Genetic History of Latin America:
Fine-scale population structure, sub-continental ancestry and phenotypic diversity

A Thesis submitted for the Degree of *Doctor of Philosophy*

Author:

Juan Camilo Chacón-Duque

Supervisors:

Dr Garrett Hellenthal

Prof Andrés Ruiz-Linares

Department of Genetics, Evolution and Environment

University College London

London, United Kingdom

March, 2018

Declaration

I, Juan Camilo Chacón-Duque 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.

Publications arising from this thesis:

- Chacón-Duque JC, Adhikari K, Fuentes-Guajardo M, Mendoza-Revilla J, … , Hellenthal G*, Ruiz-Linares A*. 2018. Latin Americans show wide-spread Converso ancestry and the imprint of local Native ancestry on physical appearance. *bioRxiv*. doi 10.1101/252155**.
  *Joint last authors. **Under review in *Nature Communications*.
- Adhikari K, Chacón-Duque JC, Mendoza-Revilla J, Fuentes-Guajardo M, Ruiz-Linares A. 2017. Genetic Diversity in the Americas. *Annual Review of Genomics and Human Genetics*. Volume 18. (Available online: August 2017). REVIEW
- Adhikari K, Mendoza-Revilla J, Chacón-Duque JC, Fuentes-Guajardo M, Ruiz-Linares A. 2016. Admixture in Latin America. *Current Opinion in Genetics & Development*, Volume 41, Pages 106-114. REVIEW

Publications not directly related to this thesis:

- Adhikari K, Fuentes-Guajardo M, Quinto-Sánchez M, Mendoza-Revilla J, Chacón-Duque JC,… Ruiz-Linares, A. 2016. A genome-wide association scan implicates DCHS2, RUNX2, GLI3, PAX1 and EDAR in human facial variation. *Nature Communications*. 7:11616
- Adhikari K, Fontanil T, Cal S, Mendoza-Revilla J, Fuentes-Guajardo M, Chacón-Duque JC,… Ruiz-Linares, A. 2016. A genome-wide association scan in admixed Latin Americans identifies loci influencing facial and scalp hair features. *Nature Communications*. 7:10815

- Chacón-Duque JC, Adhikari K, Avendaño E, Campo O, Ramirez R, Rojas W, Ruiz-Linares A, Restrepo BN, Bedoya G. 2014. “African genetic ancestry is associated with a protective effect on Dengue severity in Colombian populations”. *Infection, Genetics and Evolution*. 27: 89-95
- Adhikari K, Mendoza-Revilla J, Sohail A, Fuentes-Guajardo M, Lampert J, Chacon-Duque JC,… Ruiz-Linares, A. Submitted. A genome-wide association scan in Latin Americans underlines the convergent evolution of lighter skin pigmentation in Eurasia. (Under review in *Nature Communications*).

Statement of work

Any joint work with colleagues and supervisors reported in this thesis is clearly stated within each chapter and section. All the figures taken or adapted from published material are part of the two reviews and the preprint in which I am a co-author.

Chapter 1: Considerable amounts of the text presented in sections 1.2 to 1.4 are slightly based on two review papers I jointly wrote with Prof Ruiz-Linares (Adhikari et al. 2017; Adhikari et al. 2016c).

Chapter 2: The new algorithms described here have been developed by Dr Hellenthal and contained in the software SOURCEFIND (Chacón-Duque et al. 2018).

Chapter 3: The description of the reference datasets and part of the analyses have been adapted from Chacón-Duque et al. (2018).

Chapter 4: The whole chapter is based on and adapted from the Supplementary material from the paper in which this thesis is based (Chacón-Duque et al. 2018).

Chapter 5: The whole chapter is an extension of the main results in Chacón-Duque et al. (2018).

Chapter 6: All the analyses presented in chapter 6 were done jointly with K. Adhikari. The description of the phenotypes and part of the analyses have been adapted from Chacón-Duque et al. (2018). The phenotypes have been processed and recorded by several researchers, especially by Kaustubh Adhikari, Macarena Fuentes-Guajardo, Victor Acuña-Alonzo and Mirsha Quinto-Sanchez.

Dataset Contributions

The totality of the sample of admixed Latin Americans has been collected and genotyped by the Consortium for the Analysis of the Diversity and Evolution in Latin America – CANDELA (www.ucl.ac.uk/candela, see Chapter 1, Section 1.6 for details), led by Prof Ruiz-Linares.

Most of the reference datasets were obtained from public datasets (see Chapter 3, Table 3.1 for details). A considerable number of new reference populations were genotyped with CANDELA resources and were collected by different researchers, all of which are co-authors in the main paper derived from this thesis. Native American samples were provided by Gabriel Bedoya, Francisco Rothhammer, Mercedes Villena, René Vásquez, Elena Llop, José R. Sandoval, Alberto A. Salazar-Granara, Maria-Laura Parolin, Karla Sandoval, Rosenda I. Peñaloza-Espinosa, Hector Rangel-Villalobos, Cheryl Winkler, William Klitz, Claudio Bravi, Julio Molina, Daniel Corach, Ramiro Barrantes, Jean Michel Dugoujon, Yali Xue and Maria-Catira Bortolini. Portuguese samples were provided by Verónica Gomes, Carlos Resende, Leonor Gusmão, Antonio Amorim and Maria-Catira Bortolini. Spanish and East / South Mediterranean samples were provided by Pedro Moral.

Acknowledgments

First of all, very special thanks to my supervisors Dr Garrett Hellenthal and Prof Andrés Ruiz-Linares, they have been a wonderful team. Garrett has been just great, his unconditional support and endless discussions have been invaluable for my formation. Andrés with his vast experience has given me a full perspective of the field and of the life as an academic.

Thanks to Professors Mