, and number of

reproductive partners were not used in the end. Discarding the collection of these data to reduce the time required for each questionnaire could have had an effect in improving sample size.

6.2.3 Coquimbo pastoralists in the context of Latin America

It should be emphasised that the results presented in this thesis may not be representative of other Latin American populations or even other Chilean populations. Across Latin America, there is considerable variability in the different admixture processes and in the different proportions of European and Native American contributions involved, and also the input from different populations. Since European contact, Latin America has been historically shaped by ethnic diversity.

The proportion of African ancestry in this population was minimal but this is clearly not the case for many Latin American countries, particularly in northern South America with an estimated population of African–descendants of 150 million (Ribando, 2007). For instance, in Brazil, the country with the largest African–descendant population outside Africa, a high diversity of *LCT* haplotypes and the presence of other variants associated with lactase persistence has been reported (Friedrich et al., 2012b)

Furthermore, the amount of Spanish/Portuguese contribution, as well as the contribution of other European populations varies greatly between countries. European ancestry is usually described as very high in Argentina and south of Brazil, and low in most rural areas of Bolivia (Lizcano, 2005). There are historical records of significant migration of Italian and German populations to Brazil, Argentina and Chile after the World War II, while other migratory waves were promoted afterwards in association with conflicts in Europe (such as the splits of Yugoslavia and the U.S.S.R) (Lizcano, 2005). However, demographic and historical accounts of ancestry are mainly based on reported ancestry and heavily influenced by social factors. Additionally, there are regional differences in the Amerindian parental populations, which are underrepresented in genetic studies and public databases.

6.2.4 The conditions under which lactase persistence underwent selection

The Agricultural Communities form Chilean *"Norte Chico"* were selected for this study because of their pastoralist subsistence and the likely introduction of variants associated with lactase persistence. The genetic studies presented here disappointingly show little evidence of very recent or ongoing selection. Many of the conditions that may have been present in those places where lactase persistence underwent strong natural selection in the Neolithic, may not have been present in these communities. Child mortality is uncommon today, even in marginal populations such as this one. Similarly, reduction in fertility over the time period since the way of life was established

is likely to have reduced the differences in reproductive success to undetectable levels, which in turn are very difficult to detect by direct methods based on number of children ever born (Harrison & Morphy, 1998).

Our results did however suggest an effect of lactase persistence on BMI, which can be interpreted as a contribution to weight that might have been advantageous in periods of famine and low food availability. Today, high BMI is also achievable by other means and thus the contribution of lactase persistence may be less important than in the past. However, the assumption of low access to food in the past can be questionable, and has been criticised before (Polanyi, 1944; Sahlins, 1972).

6.3 Future research

The approach, limitations, methods and findings of this study could be revisited in the future to address the same research questions with other data, or to test the emerging hypotheses of the effect of BMI on other results. The approach to studying the evolution of lactase persistence in other populations can be repeated in other pastoralist groups with lactase–persistent and non–persistent members but less extreme admixed ancestry in other areas such as central Africa and central Asia.

Pastoralist Native Americans, like most Native American populations, are likely to have a small proportion of European ancestry. Therefore, they could offer new insights, specially employing a genomewide approach and methods of local ancestry estimations, to evaluate whether the *LCT* enhancer region is relatively more European than the rest of the genome. This line of research would also improve our knowledge about *LCT* haplotype diversity in Native Americans.

Research on other Latin American mixed populations and Native American populations is needed to develop a better characterisation of their population genetics. This is a difficult task due to the high levels of admixture and presumed low diversity in the pre–Columbian Amerindian populations (Hey, 2005; Wang et al., 2007). Methods to detect deep levels of weak population differentiation are needed, and the use of forensic STR markers could serve this purpose.

Future research could aim to improve our data and re–evaluate our results. Y chromosome, mitochondrial DNA, and specially genomewide microarrays would greatly contribute to further controls of ancestry and relatedness, and a well–designed strategy of multiple imputation could improve the significance of our models.

A great amount of data, especially the data collected about children, was not used, and could be used in the near future to explore effects of lactase persistence in birth intervals, number of reproductive partners, as well as effect of lactase persistence status of parents in the age of weaning and the number of children of their children. The same dataset could be also useful for other research topics. For instance, to explore the distribution of other variants of interest used as ancestry informative markers in this study, the reproductive effects of ancestry and wealth, and the relationship between ancestry and wealth.

With respect to methods, the availability and easy implementation of analyses based on surnames demands comparative studies to assess the agreement of their results with those based on genetic markers. If analyses based on surnames yield similar estimations to those obtained using genetic markers, surnames could be used in exploratory research, particularly for populations not repres-

ented in HapMap, 1000 genomes, or other public databases. The data collected in this thesis could be also used to implement methods of detection of isolation by distance and geographic barriers to gene flow, using Monmonier's algorithm, as has been done in other studies (Manni et al., 2004; Rodriguez Diaz et al., 2010) .

In relation to our findings, these results point towards the empirical exploration of hypotheses to explain the sex differences in the effects of lactase persistence. One possibility could be gender–bias in child nutrition which could allow male lactase–persistent individuals to develop higher BMI, diminishing the effect of other factors such as ancestry, which became relatively important for BMI in females, who did not have the nutritional input to develop a higher BMI. This hypothesis could be tested in a longitudinal study, using food journals and close monitoring of food consumption in combination with estimation of ancestry proportions and anthropometric data.

6.4 Concluding remarks

To conclude, the research questions introduced at the end of Chapter 1 are reviewed and commented on below as a summary of the key points of this thesis.

- **How frequent is lactase persistence in these communities and which of the lactase–persistence–associated alleles (nucleotide changes reported in the lactase enhancer region), are present?**

  The European *−13,910*T* was the only variant associated with lactase persistence found in sequences of the *LCT* enhancer region. The presence of this allele was strongly associated with lactose digestion determined by lactose tolerance tests in in 99.64% of the subjects. Based on the frequency of this allele, and assuming a dominance model, 38% of the subjects are predicted lactase–persistent. These results supports *−13,910*T* as causative of lactase persistence in this population.

- **Are lactase persistence frequencies affected by geography, ancestry, inbreeding, or other confounding causes?**

  Estimation of ancestry proportions using Ancestry Informative Markers and estimations of inbreeding using forensic identification markers showed no association with predicted lactase persistence, and geographic differences in allele frequencies were not significant. Nevertheless, results of these analyses were used as controls for confounding variables in further statistic tests.

- **How much milk and milk products are consumed by people in these communities and are any differences in milk consumption associated with lactase persistence? Are there symptoms from milk consumption that are correlated with persistence status?**

  An estimated average of milk consumption of 0.57 cups per day per capita was found, with only 16.9% of the subjects reported to never drink milk. Differences in average milk consumption according to predicted lactase persistence or genotype are not statistically significant. In relation to symptoms, feeling unwell with milk is not statistically associated with any genotype nor with predicted lactase persistence status. Non–persistent individuals do not avoid milk, suggesting that maybe milk avoidance is not the cause of the presumed increased relative fitness in *−13,910*T* carriers, but advantages could be result of different nutritional effects of milk consumption in lactase–persistent and non–persistent individuals.

- **Are there any differences in height, weight, and Body Mass Index associated with lactase persistence?**

Controlling for relevant confounding variables, such as age, ancestry, milk consumption, wealth and relatedness, only BMI is associated with lactase persistence, and only in males. Possible evolutionary benefits of a higher BMI can be related with episodes of famine or scarcity of resources. Some hypotheses to explain sex differences in the effects of lactase persistence could be based on higher environmental sensitivity in males, feeding sex–bias towards boys over girls, or male–male competition. This study does not provide data to test any of these hypotheses.

- **Are there any differences between number of children's births and children's deaths associated with the genotype of the parents?**

  Controlling for relevant confounding variables, such as sex, age, wealth and BMI, we found no effect of lactase persistence in total number of children, but a significant effect of BMI was found. Children's deaths were too rare to build a meaningful model. Global demographic trends towards low mortality and low fertility could make differences in reproductive success undetectable in the course of only one generation. However, a significant effect of BMI in number of children's births was found.

Based on previous knowledge of the life style and ancestry of the pastoralist populations discussed through the thesis, this research attempted to study aspects of the selection of lactase persistence, despite the inherent difficulties of such study. Contrary to what was expected, frequencies of lactase persistence were similar to those of non–pastoralist admixed populations of South America, milk consumption was not related with lactase persistence genotypes, and considerable diversity was found in *LCT* haplotypes. These results changed the focus of interest of this study, originally planned to be centred around selection, drift, and recent natural selection.

Despite these limitations and those acknowledged in section 6.2, these results suggest a relationship of lactase persistence with weight and the possible effects of this relationship in reproductive success, as a hypothetical mechanism that may have resulted in strong selection favouring lactase persistence in the past. This line of analysis seems sensible in the light of other studies suggesting links between BMI and lactase persistence (Almon et al., 2010; Corella et al., 2011; Lamri et al., 2013; Smith et al., 2008), and BMI and reproductive success (Courtiol et al., 2013; Sear, 2006). Similar studies in populations at different stages of the demographic transition could swap the roles of fertility and BMI towards mortality or height, and further studies of the contribution of lactase persistence and non-persistence to BMI and height, and its effects in mortality and fertility, could offer novel ways to understand at least one of the possible routes to the evolution of lactase persistence.

Bibliography

Acuña, M., Llop R, E., & Rothhammer E, F. (2000). Composición genética de la población chilena: las comunidades rurales de los valles de Elqui, Limarí y Choapa. *Revista Médica de Chile*, *128*(6), 593–600.

Adams, S. M., Bosch, E., Balaresque, P. L., Ballereau, S. J., Lee, A. C., Arroyo, E., López-Parra, A. M., Aler, M., Grifo, M. S. G., Brion, M., Carracedo, A., Lavinha, J., Martínez-Jarreta, B., Quintana-Murci, L., Picornell, A., Ramon, M., Skorecki, K., Behar, D. M., Calafell, F., & Jobling, M. (2008). The genetic legacy of religious diversity and intolerance: paternal lineages of Christians, Jews, and Muslims in the Iberian Peninsula. *American Journal of Human Genetics*, *83*(6), 725–736.

Adler, D., & Murdoch, D. (2014). *rgl: 3D visualization device system (OpenGL). [Software]*.
URL `http://cran.r-project.org/package=rgl`

Agueda, L., Urreizti, R., Bustamante, M., Jurado, S., Garcia-Giralt, N., Díez-Pérez, A., Nogués, X., Mellibovsky, L., Grinberg, D., & Balcells, S. (2010). Analysis of three functional polymorphisms in relation to osteoporosis phenotypes: replication in a Spanish cohort. *Calcified Tissue International*, *87*(1), 14–24.

Albers, C. (2014). Cartografía Rulamahue – Universidad de La Frontera, Temuco, Chile. Retrieved: 16 February 2014.
URL `http://www.rulamahue.cl/`

Albertsson-Wikland, K., Rosberg, S., Karlberg, J., & Groth, T. (1994). Analysis of 24-hour growth hormone profiles in healthy boys and girls of normal stature: relation to puberty. *The Journal of Clinical Endocrinology and Metabolism*, *78*(5), 1195–1201.

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–64.

Alexander, W. (2004). Clandestine Artisans or Integrated Producers? Standardization of Rural Livelihood in the Norte Chico, Chile. *Culture & Agriculture*, *26*(1), 38–51.

Alexander, W. L. (2008). *Resiliency in hostile environments: a comunidad agrícola in Chile's Norte Chico*. Cranbury: Lehigh University Press.

Alkorta-Aranburu, G., Beall, C. M., Witonsky, D. B., Gebremedhin, A., Pritchard, J. K., & Di Rienzo, A. (2012). The genetic architecture of adaptations to high altitude in Ethiopia. *PLoS genetics*, *8*(12), e1003110.

Allal, N., Sear, R., Prentice, M., & Mace, R. (2004). An evolutionary model of stature, age at first birth and reproductive success in Gambian women. *Proceedings of the Royal Society B: Biological Sciences*, *271*(1538), 465–470.

Allison, A. (1954). Protection afforded by sickle-cell trait against subtertian malarial infection. *British Medical Journalournal*, *1*(4857), 290–294.

Almon, R., Alvarez-Leon, E. E., Engfeldt, P., Serra-Majem, L., Magnuson, A., & Nilsson, T. K. (2010). Associations between lactase persistence and the metabolic syndrome in a cross-sectional study in the Canary Islands. *European Journal of Nutrition*, *49*(3), 141–146.

Alzate, H., Gonzalez, H., Guzman, J., & Guzmán, J. (1969). Lactose intolerance in South American indians. *American Journal of Clinical Nutrition*, *22*(2), 122–123.

Ampuero, G. (1978). *Cultura Diaguita*. Santiago: Departamento de extensión cultural, Ministerio de Educación.

Ampuero, G., Hidalgo, J., Schiappacasse, V., Niemeyer, H., Aldunate, A., & Solimano, I. (1989). La Cultura Diaguita chilena. In *Prehistoria*, (pp. 277–287). Santiago: Andres Bello.

Anagnostou, P., Battaggia, C., Coia, V., Capelli, C., Fabbri, C., Pettener, D., Destro-Bisol, G., & Luiselli, D. (2009). Tracing the distribution and evolution of lactase persistence in Southern Europe through the study of the T(-13910) variant. *American Journal of Human Biology*, *21*(2), 217–219.

Anderson, B., & Vullo, C. (1994). Did malaria select for primary adult lactase deficiency? *Gut*, *35*(10), 1487–1489.

Ángel, L., Calvo, E., & Muñoz, Y. (2005). Prevalencia de hipolactasia tipo adulto e intolerancia a la lactosa en adultos jóvenes. *Revista Colombiana de Gastroenterología*, *20*(4), 35–47.

Aoki, K. (1986). A stochastic model of gene-culture coevolution suggested by the culture historical hypothesis for the evolution of adult lactose absorption in humans. *Proceedings of the National Academy of Sciences of the United States of America*, *83*(9), 2929–2933.

Aoki, K. (1987). Adult lactose absorption and milk use from the standpoint of gene-culture theory. *The Japanese Journal of Genetics*, *62*(5), 445–459.

Aoki, K. (2001). Theoretical and empirical aspects of gene-culture coevolution. *Theoretical Population Biology*, *59*(4), 253–261.

Ashton, K., Tracy, M., & Queiroz, A. D. (2000). Is Bergmann's rule valid for mammals? *The American Naturalist*, *156*(4), 390–415.

Asselin, L., & Anh, V. (2008). Multidimensional poverty and multiple correspondence analysis. Tech. rep., CIRPÉE: Centre Interuniversitaire sur le Risque, les Politiques Économiques et l'Emploi., Quebec.

Astle, W., & Balding, D. J. (2009). Population Structure and Cryptic Relatedness in Genetic Association Studies. *Statistical Science*, *24*(4), 451–471.

Auricchio, S., Rubino, A., Landolt, M., Semenza, G., Prader, A., & Semeza, G. (1963). Isolated Intestinal Lactase Deficiency in the Adult. *Lancet*, *2*(7303), 324–326.

Avendaño, S., & Gallardo, H. (1986). *Las Comunidades Agricolas de la 4 Region. Una particular relacion hombre tierra*. La Serena: Sociedad Editora del Norte.

Bachmanov, A., & Beauchamp, G. (2006). Taste receptor genes. *Annual Review of Nutrition*, *27*(170), 389–414.

Baddeley, A., & Turner, R. (2005). Spatstat: an R package for analyzing spatial point patterns. *Journal of Statistical Software*, *12*(6), 1–42.

Bahamondes, M. (2003). Poverty-Environment Patterns in a Growing Economy: Farming Communities in Arid Central Chile, 1991–99. *World Development*, *31*(11), 1947–1957.

Baied, C., & Wheeler, J. (1993). Evolution of high Andean puna ecosystems: environment, climate, and culture change over the last 12,000 years in the Central Andes. *Mountain Research and Development*, *13*(2), 145–156.

Bailey, D. H., Hill, K. R., & Walker, R. S. (2014). Fitness consequences of spousal relatedness in 46 small-scale societies. *Biology Letters*, *10*(5), 2–5.

Barrai, I., Rodriguez-Larralde, A., Dipierri, J., Alfaro, E., Acevedo, N., Mamolini, E., Sandri, M., Carrieri, A., & Scapoli, C. (2012). Surnames in Chile: a study of the population of Chile through isonymy. *American Journal of Physical Anthropology*, *147*(3), 380–388.

Barrai, I., Scapoli, C., & Beretta, M. (1996). Isonymy and the genetic structure of Switzerland I. The distributions of surnames. *Annals of Human Biology*, *23*(6), 431–455.

BBC News (2013). Chile may annul 'flawed' 2012 census. Retrieved: 27 December 2013. URL `http://www.bbc.co.uk/news/world-latin-america-23611210`

Beall, C. (2001). Adaptations to altitude: A current assessment. *Annual review of anthropology*, *30*(2001), 423–456.

Beall, C. M. (2006). Andean, Tibetan, and Ethiopian patterns of adaptation to high-altitude hypoxia. *Integrative and Comparative Biology*, *46*(1), 18–24.

Beall, C. M. (2007). Two routes to functional adaptation: Tibetan and Andean high-altitude natives. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(Suppl. 1), 8655–8660.

Beall, C. M., Cavalleri, G. L., Deng, L., Elston, R. C., Gao, Y., Knight, J., Li, C., Li, J. C., Liang, Y., McCormack, M., Montgomery, H. E., Pan, H., Robbins, P., Shianna, K. V., Tam, S. C., Tsering, N., Veeramah, K. R., Wang, W., Wangdui, P., Weale, M. E., Xu, Y., Xu, Z., Yang, L., Zaman, M. J., Zeng, C., Zhang, L., Zhang, X., Zhaxi, P., & Zheng, Y. T. (2010). Natural selection on EPAS1 (HIF2alpha) associated with low hemoglobin concentration in Tibetan highlanders. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(25), 11459–11464.

Beall, C. M., Decker, M. J., Brittenham, G. M., Kushner, I., Gebremedhin, A., & Strohl, K. P. (2002). An Ethiopian pattern of human adaptation to high-altitude hypoxia. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(26), 17215–17218.

Beall, C. M., Strohl, K. P., Blangero, J., Williams-Blangero, S., Almasy, L. a., Decker, M. J., Worthman, C. M., Goldstein, M. C., Vargas, E., Villena, M., Soria, R., Alarcon, a. M., & Gonzales, C. (1997). Ventilation and hypoxic ventilatory response of Tibetan and Aymara high altitude natives. *American Journal of Physical Anthropology*, *104*(4), 427–447.

Beaumont, M. a. (2005). Adaptation and speciation: what can F(st) tell us? *Trends in ecology & evolution*, *20*(8), 435–440.

Becker, R. A., Brownrigg, R., & Wilks, A. (2013). *mapdata: Extra Map Databases. [Software]*. URL `http://cran.r-project.org/package=mapdata`

Bell, C. G., Walley, A. J., & Froguel, P. (2005). The genetics of human obesity. *Nature Reviews Genetics*, *6*(3), 221–234.

Bergmann, C. (1848). *Über die Verhältnisse der Wärmeökonomie der Thiere zu ihrer Grösse*. Gottingen: Vandenhoeck und Ruprecht.

Bersaglieri, T., Sabeti, P. C., Patterson, N., Vanderploeg, T., Schaffner, S. F., Drake, J. A., Rhodes, M., Reich, D. E., & Hirschhorn, J. N. (2004). Genetic signatures of strong recent positive selection at the lactase gene. *American Journal of Human Genetics*, *74*(6), 1111–1120.

Bertranpetit, J., & Cavalli-Sforza, L. L. (1991). A genetic reconstruction of the history of the population of the Iberian Peninsula. *Annals of Human Genetics*, *55*(1), 51–67.

Bigham, A. W., Mao, X., Mei, R., Brutsaert, T., Wilson, M. J., Julian, C. G., Parra, E. J., Akey, J. M., Moore, L. G., & Shriver, M. D. (2009). Identifying positive selection candidate loci for high-altitude adaptation in Andean populations. *Human Genomics*, *4*(2), 79–90.

Birdsey, G. M., Lewin, J., Cunningham, A., Bruford, M. W., & Danpure, C. J. (2004). Differential enzyme targeting as an evolutionary adaptation to herbivory in carnivora. *Molecular Biology and Evolution*, *21*(4), 632–646.

Bivand, R., Keitt, T., & Rowlingson, B. (2014). *rgdal: Bindings for the Geospatial Data Abstraction Library. [Software]*.
URL `http://cran.r-project.org/package=rgdal`

Bivand, R., & Lewin-Koh, N. (2014). *maptools: Tools for reading and handling spatial objects. [Software]*.
URL `http://cran.r-project.org/package=maptools`

Bivand, R., & Rundel, C. (2014). *rgeos: Interface to Geometry Engine - Open Source (GEOS). [Software]*.
URL `http://cran.r-project.org/package=rgeos`

Bivand, R. S., Pebesma, E., & Gomez-Rubio, V. (2013). *Applied spatial data analysis with {R}*. New York: Springer, 2nd ed.

Blakeslee, A. (1932). Genetics of sensory thresholds: taste for phenyl thio carbamide. *Proceedings of the National Academy of Sciences of the United States of America*, *18*, 120–130.

Bleich, S., Cutler, D., Murray, C., & Adams, A. (2008). Why is the developed world obese? *Annual Review of Public Health*, *29*, 273–295.

Blench, R. (2001). *'You Can't Go Home Again': Pastoralism in the New Millennium*. May 2001. London: Overseas Development Institute.

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

Boattini, A., & Calboli, F. C. F. (2012). *Biodem: Biodemography functions. [Software]*.
URL `http://cran.r-project.org/package=Biodem`

Boll, W., Wagner, P., & Mantei, N. (1991). Structure of the chromosomal gene and cDNAs coding for lactase-phlorizin hydrolase in humans with adult-type hypolactasia or persistence of lactase. *American Journal of Human Genetics*, *48*(5), 889–902.

Booysen, F., Berg, S. V. D., & Burger, R. (2008). Using an asset index to assess trends in poverty in seven Sub-Saharan African countries. In *Multidimensional Poverty*, 03, (pp. 1–33). Brasilia: United Nations Development Programme UNDP.

Bose, D., & Welsh, J. (1973). Lactose malabsorption in Oklahoma Indians. *American Journal of Clinical Nutrition*, *26*(12), 1320–1322.

Boyd, R. (1988). *Culture and the Evolutionary Process*. Chicago: University of Chicago Press.

Bravo, M. L., Moreno, M., Builes, J. J., Salas, A., Lareu, M. V., & Carracedo, A. (2001). Autosomal STR genetic variation in negroid Chocó and Bogotá populations. *International Journal of Legal Medicine*, *115*(2), 102–104.

Briet, F., Pochart, P., Marteau, P., Flourie, B., Arrigoni, E., & Rambaud, J. C. (1997). Improved clinical tolerance to chronic lactose ingestion in subjects with lactose intolerance: a placebo effect? *Gut*, *41*(5), 632–635.

Brines, J. (2004). Adult lactose tolerance is not an advantageous evolutionary trait. *Pediatrics*, *114*(5), 1372.

Bronberg, R., Dipierri, J. E., Alfaro, E. L., Barrai, I., Rodríguez-Larralde, A., Castilla, E. E., Colonna, V., Rodríguez-Arroyo, G., & Bailliet, G. (2009). Isonymy structure of Buenos Aires city. *Human Biology*, *81*(4), 447–461.

Browman, D., Göbel, B., & Bollig, M. (1997). Pastoral risk perception and risk definition for Altiplano herders. *Nomadic Peoples*, *1*(1), 22–36.

Bulhões, a. C., Goldani, H. A. S., Oliveira, F. S., Matte, U. S., Mazzuca, R. B., & Silveira, T. R. (2007). Correlation between lactose absorption and the C/T-13910 and G/A-22018 mutations of the lactase-phlorizin hydrolase (LCT) gene in adult-type hypolactasia. *Brazilian Journal of Medical and Biological Research*, *40*(11), 1441–1446.

Burger, J., Kirchner, M., Bramanti, B., Haak, W., & Thomas, M. G. (2007). Absence of the lactase-persistence-associated allele in early Neolithic Europeans. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(10), 3736–3741.

Burgio, G. R., Flatz, G., Barbera, C., Patané, R., Boner, A., Cajozzo, C., & Flatz, S. D. (1984). Prevalence of primary adult lactose malabsorption and awareness of milk intolerance in Italy. *The American Journal of Clinical Nutrition*, *39*(1), 100–104.

Butler, J., Reeder, D., & National Institute of Standards and Technology. U.S. Department of Commetece (2014). Short Tandem Repeat DNA Internet Database – National Institute of Standards

and Technology. U.S. Department of Commetece.
URL `http://www.cstl.nist.gov/strbase/`

Calderón-Viacava, L., Cazorla-Talleri, A., & León-Barúa, R. (1971). [Incidence of lactose malabsorption in healthy young Peruvians]. *Acta Gastroenterologica Latinoamericana*, *3*(1), 11–16.

Caldwell, E. F., Mayor, L. R., Thomas, M. G., & Danpure, C. J. (2004). Diet and the frequency of the alanine:glyoxylate aminotransferase Pro11Leu polymorphism in different human populations. *Human Genetics*, *115*(6), 504–509.

Campbell, C. D., Ogburn, E. L., Lunetta, K. L., Lyon, H. N., Freedman, M. L., Groop, L. C., Altshuler, D., Ardlie, K. G., & Hirschhorn, J. N. (2005). Demonstrating stratification in a European American population. *Nature Genetics*, *37*(8), 868–872.

Campbell, M. C., Ranciaro, A., Froment, A., Hirbo, J., Omar, S., Bodo, J. M., Nyambo, T., Lema, G., Zinshteyn, D., Drayna, D., Breslin, P. S., & Tishkoff, S. (2012). Evolution of functionally diverse alleles associated with PTC bitter taste sensitivity in Africa. *Molecular Biology and Evolution*, *29*(4), 1141–1153.

Cardoso, S., Lau, W., Eiras Dias, J., Fevereiro, P., & Maniatis, N. (2012). A candidate-gene association study for berry colour and anthocyanin content in Vitis vinifera L. *PloS One*, *7*(9), e46021.

Carnicer, J. (1993). Malabsorcio de lactosa en una poblacio pre-escolar i escolar. *Butlleti de la Societat Catalana de Pediatria*, *53*(1), 26–32.

Casellas, F., & Malagelada, J. R. (2003). Applicability of short hydrogen breath test for screening of lactose malabsorption. *Digestive Diseases and Sciences*, *48*(7), 1333–1338.

Casellas, F., Varela, E., Aparici, A., Casaus, M., & Rodríguez, P. (2009). Development, validation, and applicability of a symptoms questionnaire for lactose malabsorption screening. *Digestive Diseases and Sciences*, *54*(5), 1059–1065.

Caskey, D., Payne-Bose, D., Welsh, J. D., Gearhart, H. L., Nance, M. K., & Morrison, R. D. (1977). Effects of age on lactose malabsorption in Oklahoma Native Americans as determined by breath H2 analysis. *The American Journal of Digestive Diseases*, *22*(2), 113–116.

Castro, M., & Bahamondes, M. (1984). Un aporte antropológico al conocimiento de los mecanismos de subsistencia de las comunidades de la IV Región de Chile. *Ambiente y desarrollo*, *I*(1), 143–146.

Castro, M., & Bahamondes, M. (1986). Surgimiento y transformación del sistema comunitario: Las comunidades agrícolas, IV Región, Chile. *Ambiente y Desarrollo*, *II*(1), 111–126.

Cavalli-Sforza, L. L. (1981). *Cultural Transmission and Evolution: A Quantitative Approach*. Princeton: Princeton University Press.

Cavalli-Sforza, L. L., & Edwards, W. (1967). Phylogenetic analysis. Models and estimation procedures. *American Journal of Human Genetics*, *19*(3), 233–257.

Chakraborty, R., & Jin, L. (1993). Determination of relatedness between individuals using DNA fingerprinting. *Human Biology*, *65*(6), 875–895.

Chandrashekar, J., Mueller, K. L., Hoon, M., Adler, E., Feng, L., Guo, W., Zuker, C. S., & Ryba, N. J. (2000). T2Rs function as bitter taste receptors. *Cell*, *100*(6), 703–711.

Chang, S. H., Jobling, S., Brennan, K., & Headon, D. J. (2009). Enhanced Edar signalling has pleiotropic effects on craniofacial and cutaneous glands. *PloS One*, *4*(10), e7591.

Chen, C. C., Lu, R. B., Chen, Y. C., Wang, M. F., Chang, Y. C., Li, T. K., & Yin, S. J. (1999). Interaction between the functional polymorphisms of the alcohol-metabolism genes in protection against alcoholism. *American Journal of Human Genetics*, *65*(3), 795–807.

Chilean Ministry of National Assets. Retrieved: 5 November 2010. (2010). Oficina Técnica de Comunidades Agrícolas, Ministerio de Bienes Nacionales.
URL `http://www.comunidadesagricolas.cl/`

Clarke, J. D. (2006). Antiquity of aridity in the Chilean Atacama Desert. *Geomorphology*, *73*(1-2), 101–114.

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

Colantonio, S., Lasker, G. W., Kaplan, B., & Fuster, V. (2003). Use of Surname Models in Human Population Biology: A Review of Recent Developments. *Human Biology*, *75*(6), 785–807.

Cole, T. J. (2000). Secular trends in growth. *The Proceedings of the Nutrition Society*, *59*(2), 317–324.

Coltman, D. (1998). Birth weight and neonatal survival of harbour seal pups are positively correlated with genetic variation measured by microsatellites. *Proceedings of the Royal Society B: Biological Sciences*, *265*(1398), 803–809.

Cook, G., & Al-Torki, M. (1975). High intestinal lactase concentrations in adult Arbs in Saudi Arabia. *British Medical Journal*, *3*(5976), 135–136.

Corella, D., Arregui, M., Coltell, O., Portolés, O., Guillem-Sáiz, P., Carrasco, P., Sorlí, J. V., Ortega-Azorín, C., González, J. I., & Ordovás, J. M. (2011). Association of the LCT-13910C>T polymorphism

with obesity and its modulation by dairy products in a Mediterranean population. *Obesity*, *19*(8), 1707–1714.

Courtiol, A., Rickard, I. J., Lummaa, V., Prentice, A. M., Fulford, A. J. C., & Stearns, S. C. (2013). The Demographic Transition Influences Variance in Fitness and Selection on Height and BMI in Rural Gambia. *Current Biology*, *23*(1980), 1–6.

Crawford, D. C., Bhangale, T., Li, N., Hellenthal, G., Rieder, M. J., Nickerson, D., & Stephens, M. (2004). Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nature genetics*, *36*(7), 700–706.

Cribb, R. (1991). *Nomads in Archaeology*. Cambridge: Cambridge University Press.

Crittenden, R. G., & Bennett, L. E. (2005). Cow's milk allergy: a complex disorder. *Journal of the American College of Nutrition*, *24*(sup6), 582S–591S.

Crow, J. F., & Mange, A. P. (1965). Measurement of inbreeding from the frequency of marriages between persons of the same surname. *Biodemography and Social Biology*, *12*(4), 199–203.

Curik, I. (2003). Inbreeding, microsatellite heterozygosity, and morphological traits in Lipizzan horses. *Journal of Heredity*, *94*(2), 125–132.

Darwin, C. (1997). *The Voyage of the Beagle (Wordsworth Classics of World Literature)*. Wordsworth Editions Ltd.

Darwin, G. H. (2009). Marriages between first cousins in England and their effects. *International Journal of Epidemiology [Reprint. Original Work: Fortnightly Review 24:22-41]*, *38*(6), 1429–1439.

de Silva, E., & Stumpf, M. P. H. (2004). HIV and the CCR5-Delta32 resistance allele. *Fems microbiology letters*, *241*(1), 1–12.

Dehghan, M., Al Hamad, N., Yusufali, A., Nusrath, F., Yusuf, S., & Merchant, A. T. (2005). Development of a semi-quantitative food frequency questionnaire for use in United Arab Emirates and Kuwait based on local foods. *Nutrition Journal*, *4*(18).

Dill, J. E., Levy, M., Wells, R. F., & Weser, E. (1972). Lactase deficiency in Mexican-American males. *American Journal of Clinical Nutrition*, (25), 869–870.

Dillehay, T., & Collins, M. (1988). Early cultural evidence from Monte Verde in Chile. *Nature*, *332*(10), 150–152.

Dipierri, J. E., Alfaro, E. L., Scapoli, C., Mamolini, E., Rodriguez-Larralde, A., & Barrai, I. (2005). Surnames in Argentina: a population study through isonymy. *American Journal of Physical Anthropology*, *128*(1), 199–209.

Dray, S., & Dufour, A. B. (2007). The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software*, *22*(4), 1–20.

Dubois, L., Ohm Kyvik, K., Girard, M., Tatone-Tokuda, F., Pérusse, D., Hjelmborg, J., Skytthe, A., Rasmussen, F., Wright, M. J., Lichtenstein, P., & Martin, N. G. (2012). Genetic and environmental contributions to weight, height, and BMI from birth to 19 years of age: an international study of over 12,000 twin pairs. *PloS one*, *7*(2), e30153.

Dubroeucq, D. (2004). Land cover and land use changes in relation to social evolution—a case study from Northern Chile. *Journal of Arid Environments*, *56*(2), 193–211.

Durham, W. H. (1982). Interactions of genetic and cultural evolution: Models and examples. *Human Ecology*, *10*(3), 289–323.

Dyer, R. J. (2014). *gstudio: Analyses and functions related to the spatial analysis of genetic marker data. [Software]*.
URL `http://cran.r-project.org/package=gstudio`

Earl, D. a., & VonHoldt, B. M. (2011). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, *4*(2), 359–361.

Enattah, N. S., Jensen, T. G. K., Nielsen, M., Lewinski, R., Kuokkanen, M., Rasinpera, H., El-Shanti, H., Seo, J. K., Alifrangis, M., Khalil, I. F., Natah, A., Ali, A., Natah, S., Comas, D., Mehdi, S. Q., Groop, L., Vestergaard, E. M., Imtiaz, F., Rashed, M. S., Meyer, B., Troelsen, J., & Peltonen, L. (2008). Independent introduction of two lactase-persistence alleles into human populations reflects different history of adaptation to milk culture. *American Journal of Human Genetics*, *82*(1), 57–72.

Enattah, N. S., Sahi, T., Savilahti, E., Terwilliger, J. D., Peltonen, L., & Järvelä, I. (2002). Identification of a variant associated with adult type hypolactasia. *Nature Genetics*, *30*(2), 233–237.

Enattah, N. S., Trudeau, A., Pimenoff, V., Maiuri, L., Auricchio, S., Greco, L., Rossi, M., Lentze, M., Seo, J. K., Rahgozar, S., Khalil, I., Alifrangis, M., Natah, S., Groop, L., Shaat, N., Kozlov, A., Verschubskaya, G., Comas, D., Bulayeva, K., Mehdi, S. Q., Terwilliger, J. D., Sahi, T., Savilahti, E., Perola, M., Sajantila, A., Järvelä, I., & Peltonen, L. (2007). Evidence of still-ongoing convergence evolution of the lactase persistence T-13910 alleles in humans. *American Journal of Human Genetics*, *81*(3), 615–625.

Escoboza, P. M. L., Fernandes, M. I. M., Peres, L. C., Einerhand, A. W. C., Galvão, L. C., Martins, P., Escoboza, L., Inez, M., Wilhelmina, A., Einerhand, C., & Galva, L. C. (2004). Adult-type hypo-

lactasia: clinical, morphologic and functional characteristics in Brazilian patients at a university hospital. *Journal of Pediatric Gastroenterology and Nutrition*, *39*(4), 361–365.

Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology*, *14*(8), 2611–2620.

Evershed, R., Payne, S., & Sherratt, A. (2008). Earliest date for milk use in the Near East and southeastern Europe linked to cattle herding. *Nature*, *455*(7212), 528–531.

Evsyukov, A., & Ivanov, D. (2013). Selection Variability for Arg48His in Alcohol Dehydrogenase ADH1B among Asian Populations. *Human Biology*, *85*(4), 569–577.

Falkingham, J., & Namazie, C. (2002). *Measuring health and poverty: a review of approaches to identifying the poor*. 0. London: HSRC.

Falush, D., Stephens, M., & Pritchard, J. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, *164*(4), 1567–1587.

Falush, D., Stephens, M., & Pritchard, J. K. J. K. J. (2007). Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular Ecology Notes*, *7*(4), 574–578.

Fernández, C. I., & Flores, S. V. (2014). Lactase persistence and dairy intake in Mapuche and Mestizo populations from southern Chile. *American Journal of Physical Anthropology*, *155*(3), 482–487.

Fiebig-Wittmaack, M., Astudillo, O., Wheaton, E., Wittrock, V., Perez, C., & Ibacache, A. (2011). Climatic trends and impact of climate change on agriculture in an arid Andean valley. *Climatic Change*, *111*(3-4), 819–833.

Figueroa, R. B., Melgar, E., Jó, N., & García, O. L. (1971). Intestinal lactase deficiency in an apparently normal Peruvian population. *The American Journal of Digestive Diseases*, *16*(10), 881–889.

Filzmoser, P., Fritz, H., & Kalcher, K. (2013). *pcaPP: Robust PCA by Projection Pursuit. [Software]*. URL `http://cran.r-project.org/package=pcaPP`

Fischer, A., Gilad, Y., Man, O., & Pääbo, S. (2005). Evolution of bitter taste receptors in humans and apes. *Molecular Biology and Evolution*, *22*(3), 432–436.

Fisher, R., Corbet, A., & Williams, C. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology*, *12*(1), 42–58.

Fisher, R. A. R. (1919). The Correlation Between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh*, *52*(02), 399–433.

Fix, A. A. G. (2002). Colonization models and initial genetic diversity in the Americas. *Human Biology*, *74*(1), 1–10.

Flanagan, N., Healy, E., Ray, A., Philips, S., Todd, C., Jackson, I. J., Birch-Machin, M. a., & Rees, J. L. (2000). Pleiotropic effects of the melanocortin 1 receptor (MC1R) gene on human pigmentation. *Human Molecular Genetics*, *9*(17), 2531–2537.

Flatz, G. (1987). Genetics of lactose digestion in humans. In H. Harris, & K. Hirschhorn (Eds.) *Advances in human genetics*, vol. 16, chap. 1, (pp. 1–77). New York: Springer US.

Flatz, G., & Rotthauwe, H. (1973). Lactose nutrition and natural selection. *The Lancet*, *4*(3), 192–192.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Girón, C. G., Gordon, L., Hourlier, T., Hunt, S., Johnson, N., Juettemann, T., Kähäri, A. K., Keenan, S., Kulesha, E., Martin, F. J., Maurel, T., McLaren, W. M., Murphy, D. N., Nag, R., Overduin, B., Pignatelli, M., Pritchard, B., Pritchard, E., Riat, H. S., Ruffier, M., Sheppard, D., Taylor, K., Thormann, A., Trevanion, S. J., Vullo, A., Wilder, S. P., Wilson, M., Zadissa, A., Aken, B. L., Birney, E., Cunningham, F., Harrow, J., Herrero, J., Hubbard, T. J. P., Kinsella, R., Muffato, M., Parker, A., Spudich, G., Yates, A., Zerbino, D. R., & Searle, S. M. J. (2014). Ensembl 2014. *Nucleic Acids Research*, *42*, D749–55.

Flores, C., Maca-Meyer, N., González, A. M., Oefner, P. J., Shen, P., Pérez, J. A., Rojas, A., Larruga, J. M., Underhill, P. A., Gonzalez, A. M., & Perez, J. A. (2004). Reduced genetic structure of the Iberian peninsula revealed by Y-chromosome analysis: implications for population demography. *European Journal of Human Genetics*, *12*(10), 855–863.

Fox, J., & Hong, J. (2009). Effect Displays in {R} for Multinomial and Proportional-Odds Logit Models: Extensions to the {effects} Package. *Journal of Statistical Software*, *32*(1), 1–24.

Fox, J., & Weisberg, S. (2011). *An {R} Companion to Applied Regression*. Thousand Oaks: Sage, 2nd ed.

Freckleton, R., Harvey, P., & Pagel, M. (2003). Bergmann's rule and body size in mammals. *The American Naturalist*, *161*(5), 821–825.

Freeman, B., Smith, N., Curtis, C., Huckett, L., Mill, J., & Craig, I. W. (2003). DNA from buccal swabs recruited by mail: evaluation of storage effects on long-term stability and suitability for multiplex polymerase chain reaction genotyping. *Behavior Genetics*, *33*(1), 67–72.

Fried, M., Abramson, S., & Meyer, J. H. (1987). Passage of salivary amylase through the stomach in humans. *Digestive Diseases and Sciences*, *32*(10), 1097–1103.

Friedrich, D. C., Callegari-Jacques, S. M., Petzl-Erler, M. L., Tsuneto, L., Salzano, F. M., & Hutz, M. H. (2012a). Stability or variation? Patterns of lactase gene and its enhancer region distributions in Brazilian Amerindians. *American Journal of Physical Anthropology*, *147*(3), 427–432.

Friedrich, D. C., Santos, S. E. B., Ribeiro-dos Santos, A. K. C., & Hutz, M. H. (2012b). Several different lactase persistence associated alleles and high diversity of the lactase gene in the admixed Brazilian population. *PloS One*, *7*(9), e46520.

Fujimoto, A., Kimura, R., Ohashi, J., Omi, K., Yuliwulandari, R., Batubara, L., Mustofa, M. S., Samakkarn, U., Settheetham-Ishida, W., Ishida, T., Morishita, Y., Furusawa, T., Nakazawa, M., Ohtsuka, R., & Tokunaga, K. (2008a). A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness. *Human Molecular Geneticsenetics*, *17*(6), 835–843.

Fujimoto, A., Ohashi, J., Nishida, N., Miyagawa, T., Morishita, Y., Tsunoda, T., Kimura, R., & Tokunaga, K. (2008b). A replication study confirmed the EDAR gene to be a major contributor to population differentiation regarding head hair thickness in Asia. *Human genetics*, *124*(2), 179–185.

Fursov, M., & Novikova, O. (2008). Multitasking Software System for DNA Analysis. In *The sixth international conference on bioinformatics of genome regultation and structure*, (p. 78). Novosibirsk: Institute of Cytology and Genetics: Russian Academy of Science.

Fushan, A., Simons, C. T., Slack, J. P., & Drayna, D. (2010). Association between common variation in genes encoding sweet taste signaling components and human sucrose perception. *Chemical Senses*, *35*(7), 579–592.

Gaikwad, S., Vasulu, T. S., & Kashyap, V. K. (2006). Microsatellite diversity reveals the interplay of language and geography in shaping genetic differentiation of diverse Proto-Australoid populations of west-central India. *American Journal of Physical Anthropology*, *129*(2), 260–267.

Galanter, J. M., Fernandez-Lopez, J. C., Gignoux, C. R., Barnholtz-Sloan, J., Fernandez-Rozadilla, C., Via, M., Hidalgo-Miranda, A., Contreras, A. V., Figueroa, L. U., Raska, P., Jimenez-Sanchez, G., Zolezzi, I. S., Torres, M., Ponte, C. R., Ruiz, Y., Salas, A., Nguyen, E., Eng, C., Borjas, L., Zabala, W., Barreto, G., González, F. R., Ibarra, A., Taboada, P., Porras, L., Moreno, F., Bigham, A., Gutierrez, G., Brutsaert, T., León-Velarde, F., Moore, L. G., Vargas, E., Cruz, M., Escobedo, J., Rodriguez-Santana, J., Rodriguez-Cintrón, W., Chapela, R., Ford, J. G., Bustamante, C., Seminara, D., Shriver, M., Ziv, E., Burchard, E. G., Haile, R., Parra, E., & Carracedo, A. (2012). Development of a panel of genome-wide ancestry informative markers to study admixture throughout the Americas. *PLoS Genetics*, *8*(3), e1002554.

Gallardo, G. (2002). *Communal Land Ownership in Chile: The Agricultural Communities in the Commune of Canela, Norte Chico (1600-1998) (International Land Management Series)*. Aldershot: Ashgate.

Gallego Romero, I., Basu Mallick, C., Liebert, A., Crivellaro, F., Chaubey, G., Itan, Y., Metspalu, M., Eaaswarkhanth, M., Pitchappan, R., Villems, R., Reich, D., Singh, L., Thangaraj, K., Thomas, M., Swallow, D. M., Lahr, M. M., & Kivisild, T. (2012). Herders of Indian and European cattle share their predominant allele for lactase persistence. *Molecular Biology and Evolution*, *29*(1), 249–260.

Galvani, A. P., & Slatkin, M. (2003). Evaluating plague and smallpox as historical selective pressures for the CCR5-Delta 32 HIV-resistance allele. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(25), 15276–15279.

García-Borrón, J. C., Sánchez-Laorden, B. L., & Jiménez-Cervantes, C. (2005). Melanocortin-1 receptor structure and functional regulation. *Pigment Cell Research*, *18*(6), 393–410.

GDAL Development Team. (2011). GDAL - Geospatial Data Abstraction Library. [Software]. Open Source Geospatial Foundation.
URL `http://www.gdal.org/`

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

Gibson, G. (2007). Human evolution: thrifty genes and the dairy queen. *Current Biology*, *17*(8), R295–R296.

Gifford-Gonzalez, D., & Hanotte, O. (2011). Domesticating Animals in Africa: Implications of Genetic and Archaeological Findings. *Journal of World Prehistory*, *24*(1), 1–23.

Gillespie, J. H. (2004). *Population genetics : a concise guide*. Baltimore, Md.: Johns Hopkins University Press.

Glendinning, J. I. (1994). Is the bitter rejection response always adaptive? *Physiology & Behavior*, *56*(6), 1217–1227.

Göbel, B. (1997). 'You Have to Exploit Luck': Pastoral Household Economy and the Cultural Handling of Risk and Uncertainty in the Andean Highlands. *Nomadic peoples*, *1*(1), 37–53.

Graf, J., Hodgson, R., & van Daal, A. (2005). Single nucleotide polymorphisms in the MATP gene are associated with normal human pigmentation variation. *Human Mutation*, *25*(3), 278–284.

Greenwald, E., Samachson, J., & Spencer, H. (1963). Effect of lactose on calcium metabolism in man. *The Journal of Nutrition*, *79*(63), 531–538.

Groot, P., Bleeker, M., Pronk, J., & Arwert, F. (1989). The human α-amylase multigene family consists of haplotypes with variable numbers of genes. *Genomics*, *42*, 29–42.

Guernier, V., Hochberg, M. E., & Guégan, J. F. (2004). Ecology drives the worldwide distribution of human diseases. *PLoS Biology*, *2*(6), e141.

Guix García, J., Rodrigo Gómez, J. M., Aparisi Quereda, L., Serra Desfilis, M. A., & García-Conde Gómez, F. J. (1974). [Lactose intolerance in the Spanish population]. *Revista Española de las Enfermedades del Aparato Digestivo*, *42*(4), 367–382.

Guo, S. S. W., & Thompson, E. E. A. (1992). Performing the Exact Test of Hardy-Weinberg Proportion for Multiple Alleles. *Biometrics*, *48*(2), 361–372.

Gustafsson, A., & Lindenfors, P. (2004). Human size evolution: no evolutionary allometric relationship between male and female stature. *Journal of Human Evolution*, *47*(4), 253–266.

Gustafsson, A., Werdelin, L., Tullberg, B. S., & Lindenfors, P. (2007). Stature and sexual stature dimorphism in Sweden, from the 10th to the end of the 20th century. *American Journal of Human Biology*, *19*(6), 861–870.

Hamblin, M. T., & Di Rienzo, A. (2000). Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. *American Journal of Human Genetics*, *66*(5), 1669–1679.

Han, Y., Gu, S., Oota, H., Osier, M. V., Pakstis, A. J., Speed, W. C., Kidd, J. R., & Kidd, K. K. (2007). Evidence of positive selection on a class I ADH locus. *American Journal of Human Genetics*, *80*(3), 441–456.

Hancock, A. M., Clark, V. J., Qian, Y., & Di Rienzo, A. (2011a). Population genetic analysis of the uncoupling proteins supports a role for UCP3 in human cold resistance. *Molecular Biology and Evolution*, *28*(1), 601–614.

Hancock, A. M., Witonsky, D. B., Alkorta-Aranburu, G., Beall, C. M., Gebremedhin, A., Sukernik, R., Utermann, G., Pritchard, J. K., Coop, G., & Di Rienzo, A. (2011b). Adaptations to climate-mediated selective pressures in humans. *PLoS Genetics*, *7*(4), e1001375.

Harcourt, H., & Schreier, B. M. (2009). Diversity, Body Mass, and Latitudinal Gradients in Primates. *International Journal of Primatology*, *30*(2), 283–300.

Harding, R. M., Healy, E., Ray, J., Ellis, N. S., Flanagan, N., Todd, C., Dixon, C., Sajantila, A., Jackson, I. J., Birch-Machin, M., & Rees, J. L. (2000). Evidence for variable selective pressures at MC1R. *American Journal of Human Genetics*, *66*(4), 1351–1361.

Hardy, G. (1908). Mendelian proportions in a mixed population. *Science*, *28*(N.S.), 49–50.

Harrell, Frank, E., & Dupont, C. (2014). *Hmisc: Harrell Miscellaneous. [Software]*. URL `http://cran.r-project.org/package=Hmisc`

Harris, M. (1997). *Culture, people, nature: an introduction to general anthropology*. New York: Longman, 7th ed.

Harris, M. (1998). *Good to Eat: Riddles of Food and Culture*. Illinois: Waveland Press.

Harris, M. (2000). *The Rise of Anthropological Theory: A History of Theories of Culture*. Lanham: AltaMira Press.

Harrison, G., & Morphy, H. (1998). *Human Adaptation*. Mid Glamorgan: Oxford International.

Harvard School of Public Health (2007). Semi-Quantitative Food Frequency Questionnaire for Adults. Nutrition Department, Harvard School of Public Health HSPH. URL `https://regepi.bwh.harvard.edu/health/nutrition.html`

Harvey, C., Hollox, E., & Poulter, M. (1998). Lactase haplotype frequencies in Caucasians: association with the lactase persistence/non-persistence polymorphism. *Annals of Human Genetics*, *62*(3), 215–223.

Hedrick, P., & Verrelli, B. (2006). 'Ground truth' for selection on CCR5- Δ32. *Trends in Genetics*, *22*(6), 293–296.

Hedrick, P. P. W. (1971). A new approach to measuring genetic similarity. *Evolution*, *25*(2), 276–280.

Hedrick, P. W. (2005). A standardized genetic differentiation measure. *Evolution*, *59*(8), 1633–1638.

Hedrick, P. W. (2011). Population genetics of malaria resistance in humans. *Heredity*, *107*(4), 283–304.

Heiat, A., Vaccarino, V., & Krumholz, H. M. (2001). An evidence-based assessment of federal guidelines for overweight and obesity as they apply to elderly persons. *Archives of Internal Medicine*, *161*(9), 1194–1203.

Heiberger, R. M. (2014). *HH: Statistical Analysis and Data Display: Heiberger and Holland. [Software]*. URL `http://cran.r-project.org/package=HH`

Hess, C. G. (1990). "Moving up-Moving down": Agro-pastoral land-use patterns in the Ecuadorian Paramos. *Mountain Research and Development*, *10*(4), 333–342.

Hey, J. (2005). On the number of New World founders: a population genetic portrait of the peopling of the Americas. *PLoS Biology*, *3*(6), e193.

Hijmans, R. J. (2014). *raster: raster: Geographic data analysis and modeling. [Software]*.
URL `http://cran.r-project.org/package=raster`

Hill, C. R., Duewer, D. L., Kline, M. C., Coble, M. D., & Butler, J. M. (2013). U.S. population data for 29 autosomal STR loci. *Forensic Science International. Genetics*, *7*(3), e82–3.

Hoffman, D. (2001). Obesity in developing countries: causes and implications. *Food, Nutrition and Agriculture*, (28), 35–44.

Hoffman, G. E. (2013). Correcting for population structure and kinship using the linear mixed model: theory and extensions. *PloS One*, *8*(10), e75707.

Holden, C., & Mace, R. (1997). Phylogenetic analysis of the evolution of lactose digestion in adults. *Human Biology*, *69*(5), 605–628.

Holden, C., & Mace, R. (1999). Sexual dimorphism in stature and women's work: a phylogenetic cross-cultural analysis. *American Journal of Physical Anthropology*, *110*(1), 27–45.

Hollox, E. J., Poulter, M., Zvarik, M., Ferak, V., Krause, A., Jenkins, T., Saha, N., Kozlov, I., & Swallow, D. M. (2001). Lactase haplotype diversity in the Old World. *American Journal of Human Genetics*, *68*(1), 160–172.

Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A Program for Missing Data. *Journal of Statistical Software*, *45*(7), 1–47.

Hossain, P., Kawar, B., & El Nahas, M. (2007). Obesity and diabetes in the developing world–a growing challenge. *The New England Journal of Medicine*, *356*(3), 213–215.

Hubisz, M. J. M., Falush, D., Stephens, M., & Pritchard, J. (2009). Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources*, *9*(5), 1322–1332.

Husson, F., Josse, J., Le, S., & Mazet, J. (2013). *FactoMineR: Multivariate Exploratory Data Analysis and Data Mining with R. [Software]*.
URL `http://cran.r-project.org/package=FactoMineR`

Imtiaz, F., Savilahti, E., Sarnesto, A., Trabzuni, D., Al-Kahtani, K., Kagevi, I., Rashed, M. S., Meyer, B. F., & Järvelä, I. (2007). The T/G 13915 variant upstream of the lactase gene (LCT) is the founder allele of lactase persistence in an urban Saudi population. *Journal of Medical Genetics*, *44*(10), e89.

Ingram, C. J. E., Elamin, M. F., Mulcare, C., Weale, M. E., Tarekegn, A., Raga, T. O., Bekele, E., Elamin, F. M., Thomas, M. G., Bradman, N., & Swallow, D. M. (2007). A novel polymorphism associated

with lactose tolerance in Africa: multiple causes for lactase persistence? *Human Genetics*, *120*(6), 779–788.

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

Ingram, C. J. E., Raga, T. O., Tarekegn, A., Browning, S. L., Elamin, M. F., Bekele, E., Thomas, M. G., Weale, M. E., Bradman, N., & Swallow, D. M. (2009b). 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.

Ingstad, A. (1970). *The Norse Settlement at L'Anse Aux Meadows, Newfoundland: A Preliminary Report from the Excavations, 1961-1968*. Nuussuaq: Munksgaard.

Ingstad, H., & Friis, E. (1969). *Westward to Vinland: the discovery of pre-Columbian Norse house-sites in North America*. New York: St. Matin Press.

Instituto Nacional de Estadisticas (2007). VII Censo Nacional Agropecuario y Forestal. Instituto Nacional de Estadisticas INE, Chile. Tech. rep.

Instituto Nacional De Investigaciones Agropecuarias (2005). Estudio "Diseño, Implementacion y Seguimiento Plan integral de desarrollo del Secano, IV Región de Coquimbo". Etapa 1. Reconocimiento Detallado del Territorio a Intervenir. Tech. rep., Instituto Nacional De Investigaciones Agropecuarias, Chile.

Itan, Y., Jones, B. L., Ingram, C. J. E., Swallow, D. M., & Thomas, M. G. (2010). A worldwide correlation of lactase persistence phenotype and genotypes. *BMC Evolutionary Biology*, *10*(36).

Itan, Y., Powell, A., Beaumont, M. A., Burger, J., & Thomas, M. G. (2009). The origins of lactase persistence in Europe. *PLoS Computational Biology*, *5*(8), e1000491.

Izagirre, N., García, I., Junquera, C., de la Rúa, C., & Alonso, S. (2006). A scan for signatures of positive selection in candidate loci for skin pigmentation in humans. *Molecular Biology and Evolution*, *23*(9), 1697–1706.

Jablonski, N. G., & Chaplin, G. (2000). The evolution of human skin coloration. *Journal of Human Evolution*, *39*(1), 57–106.

Jackson, D., Mendez, C., Seguel, R., Maldonado, A., & Vargas, G. (2007). Initial Occupation of the Pacific Coast of Chile during Late Pleistocene Times. *Current Anthropology*, *48*(5), 725–731.

Jakobsson, M., & Rosenberg, N. (2007). CLUMPP: a cluster matching and permutation program for

dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, *23*(14), 1801–1806.

Jobling, M. (2001). In the name of the father: surnames and genetics. *Trends in genetics*, *17*(6), 353–357.

Jobling, M., Hollox, E., Hurles, M., Kivisild, T., & Tyler-Smith, C. (2013). *Human evolutionary genetics*. New York: Garland Science.

Johnson, J., Simoons, F., Hurwitz, R., & A (1978). Lactose malabsorption among adult Indians of the Great Basin and American Southwest. *American Journal of Clinical Nutrition*, *31*(3), 381–387.

Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, *24*(11), 1403–1405.

Jombart, T., & Ahmed, I. (2011). adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics*, *27*(21), 3070–3071.

Jones, B. L., Raga, T. O., Liebert, A., Zmarz, P., Bekele, E., Danielsen, E. T., Olsen, A. K., Bradman, N., Troelsen, J. T., & Swallow, D. M. (2013). Diversity of lactase persistence alleles in ethiopia: signature of a soft selective sweep. *American Journal of Human Genetics*, *93*(3), 538–544.

Jorde, L. B., & Morgan, K. (1987). Genetic structure of the Utah Mormons: isonymy analysis. *American Journal of Physical Anthropology*, *72*(3), 403–412.

Jost, L. (2008). G ST and its relatives do not measure differentiation. *Molecular Ecology*, *17*(18), 4015–4026.

Jost, L. (2009). D vs. GST: response to Heller and Siegismund (2009) and Ryman and Leimar (2009). *Molecular Ecology*, *18*(10), 2088–2091.

Julian, C. G., Wilson, M. J., & Moore, L. G. (2009). Evolutionary adaptation to high altitude: a view from in utero. *American Journal of Human Biology*, *21*(5), 614–622.

Kabacoff, R. (2011). *R in Action: Data analysis and graphics with R*. Shelter Island: Manning Publications.

Kamberov, Y. G., Wang, S., Tan, J., Gerbault, P., Wark, A., Tan, L., Yang, Y., Li, S., Tang, K., Chen, H., Powell, A., Itan, Y., Fuller, D., Lohmueller, J., Mao, J., Schachar, A., Paymer, M., Hostetter, E., Byrne, E., Burnett, M., McMahon, A. P., Thomas, M. G., Lieberman, D. E., Jin, L., Tabin, C. J., Morgan, B., & Sabeti, P. C. (2013). Modeling recent human evolution in mice by expression of a selected EDAR variant. *Cell*, *152*(4), 691–702.

Kanaghinis, T., Hatzioannou, J., Deliargyris, N., Danos, N., Zografos, N., Katsas, A., & Gardikas, C. (1974). Primary lactase deficiency in Greek adults. *The American Journal of Digestive Diseases*, *19*(11), 1021–1027.

Kanazawa, S. (2005). Big and tall parents have more sons: further generalizations of the Trivers-Willard hypothesis. *Journal of Theoretical Biology*, *235*(4), 583–590.

Kang, H. M., Zaitlen, N., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., & Eskin, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics*, *178*(3), 1709–1723.

Kaplan, H., & Lancaster, J. (1995). Fertility and fitness among Albuquerque men: a competitive labour market theory. In R. Dunbar (Ed.) *Human reproductive decisions: Biological and social perspectives*, (pp. 96–136). London: St. Martin Press.

Kaplan, H. S., Lancaster, J. B., Johnson, S. E., & Bock, J. (1995). Does observed fertility maximize fitness among New Mexican men? : A test of an optimality model and a new theory of parental investment in the embodied capital of offspring. *Human Nature*, *6*(4), 325–360.

Kashyap, V. K., Guha, S., Sitalaximi, T., Bindu, G. H., Hasnain, S. E., & Trivedi, R. (2006). Genetic structure of Indian populations based on fifteen autosomal microsatellite loci. *BMC Genetics*, *7*(28).

Katzmarzyk, P. T., & Leonard, W. R. (1998). Climatic influences on human body size and proportions: ecological adaptations and secular trends. *American journal of physical anthropology*, *106*(4), 483–503.

Keitt, T. (2012). *colorRamps: Builds color tables. [Software]*.
URL `http://cran.r-project.org/package=colorRamps`

King, T. E., & Jobling, M. (2009). Founders, drift, and infidelity: the relationship between Y chromosome diversity and patrilineal surnames. *Molecular Biology and Evolution*, *26*(5), 1093–1102.

Kingfisher, C. P., & Millard, a. V. (1998). "Milk makes me sick but my body needs it": conflict and contradiction in the establishment of authoritative knowledge. *Medical Anthropology Quarterly*, *12*(4), 447–466.

Kirsch, S., Weiss, B., Kleiman, S., Roberts, K., Pryor, J., Milunsky, A., Ferlin, A., Foresta, C., Matthijs, G., & Rappold, G. (2002). Localisation of the Y chromosome stature gene to a 700 kb interval in close proximity to the centromere. *Journal of Medical Genetics*, *39*(7), 507–513.

Kovats, R. S., & Hajat, S. (2008). Heat stress and public health: a critical review. *Annual Review of Public Health*, *29*, 41–55.

Kruse, T., Bolund, L., Grzeschik, K. H., Ropers, H. H., Sjöström, H., Norén, O., Mantei, N., & Semenza, G. (1988). The human lactase-phlorizin hydrolase gene is located on chromosome 2. *Febs letters*, *240*(1-2), 123–126.

Kuh, D. L., Power, C., & Rodgers, B. (1991). Secular trends in social class and sex differences in adult height. *International Journal of Epidemiology*, *20*(4), 1001–1009.

Kumanyika, S. (2003). Relative Validity of Food Frequency Questionnaire Nutrient Estimates in the Black Women's Health Study. *Annals of Epidemiology*, *13*(2), 111–118.

Kwiatkowski, D. P. (2005). How malaria has affected the human genome and what human genetics can teach us about malaria. *American Journal of Human Genetics*, *77*(2), 171–192.

Lacassie, Y., Weinberg, R., & Mönckeberg, F. (1978). Poor predictability of lactose malabsorption from clinical symptoms for Chilean populations. *The American Journal of Clinical Nutrition*, *31*(5), 799–804.

Ladas, S., Papanikos, J., & Arapakis, G. (1982). Lactose malabsorption in Greek adults: correlation of small bowel transit time with the severity of lactose intolerance. *Gut*, *23*(11), 968–973.

Lalueza-Fox, C., Gigli, E., de la Rasilla, M., Fortea, J., & Rosas, A. (2009). Bitter taste perception in Neanderthals through the analysis of the TAS2R38 gene. *Biology Letters*, *5*(6), 809–811.

Lamason, R. L., Mohideen, M.-A. P. K., Mest, J. R., Wong, A. C., Norton, H. L., Aros, M. C., Jurynec, M. J., Mao, X., Humphreville, V. R., Humbert, J. E., Sinha, S., Moore, J. L., Jagadeeswaran, P., Zhao, W., Ning, G., Makalowska, I., McKeigue, P. M., O'donnell, D., Kittles, R., Parra, E. J., Mangini, N. J., Grunwald, D. J., Shriver, M. D., Canfield, V. a., & Cheng, K. C. (2005). SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science*, *310*(5755), 1782–1786.

Lamri, A., Poli, A., Emery, N., Bellili, N., Velho, G., Lantieri, O., Balkau, B., Marre, M., & Fumeron, F. (2013). The lactase persistence genotype is associated with body mass index and dairy consumption in the D.E.S.I.R. study. *Metabolism*, *62*(9), 1323–1329.

Lango Allen, H., Estrada, K., Lettre, G., Berndt, S. I., Weedon, M. N., Rivadeneira, F., Willer, C. J., Jackson, A. U., Vedantam, S., Raychaudhuri, S., Ferreira, T., Wood, A. R., Weyant, R. J., Segrè, A. V., Speliotes, E. K., Wheeler, E., Soranzo, N., Park, J. H., Yang, J., Gudbjartsson, D., Heard-Costa, N. L., Randall, J. C., Qi, L., Vernon Smith, A., Mägi, R., Pastinen, T., Liang, L., Heid, I. M., Luan, J., Thorleifsson, G., Winkler, T. W., Goddard, M. E., Sin Lo, K., Palmer, C., Workalemahu, T., Aulchenko, Y. S., Johansson, A., Zillikens, M. C., Feitosa, M. F., Esko, T., Johnson, T., Ketkar, S.,

Kraft, P., Mangino, M., Prokopenko, I., Absher, D., Albrecht, E., Ernst, F., Glazer, N. L., Hayward, C., Hottenga, J. J., Jacobs, K. B., Knowles, J. W., Kutalik, Z., Monda, K. L., Polasek, O., Preuss, M., Rayner, N. W., Robertson, N. R., Steinthorsdottir, V., Tyrer, J. P., Voight, B. F., Wiklund, F., Xu, J., Zhao, J. H., Nyholt, D. R., Pellikka, N., Perola, M., Perry, J. R. B., Surakka, I., Tammesoo, M. L., Altmaier, E. L., Amin, N., Aspelund, T., Bhangale, T., Boucher, G., Chasman, D. I., Chen, C., Coin, L., Cooper, M. N., Dixon, A. L., Gibson, Q., Grundberg, E., Hao, K., Juhani Junttila, M., Kaplan, L. M., Kettunen, J., König, I. R., Kwan, T., Lawrence, R. W., Levinson, D. F., Lorentzon, M., McKnight, B., Morris, A. P., Müller, M., Suh Ngwa, J., Purcell, S., Rafelt, S., Salem, R. M., Salvi, E., Sanna, S., Shi, J., Sovio, U., Thompson, J. R., Turchin, M. C., Vandenput, L., Verlaan, D. J., Vitart, V., White, C. C., Ziegler, A., Almgren, P., Balmforth, A. J., Campbell, H., Citterio, L., De Grandi, A., Dominiczak, A., Duan, J., Elliott, P., Elosua, R., Eriksson, J. G., Freimer, N. B., Geus, E. J. C., Glorioso, N., Haiqing, S., Hartikainen, A. L., Havulinna, A. S., Hicks, A. a., Hui, J., Igl, W., Illig, T., Jula, A., Kajantie, E., Kilpeläinen, T. O., Koiranen, M., Kolcic, I., Koskinen, S., Kovacs, P., Laitinen, J., Liu, J., Lokki, M. L., Marusic, A., Maschio, A., Meitinger, T., Mulas, A., Paré, G., Parker, A. N., Peden, J. F., Petersmann, A., Pichler, I., Pietiläinen, K. H., Pouta, A., Ridderstrale, M., Rotter, J. I., Sambrook, J. G., Sanders, A. R., Schmidt, C. O., Sinisalo, J., Smit, J. H., Stringham, H. M., Bragi Walters, G., Widen, E., Wild, S. H., Willemsen, G., Zagato, L., Zgaga, L., Zitting, P., Alavere, H., Farrall, M., McArdle, W. L., Nelis, M., Peters, M. J., Ripatti, S., van Meurs, J. B. J., Aben, K. K., Ardlie, K. G., Beckmann, J. S., Beilby, J. P., Bergman, R. N., Bergmann, S., Collins, F. S., Cusi, D., den Heijer, M., Eiriksdottir, G., Gejman, P. V., Hall, A. S., Hamsten, A., Huikuri, H. V., Iribarren, C., Kähönen, M., Kaprio, J., Kathiresan, S., Kiemeney, L., Kocher, T., Launer, L. J., Lehtimäki, T., Melander, O., Mosley, T. H., Musk, A. W., Nieminen, M. S., O'Donnell, C. J., Ohlsson, C., Oostra, B., Palmer, L. J., Raitakari, O., Ridker, P. M., Rioux, J. D., Rissanen, A., Rivolta, C., Schunkert, H., Shuldiner, A. R., Siscovick, D. S., Stumvoll, M., Tönjes, A., Tuomilehto, J., van Ommen, G.-J., Viikari, J., Heath, A. C., Martin, N. G., Montgomery, G. W., Province, M. a., Kayser, M., Arnold, A. M., Atwood, L. D., Boerwinkle, E., Chanock, S. J., Deloukas, P., Gieger, C., Grönberg, H., Hall, P., Hattersley, A. T., Hengstenberg, C., Hoffman, W., Lathrop, G. M., Salomaa, V., Schreiber, S., Uda, M., Waterworth, D., Wright, A. F., Assimes, T. L., Barroso, I., Hofman, A., Mohlke, K. L., Boomsma, D. I., Caulfield, M. J., Cupples, L. A., Erdmann, J., Fox, C. S., Gudnason, V., Gyllensten, U., Harris, T. B., Hayes, R. B., Jarvelin, M. R., Mooser, V., Munroe, P. B., Ouwehand, W. H., Penninx, B. W., Pramstaller, P. P., Quertermous, T., Rudan, I., Samani, N. J., Spector, T. D., Völzke, H., Watkins, H., Wilson, J. F., Groop, L. C., Haritunians, T., Hu, F. B., Kaplan, R. C., Metspalu, A., North, K. E., Schlessinger, D., Wareham, N. J., Hunter, D. J., O'Connell, J. R., Strachan, D. P., Wichmann, H. E., Borecki, I. B., van Duijn, C. M., Schadt, E. E., Thorsteinsdottir, U., Peltonen, L., Uitterlinden, A. G., Visscher, P. M., Chatterjee, N., Loos, R. J. F., Boehnke, M., McCarthy, M. I., Ingelsson, E., Lindgren, C. M., Abecasis, G. R., Stefansson, K.,

Frayling, T. M., & Hirschhorn, J. N. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, *467*(7317), 832–838.

Lao, O., de Gruijter, J. M., van Duijn, K., Navarro, A., & Kayser, M. (2007). Signatures of positive selection in genes associated with human skin pigmentation as revealed from analyses of single nucleotide polymorphisms. *Annals of Human Genetics*, *71*(3), 354–369.

Lasker, G. (1977). A coefficient of relationship by isonymy: A method for estimating the genetic relationship between populations. *Human Biology*, *49*(3), 489–493.

Lasker, G., & Kaplan, B. (1985). Surnames and genetic structure: Repetition of the same pairs of names of married couples, a measure of subdivision of the population. *Human Biology*, *3*(3), 431–440.

Lauritsen, J., & Bruus, M. (2008). EpiData 3.1. A comprehensive tool for validated entry and documentation of data. [Software]. EpiData Association.
URL `http://www.epidata.dk/credit.htm`

Lawson, D. W., & Mace, R. (2008). Sibling configuration and childhood growth in contemporary British families. *International Journal of Epidemiology*, *37*(6), 1408–1421.

Lebenthal, E. (1987). Role of salivary amylase in gastric and intestinal digestion of starch. *Digestive Diseases and Sciences*, *32*(10), 1155–1157.

Lee, R., Xiong, G., & Kofonow, J. (2012). T2R38 taste receptor polymorphisms underlie susceptibility to upper respiratory infection. *The Journal of Clinical Investigation*, *122*(11), 4145–4159.

Leichter, J. (1973). Effect of dietary lactose on intestinal lactase activity in young rats. *The Journal of Nutrition*, *103*(3), 392–396.

Leichter, J., & Lee, M. (1971). Lactose intolerance in Canadian West Coast Indians. *The American Journal of Digestive Diseases*, *16*(9), 809–813.

Leis, R., Tojo, R., Pavón, P., & Douwes, A. (1997). Prevalence of lactose malabsorption in Galicia. *Journal of Pediatric Gastroenterology and Nutrition*, *25*(3), 296–300.

Leonard, W., & Crawford, M. (Eds.) (2008). *The Human Biology of Pastoral Populations*. Cambridge: Cambridge University Press.

Leonard, W. R., Snodgrass, J. J., & Sorensen, M. V. (2005). Metabolic Adaptation in Indigenous Siberian Populations. *Annual Review of Anthropology*, *34*(1), 451–471.

Leonard, W. R., Sorensen, M. V., Galloway, V., Spencer, G. J., Mosher, M. J., Osipova, L., & Spitsyn, V.

(2002). Climatic influences on basal metabolic rates among circumpolar populations. *American journal of Human Biology*, *14*(5), 609–620.

Levitt, M., Wilt, T., & Shaukat, A. (2013). Clinical Implications of Lactose Malabsorption Versus Lactose Intolerance. *Journal of Clinical Gastroenterology*, *47*(6), 471–480.

Lewontin, R. C. (2001). *The Doctrine of DNA*. London: Penguin Books Ltd.

LGC Genomics (2013). KASP genotyping chemistry: User guide and manual. LGC Genomics.

Li, H., Borinskaya, S., Yoshimura, K., Kal'ina, N., Marusin, A., Stepanov, V., Qin, Z., Khaliq, S., Lee, M. Y., Yang, Y., Mohyuddin, A., Gurwitz, D., Mehdi, S. Q., Rogaev, E., Jin, L., Yankovsky, N. K., Kidd, J. R., & Kidd, K. K. (2009). Refined geographic distribution of the oriental ALDH2*504Lys (nee 487Lys) variant. *Annals of Human Genetics*, *73*(3), 335–345.

Li, N., & Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, *165*(4), 2213–2233.

Lindenbaum, S. (2008). Review. Understanding kuru: the contribution of anthropology and medicine. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *363*(1510), 3715–3720.

Lischer, H. E. L., & Excoffier, L. (2012). PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*, *28*(2), 298–299.

Lisker, R., López-Habib, G., Lopez-Habib, G., Daltabuit, M., Rostenberg, I., Arroyo, P., & I (1974). Lactase deficiency in a rural area of Mexico. *American Journal of Clinical Nutrition*, *27*, 756–759.

Lizcano, F. (2005). Composición étnica de las tres áreas culturales del Continente Americano al comienzo del siglo XXI. *Convergencia*, (38), 185–232.

Longley, P., Webber, R., & Lloyd, D. (2007). The quantitative analysis of family names: Historic migration and the present day neighborhood structure of Middlesbrough, United Kingdom. *Annals of the Association of American Geographers*, *97*(1), 1–52.

Lorandi, A., Dillehay, T., & Netherly, P. (1988). Los diaguitas y el Tawantinsuyu. Una hipótesis de conflicto. In T. Dillehay, & P. Netherly (Eds.) *La frontera del Estado Inca*, vol. 442, (p. 235). Quito: Fundación Alexander von Humboldt.

Luca, F., Perry, G., & Rienzo, A. D. (2010). Evolutionary adaptations to dietary changes. *Annual Review of Nutrition*, *30*, 291–314.

Lumley, T., & Miller, A. (2009). *leaps: regression subset selection. [Software]*.
URL `http://cran.r-project.org/package=leaps`

Lynch, T. F. (1983). Camelid pastoralism and the emergence of Tiwanaku civilization in the South Central Andes. *World Archaeology*, *15*(1), 1–14.

Mace, R. (1991). Overgrazing overstated. *Nature*, *349*, 280–281.

Mace, R. (2000). Evolutionary ecology of human life history. *Animal Behaviour*, *59*(1), 1–10.

Mace, R. (2014). When not to have another baby: An evolutionary approach to low fertility. *Demographic Research*, *30*(April), 1074–1096.

Mace, R., Anderson, D., & Bierschenk, T. (1993). Transitions Between Cultivation and Pastoralism in Sub-Saharan Africa. *Current Anthropology*, *34*(4), 363–382.

Maggi, R., Sayagues, B., Fernandez, A., & B (1987). Lactose malabsorption and intolerance in Uruguayan population by breath hydrogen test (H2). *Journal of Pediatric Gastroenterology and Nutrition*, *6*(3), 373–376.

Malina, R., & Reyes, M. (2007). Overweight and obesity in a rural Amerindian population in Oaxaca, southern Mexico, 1968–2000. *American Journal of Human Biology*, *721*(July), 711–721.

Manco, L., Pires, S., Lopes, A. I., Figueiredo, I., Albuquerque, D., Alvarez, M., Rocha, J., & Abade, A. (2013). Distribution of the - 13910C>T polymorphism in the general population of Portugal and in subjects with gastrointestinal complaints associated with milk consumption. *Annals of Human Biology*, *40*(2), 205–208.

Manni, F., Guerard, E., & Heyer, E. (2004). Geographic patterns of (genetic, morphologic, linguistic) variation: how barriers can be detected by using Monmonier's algorithm. *Human Biology*, *76*(2), 173–190.

Mao, X., Bigham, A. W., Mei, R., Gutierrez, G., Weiss, K. M., Brutsaert, T. D., Leon-Velarde, F., Moore, L. G., Vargas, E., McKeigue, P. M., Shriver, M. D., & Parra, E. J. (2007). A genomewide admixture mapping panel for Hispanic/Latino populations. *American Journal of Human Genetics*, *80*(6), 1171–1178.

Marks, G., Hughes, M., & van der Pols, J. (2006). Relative validity of food intake estimates using a food frequency questionnaire is associated with sex, age, and other personal characteristics. *The Journal of Nutrition*, *136*(August 2005), 459–465.

Marshall, F., & Hildebrand, E. (2002). Cattle before crops: the beginnings of food production in Africa. *Journal of World Prehistory*, *16*(2), 99–143.

Martin, M. P. (1998). Genetic Acceleration of AIDS Progression by a Promoter Variant of CCR5. *Science*, *282*(5395), 1907–1911.

Mateu, E., Calafell, F., Ramos, M. D., Casals, T., & Bertranpetit, J. (2002). Can a place of origin of the main cystic fibrosis mutations be identified? *American Journal of Human Genetics*, *70*(1), 257–264.

Mazess, R. (1975). Biological adaptation: aptitudes and acclimatization. In E. S. Watts, F. E. Johnston, & G. W. Lasker (Eds.) *Biosocial interrelations in population adaptation*, McCutcheon 1964, (p. 918). The Hague: Mounton Publishers.

McTiernan, A., Wu, L., Chen, C., Chlebowski, R., Mossavar-Rahmani, Y., Modugno, F., Perri, M. G., Stanczyk, F. Z., Van Horn, L., & Wang, C. Y. (2006). Relation of BMI and physical activity to sex hormones in postmenopausal women. *Obesity*, *14*(9), 1662–1677.

Mead, S., Stumpf, M. P. H., Whitfield, J., Beck, J., Poulter, M., Campbell, T., Uphill, J. B., Goldstein, D., Alpers, M., Fisher, E. M. C., & Collinge, J. (2003). Balancing selection at the prion protein gene consistent with prehistoric kurulike epidemics. *Science*, *300*(5619), 640–643.

Mead, S., Whitfield, J., Poulter, M., Shah, P., Uphill, J., Beck, J., Campbell, T., Al-Dujaily, H., Hummerich, H., Alpers, M. P., & Collinge, J. (2008). Genetic susceptibility, evolution and the kuru epidemic. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *363*(1510), 3741–3746.

Mead, S., Whitfield, J., Poulter, M., Shah, P., Uphill, J., Campbell, T., Al-Dujaily, H., Hummerich, H., Beck, J., Mein, C., Verzilli, C., Whittaker, J., Alpers, M. P., & Collinge, J. (2009). A novel protective prion protein variant that colocalizes with kuru exposure. *The New England Journal of Medicine*, *361*(21), 2056–2065.

Medina, E., & Kaempffer, A. M. (1979). *Elementos de Salud Pública*. Santiago: Biblioteca Digital de la Universidad de Chile.

Mellafe, R. (1981). Latifundio y poder rural en Chile de los siglos XVII y XVIII. *Cuadernos de Historia*, *1*(1), 87–108.

Meloni, G. F., Colombo, C., La Vecchia, C., Pacifico, A., Tomasi, P., Ogana, A., Marinaro, M., & Meloni, T. (2001). High prevalence of lactose absorbers in Northern Sardinian patients with type 1 and type 2 diabetes mellitus. *The American Journal of Clinical Nutrition*, *73*(3), 582–585.

Meza, J. S. (2003). The Demographic-Epidemiological Transition in Chile, 1960-2001. *Revista Española de Salud Pública*, *77*(5), 605–613.

Migliano, A. B., Vinicius, L., & Lahr, M. M. (2007). Life history trade-offs explain the evolution of human pygmies. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(51), 20216–20219.

Moorad, J. (2013). A demographic transition altered the strength of selection for fitness and age-specific survival and fertility in a 19th century American population. *Evolution*, *67*(6), 1622–1634.

Moore, L. G., Charles, S. M., & Julian, C. G. (2011). Humans at high altitude: hypoxia and fetal growth. *Respiratory Physiology & Neurobiology*, *178*(1), 181–190.

Morales, C., & Parada, S. (2005). *Pobreza, desertificación y degradación de los recursos naturales*. Santiago: Cepal.

Morales, E., Azocar, L., Maul, X., Perez, C., Chianale, J., & Miquel, J. F. (2011). The European lactase persistence genotype determines the lactase persistence state and correlates with gastrointestinal symptoms in the Hispanic and Amerindian. *BMJ Open*, *1*(1), e000125.

Mulcare, C. A., Weale, M. E., Jones, A. L., Connell, B., Zeitlyn, D., Tarekegn, A., Swallow, D. M., Bradman, N., & Thomas, M. G. (2004). The T allele of a single-nucleotide polymorphism 13.9 kb upstream of the lactase gene (LCT) (C-13.9kbT) does not predict or cause the lactase-persistence phenotype in Africans. *American Journal of Human Genetics*, *74*(6), 1102–1110.

Mulder, M. B. (1998). The demographic transition: are we any closer to an evolutionary explanation? *Trends in Ecology & Evolution*, *13*(7), 266–270.

Mulligan, C. J., Robin, R. W., Osier, M. V., Sambuughin, N., Goldfarb, L. G., Kittles, R., Hesselbrock, D., Goldman, D., & Long, J. C. (2003). Allelic variation at alcohol metabolism genes (ADH1B, ADH1C, ALDH2) and alcohol dependence in an American Indian population. *Human Genetics*, *113*(4), 325–336.

Murthy, M. S., & Haworth, J. C. (1970). Intestinal lactase deficiency among east Indians. An adaptive rather than a genetically inherited phenomenon? *The American Journal of Gastroenterology*, *53*(3), 246–251.

Myles, S., Bouzekri, N., Haverfield, E., Cherkaoui, M., & JM (2005). Genetic evidence in support of a shared Eurasian-North African dairying origin. *Human Genetics*, *117*(1), 34–42.

Neel, J. (1962). Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? *American Journal of Human Genetics*, *77*(8), 694–703.

Neff, B. D. (2004). Mean d2 and divergence time: transformations and standardizations. *Journal of Heredity*, *95*(2), 165–171.

Nei, M. (1972). Genetic distance between populations. *American Naturalist*, *106*(949), 283–292.

Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences of the United States of America*, *70*(12), 3321–3323.

Nei, M., & Saitou, N. (1986). Genetic relationship of human populations and ethnic differences in reaction to drugs and food. *Progress in Clinical and Biological Research*, *214*, 21–37.

Nenadic, O., & Greenacre, M. (2007). Correspondence Analysis in R, with two- and three-dimensional graphics: The ca package. *Journal of Statistical Software*, *20*(3), 1–13.

Nettle, D. (2002a). Height and reproductive success in a cohort of British men. *Human Nature*, *13*(4), 473–491.

Nettle, D. (2002b). Women's height, reproductive success and the evolution of sexual dimorphism in modern humans. *Proceedings of the Royal Society B: Biological Sciences*, *269*(1503), 1919–1923.

Neuwirth, E. (2011). *RColorBrewer: ColorBrewer palettes*.
URL `http://cran.r-project.org/package=RColorBrewer`

Niu, T. (2004). Algorithms for inferring haplotypes. *Genetic Epidemiology*, *27*(4), 334–347.

Norton, H. L., Kittles, R., Parra, E., McKeigue, P., Mao, X., Cheng, K., Canfield, V., Bradley, D. G., McEvoy, B., & Shriver, M. D. (2007). Genetic evidence for the convergent evolution of light skin in Europeans and East Asians. *Molecular Biology and Evolution*, *24*(3), 710–722.

Novoa, J. E., & López, D. (2001). IV Región: El Escenario Geográfico Físico. In F. Squeo, G. Arancio, & J. Gutiérrez (Eds.) *Libro Rojo de la Flora Nativa de los Sitios Prioritarios para su Conservación: Región de Coquimbo*, chap. 2, (pp. 13 – 28). La Serena: Universidad de La Serena.

Nychka, D., Furrer, R., & Sain, S. (2013). *fields: Tools for spatial data. [Software]*.
URL `http://cran.r-project.org/package=fields`

Okonechnikov, K., Golosova, O., & Fursov, M. (2012). Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics*, *28*(8), 1166–1167.

Olds, L. C., & Sibley, E. (2003). Lactase persistence DNA variant enhances lactase promoter activity in vitro: functional role as a cis regulatory element. *Human Molecular Genetics*, *12*(18), 2333–2340.

Oota, H., Pakstis, A. J., Bonne-Tamir, B., Goldman, D., Grigorenko, E., Kajuna, S. L. B., Karoma, N. J., Kungulilo, S., Lu, R. B., Odunsi, K., Okonofua, F., Zhukova, O. V., Kidd, J. R., & Kidd, K. K. (2004). The evolution and population genetics of the ALDH2 locus: random genetic drift, selection, and low levels of recombination. *Annals of Human Genetics*, *68*(2), 93–109.

Orr, H. A. (2005). The genetic theory of adaptation: a brief history. *Nature Reviews Genetics*, *6*(2), 119–127.

Osier, M. V., Lu, R.-B., Pakstis, a. J., Kidd, J. R., Huang, S.-Y., & Kidd, K. K. (2004). Possible epistatic role of ADH7 in the protection against alcoholism. *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics.*, *126B*(1), 19–22.

Osier, M. V., Pakstis, A. J., Soodyall, H., Comas, D., Goldman, D., Odunsi, A., Okonofua, F., Parnas, J., Schulz, L. O., Bertranpetit, J., Bonne-Tamir, B., Lu, R. B., Kidd, J. R., & Kidd, K. K. (2002). A global perspective on genetic variation at the ADH genes reveals unusual patterns of linkage disequilibrium and diversity. *American Journal of Human Genetics*, *71*(1), 84–99.

O'Hara, R. B., & Kotze, D. J. (2010). Do not log-transform count data. *Methods in Ecology and Evolution*, *1*(2), 118–122.

Paige, D. M., Leonardo, E., Cordano, A., Nakashima, J., Adrianzen, B., & Graham, G. (1972). Lactose intolerance in Peruvian children: effect of age and early nutrition. *American Journal of Clinical Nutrition*, *25*(3), 297–301.

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, *20*(2), 289–290.

Pebesma, E. J., & Bivand, R. S. (2005). Classes and methods for spatial data in {R}. *R News*, *5*(2), 9–13.

Pemberton, T. J., & Rosenberg, N. (2014). Population-genetic influences on genomic estimates of the inbreeding coefficient: a global perspective. *Human heredity*, *77*(1-4), 37–48.

Peng, Y., Shi, H., Qi, X.-b., jie Xiao, C., Zhong, H., Ma, R.-l. Z., & Su, B. (2010). The ADH1B Arg47His polymorphism in east Asian populations and expansion of rice domestication in history. *BMC Evolutionary Biology*, *10*(15).

Pereira, R., Phillips, C., Pinto, N., Santos, C., dos Santos, S. E. B., Amorim, A., Carracedo, A., & Gusmão, L. (2012). Straightforward inference of ancestry and admixture proportions through ancestry-informative insertion deletion multiplexing. *PloS One*, *7*(1), e29684.

Perry, G. H., Dominy, N. J., Claw, K. G., Lee, A. S., Fiegler, H., Redon, R., Werner, J., Villanea, F., Mountain, J. L., Misra, R., Carter, N. P., Lee, C., & Stone, A. C. (2007). Diet and the evolution of human amylase gene copy number variation. *Nature Genetics*, *39*(10), 1256–1260.

Perusse, D. (1993). Cultural and reproductive success in industrial societies: Testing the relationship at the proximate and ultimate levels. *Behavioral and Brain Sciences*, *16*(2), 267–322.

Pettener, D., Pastor, S., & Tarazona-Santos, E. (1998). Surnames and genetic structure of a high-

altitude Quechua community from the Ichu River Valley, Peruvian Central Andes, 1825-1914. *Human Biology*, *70*(5), 865–887.

Peuhkuri, K., Poussa, T., & Korpela, R. (1998). Comparison of a portable breath hydrogen analyser (Micro H2) with a Quintron MicroLyzer in measuring lactose maldigestion, and the evaluation of a Micro H2 for diagnosing hypolactasia. *Scandinavian Journal of Clinical and Laboratory Investigation*, *58*(3), 217–24.

Peuhkuri, K., & Vapaatalo, H. (2000). Lactose intolerance—a confusing clinical diagnosis. *The American Journal of Clinical Nutrition*, *9*(1), 599–603.

Phillips, P. (2001). *Method and Theory in American Archaeology*. Tuscaloosa: University of Alabama Press.

Pier, G., Grout, M., & Zaidi, T. (1998). Salmonella typhi uses CFTR to enter intestinal epithelial cells. *Nature*, *393*(6680), 79–82.

Pinto-Cisternas, J., Pineda, L., & Barrai, I. (1985). Estimation of inbreeding by isonymy in Iberoamerican populations: an extension of the method of Crow and Mange. *American journal of human genetics*, *37*(2), 373–85.

Polanyi, K. (1944). *The Great Transformation: The Political and Economic Origins of Our Time*. Boston: Beacon Press.

Poolman, E. M., & Galvani, A. P. (2007). Evaluating candidate agents of selective pressure for cystic fibrosis. *Journal of the Royal Society Interface*, *4*(12), 91–98.

Porras-Hurtado, L., Ruiz, Y., Santos, C., Phillips, C., Carracedo, A., & Lareu, M. V. (2013). An overview of STRUCTURE: applications, parameter settings, and supporting software. *Frontiers in genetics*, *4*, 98.

Poulter, M., Hollox, E., Harvey, C. B., Mulcare, C., Peuhkuri, K., Kajander, K., Sarner, M., Korpela, R., Swallow, D. M., & C (2003). The causal element for the lactase persistence/non-persistence polymorphism is located in a 1 Mb region of linkage disequilibrium in Europeans. *Annals of Human Genetics*, *67*(4), 298–311.

Power, M. L., & Schulkin, J. (2008). Sex differences in fat storage, fat metabolism, and the health risks from obesity: possible evolutionary origins. *The British Journal of Nutrition*, *99*(5), 931–940.

Pretto, F., Silveira, T., Menegaz, V., & Oliveira, J. (2002). Lactose malabsorption in children and adolescents: diagnosis through breath hydrogen test using cow milk. *Jornal de Pediatria Rio de Janeiro*, *78*(3), 213–218.

Pritchard, J., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945–959.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. a. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–575.

Quantum GIS Development Team (2011). Quantum GIS Geographic Information System [Software]. Open Source Geospatial Foundation Project.
URL `http://scholar.google.com/scholar?q=Quantum+GIS#0`

Quinque, D., Kittler, R., Kayser, M., Stoneking, M., & Nasidze, I. (2006). Evaluation of saliva as a source of human DNA for population and association studies. *Analytical biochemistry*, *353*(2), 272–277.

R Core Development Team (2013). R: A Language and Environment for Statistical Computing. [Software]. R Foundation for Statistical Computing.
URL `http://www.r-project.org/`

Rajkumar, R., & Kashyap, V. K. (2004). Genetic structure of four socio-culturally diversified caste populations of southwest India and their affinity with related Indian and global groups. *BMC Genetics*, *5*(23).

Ranciaro, A., Campbell, M. C., Hirbo, J. B., Ko, W. Y., Froment, A., Anagnostou, P., Kotze, M. J., Ibrahim, M., Nyambo, T., Omar, S., & Tishkoff, S. (2014). Genetic Origins of Lactase Persistence and the Spread of Pastoralism in Africa. *American Journal of Human Genetics*, *94*(4), 496–510.

Rasinperä, H., Forsblom, C., Enattah, N. S., Halonen, P., Salo, K., Victorzon, M., Mecklin, J., Järvinen, H., Enholm, S., Sellick, G., Alazzouzi, H., Houlston, R., Robinson, J., Groop, P., Tomlinson, I., Schwartz, S., Aaltonen, L., & Järvelä, I. (2005). The C/C-13910 genotype of adult-type hypolactasia is associated with an increased risk of colorectal cancer in the Finnish population. *Gut*, *54*(5), 643–647.

Raymond, M., & Rousset, F. (1995). GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *Journal of Heredity*, *86*(3), 248–249.

Reher, D. (2012). Population And The Economy During The Demographic Transition. *Economic Affairs*, *32*(1), 10–16.

Reher, D. S. (2004). The demographic transition revisited as a global process. *Population, Space and Place*, *10*(1), 19–41.

Relethford, J. (1988). Estimation of kinship and genetic distance from surnames. *Human Biology*, *60*(3), 475–492.

Relethford, J. H. (2002). Apportionment of global human genetic diversity based on craniometrics and skin color. *American Journal of Physical Anthropology*, *118*(4), 393–398.

Revelle, W. (2014). *psych: Procedures for Psychological, Psychometric, and Personality Research. [Software]*. Northwestern University, Evanston, Illinois.
URL `http://cran.r-project.org/package=psych`

Reynolds, J., Weir, B. S., & Cockerham, C. C. (1983). Estimation of the coancestry coefficient: basis for a short-term genetic distance. *Genetics*, *105*(3), 767–779.

Ribando, C. (2007). Afro-Latinos in Latin America and considerations for US policy. Tech. rep., US Congressional Research Service.

Roberts, D. F. (1953). Body weight, race and climate. *American Journal of Physical Anthropology*, *11*(4), 533–558.

Rodríguez, J., Becker A., C., González C., P., Troncoso M., A., & Pavlovic B., D. (2004). La Cultura Diaguita en el Valle del Río Illapel. *Chungará (Arica)*, *36*(2), 739–751.

Rodriguez Diaz, R., Blanco Villegas, M., Díaz, R. R., & Villegas, M. B. (2010). Genetic structure of a rural region in Spain: Distribution of surnames and gene flow. *Human Biology*, *82*(3), 301–314.

Rodriguez-Larralde, A., Dipierri, J., Gomez, E. A., Scapoli, C., Mamolini, E., Salvatorelli, G., De Lorenzi, S., Carrieri, A., & Barrai, I. (2011). Surnames in bolivia: a study of the population of bolivia through isonymy. *American Journal of Physical Anthropology*, *144*(2), 177–184.

Rodríguez-Larralde, A., Formica, G., Scapoli, C., Beretta, M., Mamolini, E., & Barrai, I. (1993). Microevolution in Perugia: isonymy 1890-1990. *Annals of Human Biology*, *20*(3), 261–274.

Rodriguez-Larralde, A., Scapoli, C., Beretta, M., Nesti, C., Mamolini, E., & Barrai, I. (1998). Isonymy and the genetic structure of Switzerland. II. Isolation by distance. *Annals of Human Biology*, *25*(6), 533–540.

Rogers, A. (1991). Doubts about isonymy. *Human Biology*, *63*(5), 663–668.

Romeo, G., Devoto, M., & Galietta, L. (1989). Why is the cystic fibrosis gene so frequent? *Human Genetics*, *84*(95), 1–5.

Rosado, J., López, P., & Palma, M. (1994a). Mala digestión e intolerancia a la lactosa en adultos mexicanos. Importancia de evaluarlas con dosis habituales de leche. *Revista de Investigacion Clinica*, *46*(3), 203–208.

Rosado, J. L., Gonzalez, C., Valencia, M. E., López, P., Palma, M., López, B., Mejía, L., Báez, M. C., & Lopez, P. (1994b). Lactose maldigestion and milk intolerance: a study in rural and urban Mexico using physiological doses of milk. *The Journal of Nutrition*, *124*(7), 1052–1059.

Rosenberg, N. a. (2003). Distruct: a Program for the Graphical Display of Population Structure. *Molecular Ecology Notes*, *4*(1), 137–138.

Roullier, C., Benoit, L., McKey, D. B., & Lebot, V. (2013). Historical collections reveal patterns of diffusion of sweet potato in Oceania obscured by modern plant movements and recombination. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(6), 2205–2210.

Rousset, F. (2008). genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Molecular Ecology Resources*, *8*(1), 103–106.

Rubi-Castellanos, R., Anaya-Palafox, M., Mena-Rojas, E., Bautista-España, D., Muñoz Valle, J. F., & Rangel-Villalobos, H. (2009a). Genetic data of 15 autosomal STRs (Identifiler kit) of three Mexican Mestizo population samples from the States of Jalisco (West), Puebla (Center), and Yucatan (Southeast). *Forensic Science International. Genetics*, *3*(3), e71–e76.

Rubi-Castellanos, R., Martínez-Cortés, G., Muñoz Valle, J. F., González-Martín, A., Cerda-Flores, R. M., Anaya-Palafox, M., & Rangel-Villalobos, H. (2009b). Pre-Hispanic Mesoamerican demography approximates the present-day ancestry of Mestizos throughout the territory of Mexico. *American Journal of Physical Anthropology*, *139*(3), 284–294.

Ruff, C. (2002). Variation in human body size and shape. *Annual Review of Anthropology*, *31*(1), 211–232.

Sabeti, P., Reich, D., Higgins, J., Levine, H., Richter, D., Schaffner, S., Gabriel, S., Platko, J., Patterson, N., McDonald, G., & Others (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature*, *419*(6909), 832–837.

Sabeti, P. C., Walsh, E., Schaffner, S. F., Varilly, P., Fry, B., Hutcheson, H. B., Cullen, M., Mikkelsen, T. S., Roy, J., Patterson, N., Cooper, R., Reich, D., Altshuler, D., O'Brien, S., & Lander, E. S. (2005). The case for selection at CCR5-Delta32. *PLoS biology*, *3*(11), e378.

Sahi, T. (1974). The inheritance of selective adult-type lactose malabsorption. *Scandinavian Journal of Gastroenterology*, *30*(Supplement), 1–73.

Sahi, T. (1994). Genetics and Epidemiology of Adult-type Hypolactasia. *Scandinavian Journal of Gastroenterology*, *29*(s202), 7–20.

Sahi, T., & Launiala, K. (1977). More evidence for the recessive inheritance of selective adult type lactose malabsorption. *Gastroenterology*, *73*(2), 231–232.

Sahlins, M. D. (1972). *Stone Age Economics*. New Jersey: Transaction Publishers.

Sahoo, S., & Kashyap, V. K. (2005). Influence of language and ancestry on genetic structure of contiguous populations: a microsatellite based study on populations of Orissa. *BMC Genetics*, *6*(4).

Salo, P., Kääriäinen, H., Page, D. C., & de la Chapelle, A. (1995). Deletion mapping of stature determinants on the long arm of the Y chromosome. *Human Genetics*, *95*(3), 283–286.

Sanger, F., & Nicklen, S. (1977). DNA sequencing with chain-terminating. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(12), 5463–5467.

Santander, A. (1993). Contribucion al Estudio del Impacto de la Economía Minero-Cuprifera en el Desmonte o tala de la Vegetacion Arborea y Arbustiva, 1601-1900. Tech. rep., Dirección de Aguas, Illapel.

Schousboe, K., Willemsen, G., Kyvik, K. O., Mortensen, J., Boomsma, D. I., Cornes, B. K., Davis, C. J., Fagnani, C., Hjelmborg, J., Kaprio, J., De Lange, M., Luciano, M., Martin, N. G., Pedersen, N., Pietiläinen, K. H., Rissanen, A., Saarni, S., Sø rensen, T. I., Van Baal, G. C. M., & Harris, J. R. (2003). Sex differences in heritability of BMI: a comparative study of results from twin studies in eight countries. *Twin Research*, *6*(5), 409–421.

Sear, R. (2006). Size-dependent reproductive success in Gambian men: Does height or weight matter more? *Biodemography and Social Biology*, *53*(3-4), 172–188.

Ségurel, L., Lafosse, S., Heyer, E., & Vitalis, R. (2010). Frequency of the AGT Pro11Leu polymorphism in humans: Does diet matter? *Annals of Human Genetics*, *74*(1), 57–64.

Sevá-Pereira, A., Magalhães, A., & Pereira, R. (1983). Primary adult lactose malabsorption, a common genetic trait among Southeastern Brazilians. *Revista Brasileira de Genetica*, *6*(4), 717–759.

Sharpsteen, C., & Bracken, C. (2013). *tikzDevice: R Graphics Output in LaTeX Format. [Software]*. URL `http://cran.r-project.org/package=tikzDevice`

Shaw, R. R. (1960). An index of consanguinity based on the use of the surname in Spanish-speaking countries. *Journal of Heredity*, *51*(5), 1–2.

Shi, P., Zhang, J., Yang, H., & Zhang, Y. P. (2003). Adaptive diversification of bitter taste receptor genes in Mammalian evolution. *Molecular Biology and Evolution*, *20*(5), 805–814.

Shigemura, N., Shirosaki, S., Sanematsu, K., Yoshida, R., & Ninomiya, Y. (2009). Genetic and molecular basis of individual differences in human umami taste perception. *PloS One*, *4*(8), e6717.

Silventoinen, K. (2003). Determinants of variation in adult body height. *Journal of Biosocial Science*, *35*(2), 263–285.

Silventoinen, K., Lahelma, E., & Rahkonen, O. (1999). Social background, adult body-height and health. *International Journal of Epidemiology*, (28), 911–918.

Silventoinen, K., Sammalisto, S., Perola, M., Boomsma, D. I., Cornes, B. K., Davis, C., Dunkel, L., De Lange, M., Harris, J. R., Hjelmborg, J. V. B., Luciano, M., Martin, N. G., Mortensen, J., Nisticò, L., Pedersen, N. L., Skytthe, A., Spector, T. D., Stazi, M. A., Willemsen, G., & Kaprio, J. (2003). Heritability of adult body height: a comparative study of twin cohorts in eight countries. *Twin Research*, *6*(5), 399–408.

Simonson, T. S., Yang, Y., Huff, C. D., Yun, H., Qin, G., Witherspoon, D. J., Bai, Z., Lorenzo, F. R., Xing, J., Jorde, L. B., Prchal, J. T., & Ge, R. (2010). Genetic evidence for high-altitude adaptation in Tibet. *Science*, *329*(5987), 72–75.

Simoons, F. (1969). Primary adult lactose intolerance and the milking habit: a problem in biological and cultural interrelations. I Review of the Medical Research. *Digestive Diseases and Sciences*, *14*(12), 819–836.

Simoons, F. (1978). Lactose Malabsorption in Africa. *African Economic History*, *5*(5), 16–34.

Simoons, F. J. (1970). Primary adult lactose intolerance and the milking habit: a problem in biologic and cultural interrelations. II. A culture historical hypothesis. *The American Journal of Digestive Diseases*, *15*(8), 695–710.

Simoons, F. J. (1994). *Eat Not this Flesh: Food Avoidances from Prehistory to the Present*. Wisconsin: University of Wisconsin Press.

Smith, C., & Fretwell, S. (1974). The optimal balance between size and number of offspring. *American Naturalist*, *108*(962), 499–506.

Smith, G., Lawlor, D., Timpson, N., & Baban, J. (2008). Lactase persistence-related genetic variant: population substructure and health outcomes. *European Journal of Human Genetics*, *17*(3), 357–367.

Soranzo, N., Bufe, B., Sabeti, P. C., Wilson, J. F., Weale, M. E., Marguerie, R., Meyerhof, W., & Goldstein, D. B. (2005). Positive selection on a high-sensitivity allele of the human bitter-taste receptor TAS2R16. *Current Biology*, *15*(14), 1257–1265.

South, A. (2011). rworldmap: A New R package for Mapping Global Data. *The R Journal*, *3*(1), 35–43.

Sowers, M. F., & Winterfeldt, E. (1975). Lactose intolerance among Mexican Americans. *The American Journal of Clinical Nutrition*, *28*(7), 704–705.

Speakman, J. R. (2006). Thrifty genes for obesity and the metabolic syndrome–time to call off the search? *Diabetes & vascular disease research : official journal of the International Society of Diabetes and Vascular Disease*, *3*(1), 7–11.

Steckel, R. (1995). Stature and the Standard of Living. *Journal of Economic Literature*, *33*(4), 1903–1940.

Stephens, J. C., Reich, D. E., Goldstein, D. B., Shin, H. D., Smith, M. W., Carrington, M., Winkler, C., Huttley, G., Allikmets, R., Schriml, L., Gerrard, B., Malasky, M., Ramos, M. D., Morlot, S., Tzetis, M., Oddoux, C., di Giovine, F. S., Nasioulas, G., Chandler, D., Aseev, M., Hanson, M., Kalaydjieva, L., Glavac, D., Gasparini, P., Kanavakis, E., Claustres, M., Kambouris, M., Ostrer, H., Duff, G., Baranov, V., Sibul, H., Metspalu, A., Goldman, D., Martin, N., Duffy, D., Schmidtke, J., Estivill, X., O'Brien, S. J., & Dean, M. (1998). Dating the origin of the CCR5-Delta32 AIDS-resistance allele by the coalescence of haplotypes. *American Journal of Human Genetics*, *62*(6), 1507–1515.

Stephens, M., & Donnelly, P. (2003). A comparison of bayesian methods for haplotype reconstruction from population genotype data. *American Journal of Human Genetics*, *73*(5), 1162–1169.

Stephens, M., & Scheet, P. (2005). Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *American Journal of Human Genetics*, *76*(3), 449–462.

Stephens, M., Smith, N. J., & Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, *68*(4), 978–989.

Stinson, S. (1985). Sex differences in environmental sensitivity during growth and development. *American Journal of Physical Anthropology*, *28*(S6), 123–147.

Stock, J., & Migliano, A. (2009). Stature, mortality, and life history among indigenous populations of the Andaman Islands, 1871–1986. *Current Anthropology*, *50*(2003), 713–725.

Storey, A., Ramírez, J. M., Quiroz, D., Burley, D. V., Addison, D. J., Walter, R., Anderson, A. J., Hunt, T. L., Athens, J. S., Huynen, L., & Matisoo-Smith, E. (2007). Radiocarbon and DNA evidence for a pre-Columbian introduction of Polynesian chickens to Chile. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(25), 10335–10339.

Story, M., Evans, M., Fabsitz, R. R., Clay, T. E., Holy Rock, B., & Broussard, B. (1999). The epidemic

of obesity in American Indian communities and the need for childhood obesity-prevention programs. *The American Journal of Clinical Nutrition*, *69*(4 Suppl), 747S–754S.

Story, M., Stevens, J., Himes, J., Stone, E., Holy Rock, B., Ethelbah, B., & Davis, S. (2003). Obesity in American-Indian children: prevalence, consequences, and prevention. *Preventive Medicine*, *37*, S3–S12.

Strickland, S. (1993). Human nutrition in Mongolia: Maternal mortality and rickets. *Nomadic Peoples*, (33), 231–239.

Sulem, P., Gudbjartsson, D. F., Stacey, S. N., Helgason, A., Rafnar, T., Magnusson, K. P., Manolescu, A., Karason, A., Palsson, A., Thorleifsson, G., Jakobsdottir, M., Steinberg, S., Pálsson, S., Jonasson, F., Sigurgeirsson, B., Thorisdottir, K., Ragnarsson, R., Benediktsdottir, K. R., Aben, K. K., Kiemeney, L., Olafsson, J. H., Gulcher, J., Kong, A., Thorsteinsdottir, U., & Stefansson, K. (2007). Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nature Genetics*, *39*(12), 1443–1452.

Sverrisdóttir, O., Timpson, A., Toombs, J., Lecoeur, C., Froguel, P., Carretero, J., Arsuaga, J., Gotherstrom, A., & Thomas, M. (2014). Direct estimates of natural selection in Iberia indicate calcium absorption was not the only driver of lactase persistence in Europe. *Molecular Biology and Evolution*, *31*(4), 975–983.

Swagerty, D. L., Walling, A. D., Klein, R. M., & Swagery, D. (2002). Lactose intolerance. *American Family Physician*, *65*(9), 1845–1850.

Swallow, D. M. (2003). Genetics of lactase persistence and lactose intolerance. *Annual review of genetics*, *37*, 197–219.

Sykes, B., & Irven, C. (2000). Surnames and the Y chromosome. *American Journal of Human Genetics*, *66*(4), 1417–9.

Tan, J., Yang, Y., Tang, K., Sabeti, P. C., Jin, L., & Wang, S. (2013). The adaptive variant EDARV370A is associated with straight hair in East Asians. *Human genetics*, *132*(10), 1187–91.

Teves, P., Medina, J., Espinoza, Z., & EM (2001). Análisis de la prueba de tolerancia a la lactosa. *Revista Peruana de Gastroenterología*, *21*, 282 –286.

The 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, *491*(7422), 56–65.

Therneau, T. (2012). *coxme: Mixed Effects Cox Models. [Software]*.
URL `http://cran.r-project.org/package=coxme`

Thompson, F. E., & Subar, A. F. (2008). Dietary Assessment Methodology. In A. Coulston, C. Boushey, & M. Ferruzzi (Eds.) *Nutrition in the Prevention and Treatment of Disease*, chap. 1. London: Elsevier, 3rd ed.

Tishkoff, S., Varkonyi, R., Cahinhinan, N., Abbes, S., Argyropoulos, G., Destro-Bisol, G., Drousiotou, A., Dangerfield, B., Lefranc, G., Loiselet, J., Piro, A., Stoneking, M., Tagarelli, A., Tagarelli, G., Touma, E. H., Williams, S. M., & Clark, G. (2001). Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. *Science*, *293*(5529), 455–462.

Tishkoff, S. A., Reed, F. A., Ranciaro, A., Voight, B. F., Babbitt, C. C., Silverman, J. S., Powell, K., Mortensen, H. M., Hirbo, J. B., Osman, M., Ibrahim, M., Omar, S. A., Lema, G., Nyambo, T. B., Ghori, J., Bumpstead, S., Pritchard, J. K., Wray, G. A., & Deloukas, P. (2007). Convergent adaptation of human lactase persistence in Africa and Europe. *Nature Genetics*, *39*(1), 31–40.

Troelsen, J. T. (2005). Adult-type hypolactasia and regulation of lactase expression. *Biochimica et Biophysica Acta*, *1723*(1-3), 19–32.

Troncoso, A., & Pavlovic, D. (2013). Historia, saberes y prácticas: un ensayo sobre el desarrollo de las comunidades alfareras del Norte Semiárido Chileno. *Revista Chilena de Antropología*, *0*(27), 101–140.

Tserendolgor, U., Mawson, J., MacDonald, A., & Oyunbileg, M. (1998). Prevalence of rickets in Mongolia. *Asia Pacific Journal of Clinical Nutrition*, *7*(3/4), 325–328.

Tu, G. C., & Israel, Y. (1995). Alcohol consumption by orientals in North America is predicted largely by a single gene. *Behavior Genetics*, *25*(1), 59–65.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3), 1–67.

Vázquez, C., Escobar, H., Polanco, I., Codoceo, R., & Vitoria, J. C. (1975). [Malabsorption of carbohydrates in children]. *Anales Españoles de Pediatría*, *8*(2), 105–194.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S*. New York: Springer, 4th ed.

Vergara, R., Toro, H., Bonilla, D., & Meneses, J. (2005). *Población y Asentamientos Humanos en el Ámbito de las Comunidades Agrícolas - Región de Coquimbo*. Santiago: INE Instituto Nacional de Estadísticas.

Vettorazzi, C., Canales, D., Rosales, F., Barillas-Mury, C., Woert, J., Pineda, O., & Solomons, N.

(1992). Milk, lactose and ethanol as dietary factors in cataractogenesis in Guatemala. A case-control study. *International Journal of Food Sciences and Nutrition*, *43*(3), 155–162.

Voight, B. F., & Pritchard, J. K. (2005). Confounding from cryptic relatedness in case-control association studies. *PLoS Genetics*, *1*(3), e32.

Wang, S., Lewis, C. M., Jakobsson, M., Ramachandran, S., Ray, N., Bedoya, G., Rojas, W., Parra, M. V., Molina, J. a., Gallo, C., Mazzotti, G., Poletti, G., Hill, K., Hurtado, A. M., Labuda, D., Klitz, W., Barrantes, R., Bortolini, M. C., Salzano, F. M., Petzl-Erler, M. L., Tsuneto, L. T., Llop, E., Rothhammer, F., Excoffier, L., Feldman, M. W., Rosenberg, N. a., & Ruiz-Linares, A. (2007). Genetic variation and population structure in native Americans. *PLoS Genetics*, *3*(11), e185.

Wang, X., Thomas, S. D., & Zhang, J. (2004). Relaxation of selective constraint and loss of function in the evolution of human bitter taste receptor genes. *Human Molecular Genetics*, *13*(21), 2671–2678.

Warnes, G. R., Bolker, B., Gorjanc, G., Grothendieck, G., Korosec, A., Lumley, T., MacQueen, D., Magnusson, A., & Rogers, J. (2014). *gdata: Various R programming tools for data manipulation. [Software]*.
URL `http://cran.r-project.org/package=gdata`

Weatherall, D. J. (2008). Genetic variation and susceptibility to infection: the red cell and malaria. *British Journal of Haematology*, *141*(3), 276–286.

Wei, T. (2013). *corrplot: Visualization of a correlation matrix. [Software]*.
URL `http://cran.r-project.org/package=corrplot`

Weinberg, W. (1909). Über Vererbungsgesetze beim Menschen. *Zeitschrift für Induktive Abstammungs- und Vererbungslehre*, *2*(1), 276–330.

Weiss, K. M., Ferrell, R. F., & Hanis, C. L. (1984). A new world syndrome of metabolic diseases with a genetic and evolutionary basis. *American Journal of Physical Anthropology*, *27*(S5), 153–178.

Weiss, V. (1980). Inbreeding and genetic distance between hierarchically structured populations measured by surname frequencies. *Mankind Quarterly*, *21*, 135–149.

Wells, J. C. K. (2002). Thermal environment and human birth weight. *Journal of Theoretical Biology*, *214*(3), 413–425.

Weng, H., Bastian, L., Taylor, D., Moser, B., & Ostbye, T. (2004). Number of children associated with obesity in middle-aged women and men: results from the Health and Retirement Study. *Journal of Women's Health*, *13*(1), 85–91.

Westreicher, C., Mérega, J., & Palmili, G. (2006). Review of the literature on Pastoral Economics and Marketing: South America. Tech. rep., World Initiative for Sustainable Pastoralism.

Westreicher, C. A., Mérega, J. L., & Palmili, G. (2007). The Economics of Pastoralism: Study on Current Practices in South America. *Nomadic Peoples*, *11*(2), 87–105.

Wheeler, J. (1995). Evolution and present situation of the South American Camelidae. *Biological Journal of the Linnean Society*, *54*(3), 271–295.

Wickham, & Hadley (2007). Reshaping data with the reshape package. *Journal of Statistical Software*, *21*(12).

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. New York: Springer New York, 1st ed.

Wilson, C. (2011). Understanding global demographic convergence since 1950. *Population and Development Review*, *37*(2), 375–388.

Wiuf, C. (2001). Do delta F508 heterozygotes have a selective advantage? *Genetical research*, *78*(1), 41–47.

Wood, A., White, I., & Royston, P. (2008). How should variable selection be performed with multiply imputed data? *Statistics in medicine*, *27*(17), 3227–3246.

Wooding, S. (2006). Phenylthiocarbamide: a 75-year adventure in genetics and natural selection. *Genetics*, *168*(3), 1097–1104.

Wooding, S., Bufe, B., Grassi, C., Howard, M. T., Stone, A. C., Vazquez, M., Dunn, D. M., Meyerhof, W., Weiss, R. B., & Bamshad, M. J. (2006). Independent evolution of bitter-taste sensitivity in humans and chimpanzees. *Nature*, *440*(7086), 930–934.

Wooding, S., Kim, U.-K., Bamshad, M. J., Larsen, J., Jorde, L. B., & Drayna, D. (2004). Natural selection and molecular evolution in PTC, a bitter-taste receptor gene. *American Journal of Human Genetics*, *74*(4), 637–646.

Woteki, C., Weser, E., & Young, E. (1977). Lactose malabsorption in Mexican-American adults. *The American Journal of Clinical Nutrition*, *30*(4), 470–475.

Wright, K. (2013). *corrgram: Plot a correlogram. [Software]*.
URL `http://cran.r-project.org/package=corrgram`

Wright, S. (1922). Coefficients of inbreeding and relationship. *The American naturalist*, *56*(645), 330–338.

Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding and selection in evolution. In *Proceedings of the 6th International Congress of Genetics*, (pp. 356–366). Genetics Society of America.

Wright, S. (1949). The genetical structure of populations. *Annals of Eugenics*, *15*(1), 323–353.

Wright, S. (1965). The interpretation of population structure by F-statistics with special regard to systems of mating. *Evolution*, *19*(3), 395–420.

Yanez, A. P., Angulo, J. P., & Fernandez, C. J. (1971). Malabsorción de lactosa en estudiantes espanoles. I. Toleranzia intestinal a la sobrecarga oral de lactosa. *Revista Española de las Enfermedades del Aparato Digestivo*, *35*, 925–938.

Yasuda, N., & Furusho, T. (1971). Random and nonrandom inbreeding revealed from isonymy study. I. Small cities of Japan. *American Journal of Human Genetics*, *23*(3), 303–316.

Young, G., Zavala, H., Wandel, J., Smit, B., Salas, S., Jimenez, E., Fiebig, M., Espinoza, R., Diaz, H., & Cepeda, J. (2009). Vulnerability and adaptation in a dryland community of the Elqui Valley, Chile. *Climatic Change*, *98*(1-2), 245–276.

Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S., & Buckler, E. S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, *38*(2), 203–208.

Zeileis, A., Kleiber, C., & Jackman, S. (2008). Regression models for count data in R. *Journal of Statistical Software*, *27*(8), 2–25.

Zhao, K., Aranzana, M. J., Kim, S., Lister, C., Shindo, C., Tang, C., Toomajian, C., Zheng, H., Dean, C., Marjoram, P., & Nordborg, M. (2007). An Arabidopsis example of association mapping in structured samples. *PLoS Genetics*, *3*(1), e4.

Appendix A

Copyright clearance

This Appendix contains written permission from the copyright holders to reproduce the contents in Figure 1.1 (Adapted from Tishkoff et al. 2007 by permission from Macmillan Publishers Ltd: Nature Genetics copyright ©2007 Nature Publishing Group.), and Table 1.1 (Adapted from Simoons 1978 by permission from Springer Science and Business Media: The American Journal of Digestive Diseases copyright ©1978 Springer).

NATURE PUBLISHING GROUP LICENSE TERMS AND CONDITIONS

Dec 08, 2014

This is a License Agreement between Nicolas Montalva ("You") and Nature Publishing Group ("Nature Publishing Group") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Nature Publishing Group, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	3337080694869
License date	Feb 27, 2014
Order Content Publisher	Nature Publishing Group
Order Content Publication	Nature Genetics
Order Content Title	Convergent adaptation of human lactase persistence in Africa and Europe
Order Content Author	Sarah A Tishkoff, Floyd A Reed, Alessia Ranciaro, Benjamin F Voight, Courtney C Babbitt, Jesse S Silverman, Kweli Powell, Holly M Mortensen, Jibril B Hirbo, Maha Osman, Muntaser Ibrahim, Sabah A Omar, Godfrey Lema, Thomas B Nyambo, Jilur Ghori, Suzannah Bumpstead, Jonathan K Pritchard, Gregory A Wray, Panos Deloukas
Order Content Date	Dec 10, 2006
Volume number	39
Issue number	1
Type of Use	reuse in a dissertation / thesis
Requestor type	academic/educational
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables /illustrations	1
High-res required	no
Figures	Figure 1
Author of this NPG article	no
Your reference number	None
Title of your thesis / dissertation	Adaptation to milk drinking and evolution of lactase persistence in pastoralist goat herders in central-northern Chile
Expected completion date	Sep 2014
Estimated size (number of pages)	200
Total	**0.00 GBP**
Terms and Conditions	

Terms and Conditions for Permissions

Nature Publishing Group hereby grants you a non-exclusive license to reproduce this material for this purpose, and for

no other use,subject to the conditions below:

1. NPG warrants that it has, to the best of its knowledge, the rights to license reuse of this material. However, you should ensure that the material you are requesting is original to Nature Publishing Group and does not carry the copyright of another entity (as credited in the published version). If the credit line on any part of the material you have requested indicates that it was reprinted or adapted by NPG with permission from another source, then you should also seek permission from that source to reuse the material.

2. Permission granted free of charge for material in print is also usually granted for any electronic version of that work, provided that the material is incidental to the work as a whole and that the electronic version is essentially equivalent to, or substitutes for, the print version.Where print permission has been granted for a fee, separate permission must be obtained for any additional, electronic re-use (unless, as in the case of a full paper, this has already been accounted for during your initial request in the calculation of a print run).NB: In all cases, web-based use of full-text articles must be authorized separately through the 'Use on a Web Site' option when requesting permission.

3. Permission granted for a first edition does not apply to second and subsequent editions and for editions in other languages (except for signatories to the STM Permissions Guidelines, or where the first edition permission was granted for free).

4. Nature Publishing Group's permission must be acknowledged next to the figure, table or abstract in print. In electronic form, this acknowledgement must be visible at the same time as the figure/table/abstract, and must be hyperlinked to the journal's homepage.

5. The credit line should read:
   Reprinted by permission from Macmillan Publishers Ltd: [JOURNAL NAME] (reference citation), copyright (year of publication)
   For AOP papers, the credit line should read:
   Reprinted by permission from Macmillan Publishers Ltd: [JOURNAL NAME], advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].XXXXX)

   Note: For republication from the *British Journal of Cancer,* the following credit lines apply.
   Reprinted by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME] (reference citation), copyright (year of publication)For AOP papers, the credit line should read:
   Reprinted by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME], advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].XXXXX)

6. Adaptations of single figures do not require NPG approval. However, the adaptation should be credited as follows:

   Adapted by permission from Macmillan Publishers Ltd: [JOURNAL NAME] (reference citation), copyright (year of publication)

   Note: For adaptation from the *British Journal of Cancer,* the following credit line applies.
   Adapted by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME] (reference citation), copyright (year of publication)

7. Translations of 401 words up to a whole article require NPG approval. Please visit http://www.macmillanmedicalcommunications.com for more information.Translations of up to a 400 words do not require NPG approval. The translation should be credited as follows:

   Translated by permission from Macmillan Publishers Ltd: [JOURNAL NAME] (reference citation), copyright (year

of publication).

Note: For translation from the *British Journal of Cancer,* the following credit line applies.
Translated by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME] (reference citation), copyright (year of publication)

We are certain that all parties will benefit from this agreement and wish you the best in the use of this material. Thank you.
Special Terms:
v1.1

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

Gratis licenses (referencing $0 in the Total field) are free. Please retain this printable license for your reference. No payment is required.

SPRINGER LICENSE
TERMS AND CONDITIONS

Dec 08, 2014

This is a License Agreement between Nicolas Montalva ("You") and Springer ("Springer") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Springer, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	3337260056725
License date	Feb 27, 2014
Order Content Publisher	Springer
Order Content Publication	American Journal of Digestive Diseases
Order Content Title	The geographic hypothesis and lactose malabsorption
Order Content Author	Frederick J. Simoons PhD
Order Content Date	Jan 1, 1978
Volume number	23
Issue number	11
Type of Use	Thesis/Dissertation
Portion	Figures
Author of this Springer article	No
Order reference number	None
Original figure numbers	Table 1
Title of your thesis / dissertation	Adaptation to milk drinking and evolution of lactase persistence in pastoralist goat herders in central-northern Chile
Expected completion date	Sep 2014
Estimated size(pages)	200
Total	**0.00 GBP**

Terms and Conditions

Introduction

The publisher for this copyrighted material is Springer Science + Business Media. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at http://myaccount.copyright.com).

Limited License

With reference to your request to reprint in your thesis material on which Springer Science and Business Media control the copyright, permission is granted, free of charge, for the use indicated in your enquiry.

Licenses are for one-time use only with a maximum distribution equal to the number that you identified in the licensing process.

This License includes use in an electronic form, provided its password protected or on the university's intranet or repository, including UMI (according to the definition at the Sherpa website: http://www.sherpa.ac.uk/romeo/). For any

other electronic use, please contact Springer at (permissions.dordrecht@springer.com or permissions.heidelberg@springer.com).
The material can only be used for the purpose of defending your thesis limited to university-use only. If the thesis is going to be published, permission needs to be re-obtained (selecting "book/textbook" as the type of use).
Although Springer holds copyright to the material and is entitled to negotiate on rights, this license is only valid, subject to a courtesy information to the author (address is given with the article/chapter) and provided it concerns original material which does not carry references to other sources (if material in question appears with credit to another source, authorization from that source is required as well).
Permission free of charge on this occasion does not prejudice any rights we might have to charge for reproduction of our copyrighted material in the future.
Altering/Modifying Material: Not Permitted
You may not alter or modify the material in any manner. Abbreviations, additions, deletions and/or any other alterations shall be made only with prior written authorization of the author(s) and/or Springer Science + Business Media. (Please contact Springer at (permissions.dordrecht@springer.com or permissions.heidelberg@springer.com)
Reservation of Rights
Springer Science + Business Media reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.
Copyright Notice:Disclaimer
You must include the following copyright and permission notice in connection with any reproduction of the licensed material: "Springer and the original publisher /journal title, volume, year of publication, page, chapter/article title, name(s) of author(s), figure number(s), original copyright notice) is given to the publication in which the material was originally published, by adding; with kind permission from Springer Science and Business Media"
Warranties: None
Example 1: Springer Science + Business Media makes no representations or warranties with respect to the licensed material.
Example 2: Springer Science + Business Media makes no representations or warranties with respect to the licensed material and adopts on its own behalf the limitations and disclaimers established by CCC on its behalf in its Billing and Payment terms and conditions for this licensing transaction.
Indemnity
You hereby indemnify and agree to hold harmless Springer Science + Business Media and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.
No Transfer of License
This license is personal to you and may not be sublicensed, assigned, or transferred by you to any other person without Springer Science + Business Media's written permission.
No Amendment Except in Writing
This license may not be amended except in a writing signed by both parties (or, in the case of Springer Science + Business Media, by CCC on Springer Science + Business Media's behalf).
Objection to Contrary Terms
Springer Science + Business Media hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment terms and conditions (which are incorporated herein), comprise the entire agreement between you and Springer Science + Business Media (and CCC) concerning this licensing transaction. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall control.
Jurisdiction
All disputes that may arise in connection with this present License, or the breach thereof, shall be settled exclusively by arbitration, to be held in The Netherlands, in accordance with Dutch law, and to be conducted under the Rules of the 'Netherlands Arbitrage Instituut' (Netherlands Institute of Arbitration).***OR:***
All disputes that may arise in connection with this present License, or the breach thereof, shall be settled exclusively by arbitration, to be held in the Federal Republic of Germany, in accordance with German law.
Other terms and conditions:

v1.3

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

Gratis licenses (referencing $0 in the Total field) are free. Please retain this printable license for your reference. No payment is required.

Appendix B

List of R packages

This Appendix contains a list of the R packages mentioned through the thesis. The list is grouped by categories, with references in brackets when a specific citation format is suggested by the package documentation according to R citation utilities.

- **Correspondence and Principal Component Analyses:**
  - "`ca`" (Nenadic & Greenacre, 2007).
  - "`FactoMineR`" (Husson et al., 2013).
  - "`pcaPP`" (Filzmoser et al., 2013).
- **Data management and miscellaneous summary tools:**
  - "`gdata`" (Warnes et al., 2014).
  - "`Hmisc`" (Harrell, Frank & Dupont, 2014).
  - "`psych`" (Revelle, 2014).
  - "`reshape`" (Wickham & Hadley, 2007).
- **Geography, mapping and spatial analysis:**
  - "`fields`" (Nychka et al., 2013).
  - "`mapdata`" (Becker et al., 2013).
  - "`maps`" (Becker et al., 2013).
  - "`maptools`" (Bivand & Lewin-Koh, 2014).

  - "raster" (Hijmans, 2014).
  - "rgeos" (Bivand & Rundel, 2014).
  - "rgdal" (Bivand et al., 2014).
  - "rworldmap" (South, 2011).
  - "sp" (Bivand et al., 2013; Pebesma & Bivand, 2005).
  - "spatstat" (Baddeley & Turner, 2005).
- **Graphics and plotting tools:**
  - "ColorRamps" (Keitt, 2012).
  - "ggplot2" (Wickham, 2009).
  - "HH" (Heiberger, 2014).
  - "lattice" (Wickham & Hadley, 2007).
  - "RColorBrewer" (Neuwirth, 2011).
  - "rgl" (Adler & Murdoch, 2014).
  - "tikzDevice" (Sharpsteen & Bracken, 2013).
- **Multiple imputation:**
  - "Amelia" (Honaker et al., 2011).
  - "mice" (van Buuren & Groothuis-Oudshoorn, 2011).
- **Population genetics:**
  - "ade4" (Dray & Dufour, 2007).
  - "adegenet" (Jombart, 2008; Jombart & Ahmed, 2011).
  - "ape" (Paradis et al., 2004).
  - "gstudio" (Dyer, 2014).
- **Regression and model fitting:**
  - "car" (Fox & Weisberg, 2011).
  - "coxme" (Therneau, 2012).
  - "effects" (Fox & Hong, 2009).
  - "leaps" (Lumley & Miller, 2009).

  - "`MASS`" (Venables & Ripley, 2002).
  - "`pscl`" (Zeileis et al., 2008).

- **Surnames analyses:**
  - "`Biodem`" (Boattini & Calboli, 2012).ontent...

- **Graphic matrices and correlograms:**
  - "`corrgram`" (Wright, 2013).
  - "`corrplot`" (Wei, 2013).

Appendix C

Questionnaires

Appendix: *Questionnaires*

Questionnaire: Genetic and Cultural Adaptation to Milk Consumption in Goat Herders from the Semiarid Region of Coquimbo, Chile

Instructions for the survey taker before starting:

Thank the interviewee for his/her Participation
Make clear that the information will be used only in scientific publications and without giving any details about the identity of the interviewee.
Explain the topics included in the questionnaire: Family history, dietary habits, and livelihood in the communities.
Make clear that the interviewee can decline and be removed of this research whenever he/she wants, until 3 months of our visit.

Example: "Thank you for helping us with this questionnaire. Before proceeding I would like to remind you that your information will be used in scientific publications only, without providing your personal information. The questionnaire I will ask you to fill in includes information about your family and its history, the daily life in the Agricultural Communities, and your diet and health status. I would also like to remind you that you can decline to collaborate with us at any time, until your information has been anonymised, 3 months after our visit."

Personal Details (This section will be removed of the questionnaire after collate your data. You name will be not recorded in our databases):

DATE: _____/_____/_____ TIME: _____:_____

NAME:
_____

CODE: ____ _____

PARTNER INCLUDED?:
○ No
○ Yes → Code _____

OTHER RELATIVES INCLUDED?:
○ No
○ Yes → Codes?
_____ _____ _____ _____ _____

I. Personal Details:

1. Code : ____ _____
2. Birth Date : ____/____/____
3. Sex : ○ Male → ***go to question 4***
○ Female

3.1. When did you have your first menstrual cycle (menarche)?

Actual age _____ years. Or range:

○ 10 years or younger
○ 11-12 years old
○ 13-14 years old
○ 15 years or older
○ Never/Not yet

3.2. When did you have your last menstrual cycle (menopause)?

Actual age _____ years. Or range:

○ 40 years or younger
○ 40-50 years old
○ 50-60 years old
○ 60 years or older
○ Never/Not yet
○ N/A (e.g. uterectomy)

4. Community of Residence:
_____

5. How long have you been living in this community?

_____ years.

6. Where were you living when you were born? (Not birth place, but residency at birth).

○ This community.
○ Other Agricultural Community. Which?

_____
○ Outside. Where?

_____

7. Have you ever had children?

○ No → go to question 9
○ Yes

8. How many children did you have? Please include deceased children.

_____ Children:
_____ Living males.
_____ Deceased males.
_____ Living females.
_____ Deceased females.

II. About your children

The following questions are related to your children and their milk consumption habits. We will start from the first to be born all the way through to your last born child.

Survey taker: Be sure to fill a "Child Questionnaire" for each child.
□ Check when all the "Child Questionnaires" have been filled.

III. About your father

9. Where did he live when he was born? (Not birth place, but residency at birth)

○ This community.
○ Other Agricultural Community. Which?
_____
○ Outside. Where?
_____

10. Father's father (paternal grandfather) residency at birth:

○ This community.
○ Other Agricultural Community. Which?
_____
○ Outside. Where?
_____

11. Father's mother (paternal grandmother) residency at birth:

○ This community.
○ Other Agricultural Community. Which?
_____
○ Outside. Where?
_____

12. Does/Did your father feel unwell after consuming milk or milk products?

○ No
○ Yes, Please provide details on symptoms and products.
_____

IV. About your mother

13. Where did she live when she was born? (Not birth place, but residency at birth)

○ This community.
○ Other Agricultural Community. Which?
_____
○ Outside. Where?
_____

14. Mother's father (maternal grandfather) residency at birth:

○ This community.
○ Other Agricultural Community. Which?
_____
○ Outside. Where?
_____

15. Mother's mother (maternal grandmother) residency at birth:

○ This community.
○ Other Agricultural Community. Which?
_____
○ Outside. Where?
_____

16. Does/Did your mother feel unwell after having milk or milk products?

○ No
○ Yes, Please provide details on symptoms and products.
_____

17. Check if you or your households have access to the following goods/services:

□ Tap Water
□ Electricity
□ Ceiling
□ Floor
□ Sewerage
□ Hot water
□ Mobile phone
□ Washing machine
□ Fridge
□ Television
□ Computer
□ Car
□ Commoner rights
□ Livestock/Cattle
□ Irrigated plot (hijuela)
□ Rain-fed plot (lluvia)

18. Regarding the conditions of your residence, you are:

○ Living in your own house / the house of your partner, parents, or children.
○ Living as a guest
○ Living as Lodger/Tenant

19. Do you drink milk? (Check NO for lactose-free, or substitutes such as rice or soya milk)

○ No → ***go to question 23***
○ Yes

20. Indicate your regular consumption of milk during a year in summer and winter.

Summer	**Winter**
○	○ Never
○	○ Less than once per month
○	○ 1-3 glasses per month
○	○ 1 glass per week
○	○ 2-4 glasses per week
○	○ 5-6 glasses per week
○	○ 1 glass per day
○	○ 2-3 glasses per day
○	○ 4 or more glasses per day

21. Did you drink milk the last week? (last 7 days)

○ No → ***go to question 24***
○ Yes, How many days? ___ days

22. In average, how many glasses did you drink per day? _____ glasses

23. Did you drink milk yesterday?

○ **No → *go to question 24***
○ **Yes, How much? ___ glasses/cups.**

24. Do you eat dairy products?

○ No → ***go to question 26***
○ Yes, Please check which:

□ Fresh Cheese
□ Mature Cheese
□ Butter (not margarine)
□ Yoghurt
□ Evaporated/Condensed/Dulce de Leche
□ Other _____

24. Indicate your regular consumption of milk products during a year.

Summer	**Winter**
○	○ Never
○	○ Less than once per month
○	○ 1-3 servings per month
○	○ 1 serving per week
○	○ 2-4 servings per week
○	○ 5-6 servings per week
○	○ 1 serving per day
○	○ 2-3 servings per day
○	○ 4 or more servings per day

25. Did you eat any of these products within the last week? (last 7 days)

○ No
○ Yes, How many days? ___ days

26. Did you eat any of these products Yesterday?

○ No
○ Yes, How much? ___ servings

26. Do you feel unwell after eating certain specific foods?

○ No
○ Yes, Please check which:
□ Grains (e.g. corn, rice)
□ Wheat (e.g. bread)
□ Legumes (e.g. beans, lentils)
□ Seasoning (e.g. garlic, onion, pepper)
□ Green leafy vegetables (e.g. cabbage, celery)
□ High glycemic fruits (e.g. tangerines, grapes, papaya)
□ Sugars (e.g. sweets, chocolate, cakes)
□ Alcoholic beverages
□ Fats and fried food
□ Milk and/or milk products
□ Other
_____
Please provide details on symptoms:

_____

27. Do you feel unwell after drinking milk or eating dairy products?

○ No → ***end of the questionnaire***

○ Yes, please check symptoms below:

	Fresh Milk	Milk prod.
Belching	□	□
Reflux	□	□
Stomach ache	□	□
Constipation	□	□
Diarrhoea	□	□
Flatulence	□	□
Nausea	□	□
Vomiting	□	□
Bloating	□	□
Cramps	□	□
Wind	□	□
Other_____	□	□

→ ***end of the questionnaire***

Questionnaire: Genetic and Cultural Adaptation to Milk Consumption in Goat Herders from the Semiarid Region of Coquimbo, Chile

CHILD QUESTIONNAIRE

Instructions for the survey taker:

A new questionnaire form should be used for each additional child in the main questionnaire. Please, make sure you link each child's questionnaire with the code corresponding to the participant.

Personal Details (This section will be removed from the questionnaire after we collate your data. You name will be not recorded in our databases):

PARTICIPANT'S NAME: _____

PARTICIPANT'S CODE: _____ _____

CHILD'S NAME: _____

CHILD BIRTH RANK (1st, 2nd, 3rd, etc.):_____

CHILD'S CODE: ____ _____ ____

✂----------

1. Main questionnaire Code: ____ _____
2. Child code : ____ _____ ____
3. Birth Date : ____/____/____

4. Sex :
○ Male
○ Female

5. Birth Rank (1st, 2nd...) : _____

6. Is he/she alive?

○ Yes → ***go to question 9***
○ No

7. Death date : ____/____/____
8. Cause of Death : _____

9. Where was he/she living when he/she was born? (Not birth place, but residency at birth).

○ This community.
○ Other Agricultural Community. Which? _____
○ Outside. Where? _____

10. ***(Survey taker: Skip this question for the first born child)*** **Is this child half-sibling of any of the children reported before?**

○ No → ***go to question 12***
○ Yes, Identify half-sibling by their Birth ranks:

_____

11. Where was the other parent of at his/her birth?

○ This community.
○ Other Agricultural Community. Which? _____
○ Outside. Where? _____

12. Are you breastfeeding or Did you breastfeed this child?

○ Yes, when baby.
○ Yes, she/he is still been Breastfed.
→ ***go to question 14***
○ No.

13. How old was he/she when weaning?

_____ months old (0 if was not breastfed)

14. Did you feed him/her with other milks during and/or after weaning?

○ No
○ Yes, How old was he/she? _____

15. Does he/she drink milk now?

○ No
○ Yes

16. Does he/she eat dairy products?

○ No
○ Yes

17. Have you noticed that he/she feels unwell after eating some specific foods?

○ No
○ Yes, Please check which:

□ Grains (e.g. corn, rice)
□ Wheat (e.g. bread)
□ Legumes (e.g. beans, lentils)
□ Seasoning (e.g. garlic, onion, pepper)
□ Green leafy vegetables (e.g. cabbage,celery)
□ High glycemic fruits (e.g. tangerines, grapes, papaya)
□ Sugars (e.g. sweets, chocolate, cakes)
□ Alcoholic beverages
□ Fats and fried food
□ Milk and/or milk products
□ Other _____

Please provide details on symptoms:
_____

18. Does he/she feel unwell after drinking milk or eating dairy products?

○ No
○ Yes, please check symptoms below:

	Fresh Milk	Milk Products
Belching	□	□
Reflux	□	□
Stomach ache	□	□
Constipation	□	□
Diarrhoea	□	□
Flatulences	□	□
Nausea	□	□
Vomiting	□	□
Bloating	□	□
Cramps	□	□
Winds	□	□
Other_____	□	□

19. Does your child have children? (Your grandchildren)

○ No
○ Yes, please give details below:

_____ Grandchildren:
_____ Living males.
_____ Deceased males.
_____ Living females.
_____ Deceased females.

→ End of the "Child Questionnaire". Fill another copy of this questionnaire for the next child, or continue with the main questionnaire if this was the last born child

Appendix D

Information sheet and informed consent form

UCL DEPARTMENT OF ANTHROPOLOGY
University College London, 14 Taviton Street, London WC1H 0BW, UK
Tel +44 (0)20 7679 8633.

Information Sheet for Adult participants from the Agricultural Communities in Research Studies

You will be given a copy of this information sheet.

Title of Project: **Genetic and Cultural Adaptation to Milk Consumption in Goat Herders from the Semiarid Region of Coquimbo, Chile.**

This study has been approved by the UCL Research Ethics Committee
(Project ID Number): **2967/001**

This study has been approved by the University of Chile Social Sciences Research Ethics Committee (CEDEA)
(Project ID Number): **078/2010**

Name	Nicolás Montalva Rivera
Work Address	UCL Departament of Anthropology 14 Taviton Street London WC1 0BW Reino Unido
Contact Details	**In UK:** UCL Departament of Anthropology 14 Taviton Street London WC1 0BW United Kingdom Tel: +44 (0) 20 7679 8633 email: nicolas.montalva.09@ucl.ac.uk **In Chile:** Universidad de Chile Departamento de Antropología Laboratorio de Antropología Física Av. Ignacio Carrera Pinto 1045 Ñuñoa Santiago de Chile Tel: +56 (2) 9787758

UCL DEPARTMENT OF ANTHROPOLOGY
University College London, 14 Taviton Street, London WC1H 0BW, UK
Tel +44 (0)20 7679 8633.

We would like to invite you to participate in this research project.

Details of the Study:

We have invited you to take part in a scientific study that aims to know the relation between consumption of goat's milk and lactose tolerance in your community. We can know if a person have tolerance to lactose (natural sugar in milk) through a DNA analysis using cells from your mouth. To do this analysis, we will ask you to perform various procedures described below, that will allow us to know if you are lactose tolerant. In addition, to know your dietary habits and your family history, we will ask you to respond a questionnaire with questions on these topics. At last, we will measure your weight and heigh.

There are not direct benefits for you if you decide to take part on this study, but you will be able to know the results of the DNA test if you want. Your collaboration will allow us to improve our knowledge about lactose tolerance, dietary habits, and goat breeding practices at this community. Nor is there any danger in take part of the sample collection. There is no payment nor monetary cost to you for participating in this study.

Specifically, we are requesting the participation of men and women over 18 years old, from the Agricultural Communities of the Coquimbo Region.

As participant, you will be gathered in an appropriate, comfortable, clean and safe space. The procedures will be taken simultaneously, and will take between 20 and 50 minutes. You will be addresses by a member of the research staff, who will start confirming your age and origin. Then, we will ask for your participation in the following procedures:

1.- Collection of buccal cells samples.

To take samples of cells from your mouth, we will ask you to rub 10 cotton swabs the inner side of your cheek, 20 seconds each. We will place each swab in a plastic tube with a storage liquid.

2.- Questionnaires.

We will ask you to complete a survey about your family history, your reproductive history, and about your dietary habits in relation to milk and dairy products.

3.- Measurements of weight and heigh.

We will measure your weight and heigh using a common scale and stadiometer.

UCL DEPARTMENT OF ANTHROPOLOGY
University College London, 14 Taviton Street, London WC1H 0BW, UK
Tel +44 (0)20 7679 8633.

Safeguards and Risks.

• **Information:** We will inform you all the aims and characteristics of this research in full and in a non-techinical language. In addition, you will have chances to ask whatever you wish.

• **Anonymity:** Your name will not appear in any communication or report of this project. We will handle of your information in a confidential way. All the measurement, samples, questionnaires and interviews related with each participant will be identified with a code. Our records we will not have your name, and your individuals details will be only known by the research staff and by you, in case you want to know the results.

• **Free:** This study is free, and does not have any cost for you nor for your health provider.

• **Not Remunerated:** This study does not have any economic benefit or payment for you.

• **Voluntary:** You are not obliged to take part in this study in any way, and your participation is completely up to you. You will not get any payment if you decide to take part. You have the right to withdraw your participation at anytime (during our visit, or even after the data collection) without giving any reason.

• **Risks:** These procedures do not represent risks for you. However, if you are damaged as consecuence of any fortuitous accident during your participation in this study, we will cover the cost of any medical requirement that you may incur in the corresponding health service according to your health provider.

• **Biological Samples (DNA):** We will use the samples from your mouth, and the genetic material extracted from them, only with the above mentioned purposes. We will destroy all the biological material once the study has been concluded. The materials will be destroyed according to the protocols used at the laboratories of Biology at UCL, inactivating all the biological material by autoclave sterilization.

We will disclose the obtained data only in a general and not specific way, and only in scientific publications. We will not use any data with other purposes than the explained above. All the data will be confidential, and we will never mention your name

UCL DEPARTMENT OF ANTHROPOLOGY
University College London, 14 Taviton Street, London WC1H 0BW, UK
Tel +44 (0)20 7679 8633.

nor any other data that may allow your identification.

I, the researcher in charge of this study, will keep a consent from you to participate in this study, and you will be provided with a copy. The sponsor of this research is the Department of Anthropology at University College London, with collaboration of the Department of Anthropology at Universidad de Chile. You can contact the researcher in charge of this study at any time, using the contact details mentioned above.

Please discuss the information above with others if you wish or ask us if there is anything that is not clear or if you would like more information.

It is up to you to decide whether to take part or not; choosing not to take part will not disadvantage you in any way. If you do decide to take part you are still free to withdraw at any time and without giving a reason.

All data will be collected and stored protecting your privacy and anonymity, in accordance with the British legislation of data protection (Data Protection Act 1998.)

UCL DEPARTMENT OF ANTHROPOLOGY
University College London, 14 Taviton Street, London WC1H 0BW, UK
Tel +44 (0)20 7679 8633.

Informed Consent Form for Adult participants from the Agricultural Communities in Research Studies

Please complete this form after you have read the Information Sheet and/or listened to an explanation about the research.

Title of Project: **Genetic and Cultural Adaptation to Milk Consumption in Goat Herders from the Semiarid Region of Coquimbo, Chile.**

This study has been approved by the UCL Research Ethics Committee
(Project ID Number): **2967/001**

This study has been approved by the University of Chile Social Sciences Research Ethics Committee (CEDEA)
(Project ID Number): **078/2010**

Thank you for your interest in taking part in this research. Before you agree to take part, the person organising the research must explain the project to you.

If you have any questions arising from the Information Sheet or explanation already given to you, please ask the researcher before you to decide whether to join in. You will be given a copy of this Consent Form to keep and refer to at any time.

Participant's Statement

I _____

- have read or listen to the notes written above and the Information Sheet, and understand what the study involves.
- understand that if I decide at any time that I no longer wish to take part in this project, I can notify the researchers involved and withdraw immediately.
- declare my participation has not been forced by the research staff nor by thirds parts.
- understand that I must not take part if I am under 18 years old.
- understand that I can get the results of my DNA test only with the provided code, since my name is not recorded nor linked with the data.
- understand that I will not be charged nor paid for my collaboration in this study.
- consent to the processing of my personal information for the purposes of this research study.
- understand that such information will be treated as strictly confidential and handled in accordance with the provisions of the British legislation on personal data (Data Protection Act 1998.)
- agree that the research project named above has been explained to me to my satisfaction and I agree to take part in

UCL DEPARTMENT OF ANTHROPOLOGY
University College London, 14 Taviton Street, London WC1H 0BW, UK
Tel +44 (0)20 7679 8633.

this study.

Do you want to know the results of your DNA test? YES NO

Signature: _____ Date: _____

Researcher's Statement

I, Nicolás Montalva,

- have provided the participant a copy of the information sheet, allow him/her to ask for help if he/she cannot read, and have answered all his/her questions about this research.
- think that he/she has understood the provided information, including risks, benefits and rights in regard to his/her participation.
- have provided enough information to allow the participant to take an informed decision about his/her participation.
- declare that I have not forced or influenced the participant's decision in any way.
- will not charge nor pay to the participant his/her collaboration in this study.
- declare that personal information will be treated as strictly confidential and handled in accordance with the provisions of the British legislation on personal data (Data Protection Act 1998.)

Signature: _____ Date: _____

Appendix E

Selection coefficient and sample size to detect deviations from HWE

Rule: Given a sample size n, and a significance level defined by the value of the χ^2−statistic, significant deviation from HWE will be observed only if selection coefficient $s \geq \sqrt{\chi^2/n}$.

Proof. Relative fitnesses of each genotype (2 alleles) are $A_1A_1 = 1$, $A_1A_2 = 1-hs$, $A_2A_2 = 1-s$, where s is the selection coefficient and h the heterozygous effect (Gillespie, 2004). Therefore, observed absolute frequencies for a population deviated from HWE due to selection are those expected under HWE $\times$ relative fitness $\times$ sample size (n), as shown in Table E.1.

Genotype	HWE	Under selection
A_1A_1	np^2	np^2
A_1A_2	$n2pq$	$n(2pq - hs)$
A_2A_2	nq^2	$n(q^2 - s)$

Table E.1. Absolute frequencies for each genotype (2 alleles) under HWE and under selection. The model assumes no other evolutionary forces.

Under a model of full dominance $h = 0$, and thus expected and observed values of A_1A2 are equal. Deviation from HWE is normally tested a χ^2−test. A population is considered to be deviated from HWE if the result of the χ^2−test statistic is higher than the expected by chance (given a desired significance level with 1 degree of freedom in this case). The χ^2−test statistic is:

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(np^2 - np^2)^2}{np^2} + \frac{(n2pq - n2pq)^2}{n2pq} + \frac{(n(q^2 - s))^2}{nq^2}$$

And because, by definition, relative frequencies add up to 1:

$$\chi^2 = \frac{(np^2 - np^2)^2 + (n2pq - n2pq)^2 + n(q^2 - s)^2}{n}$$

Because there is no difference between HWE expectations and the absolute frequencies of the genotypes A_1A_1 and A_1A_2, the first two terms are zero. And, by distributive property, the third terms is:

$$\chi^2 = \frac{n(q^2 - s)^2}{n} = \frac{(ns)^2}{n} = ns^2$$

Dividing by n, $s^2 \geq \chi^2/n$. Therefore, to be detectable by the χ^2–test (i.e. to yield a value of the χ^2 statistic greater than the required for the desired significance level with 1 degree of freedom), s needs to be:

$$s \geq \sqrt{\frac{\chi^2}{n}}$$

□

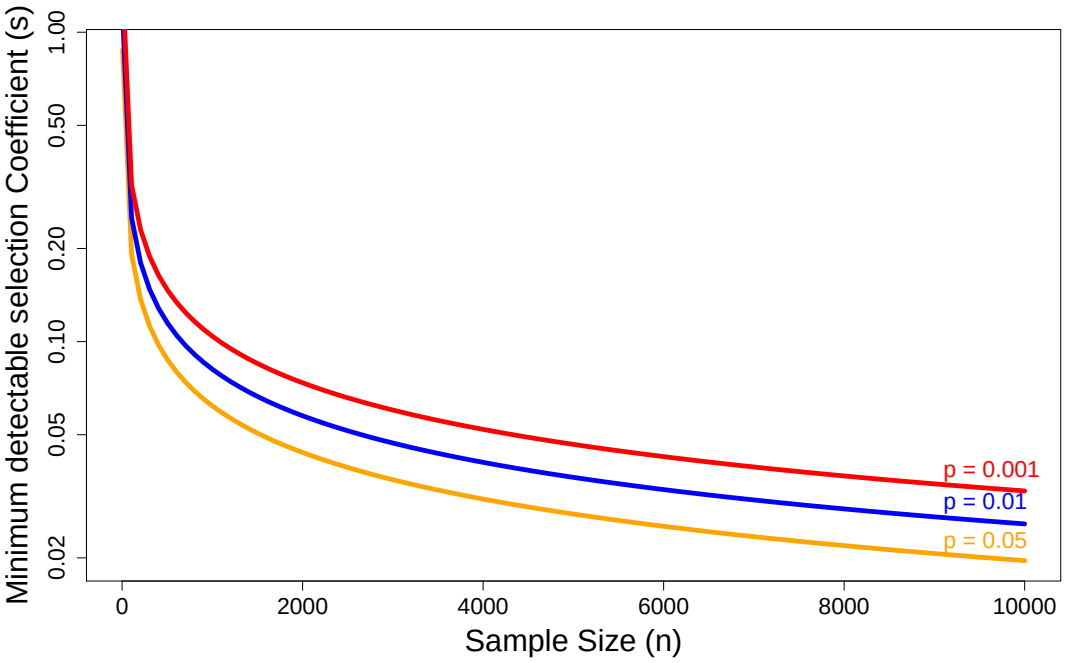

Figure E.1. Plot of the function $f(x) = \sqrt{\chi^2/x}$ describing the minimum selection coefficient s detectable by HWE as a function of sample size n. Each curve corresponds to different values of χ^2 for each annotated significance level.

Appendix F

Analysis of surnames to determine sites for collection of samples

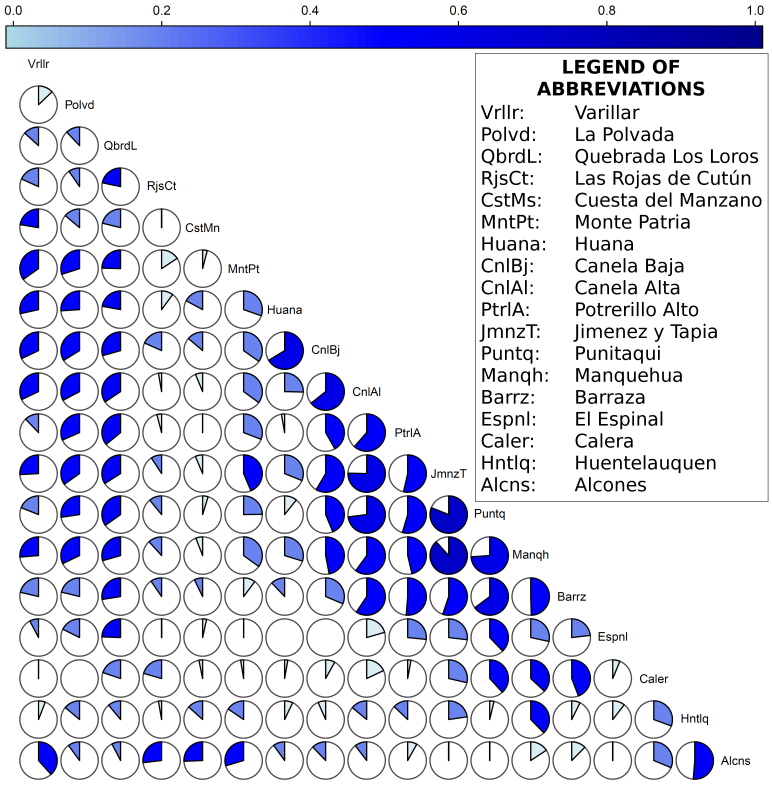

Figure F.1. Correlation between surname frequencies in each pair of communities. All the communities with more than 500 registered commoners were included.

The small size and the expected high inbreeding of the people from the Agricultural Communities presents a problem that has to do with including closely related individuals in one sample, with a potential for distortion in any association study. Before starting the pilot fieldwork, surnames

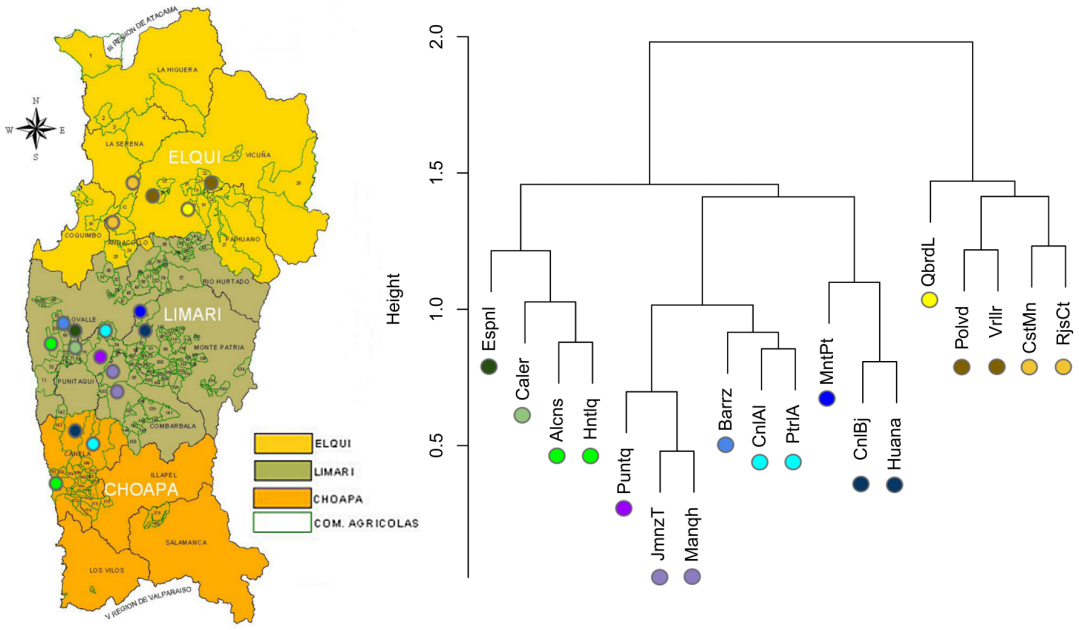

Figure F.2. *Left:* Map of the Region of Coquimbo, showing the location of the communities included in the analysis. *Right:* Hierarchical Cluster Analyses of Hedrick standardised kinship coefficient matrix per community.

were used as markers of relatedness to assess the extent of this problem and decide villages to be included in the pilot study.

Our surnames analysis was done based on the lists from the Land Registry authority, available from the website of the regulatory governmental body in charge of the Agricultural Communities OTCA (Chilean Ministry of National Assets. Retrieved: 5 November 2010., 2010). Presuming that the census includes all the people living under a single land registration and non–registered commoners, we should expect that the census figures represent the population sizes more accurately. Census data were used to select a total of 18 communities with more of 500 people, and then, lists of names were retrieved from the Land Registry records for each of these 18 communities.

Our approach was based on a matrix of distances build from the correlations on surnames in all the pairs of communities. Closely related communities can be identified from the data in Figure F.1. This matrix can be used to perform a hierarchical cluster analysis according to the method described by Hedrick (1971) and Weiss (1980).

Some patterns in the surname distribution can be identified plotting the clustering analysis in a dendogram (Figure F.2). The cluster groups the Communities of Quebrada Los Loros, Polvada, Varillar, Cuesta del Manzano; Rojas de Cutun, as outgroup, in a clade that is geographically related to the Elqui Valley, the northernmost area in the Region of Coquimbo. Patterns on other clusters are not as clear, and seem to not be related with spatial distribution, but there is a trend to group communities closer to the coastline (West) in the same cluster.

In order to answer our initial question, we could confirm a very small population size. Nonetheless, sampling problems were prevented by selecting larger communities, avoiding those with similar surname distribution (in the same clusters), and by visiting villages geographically dispersed from one another.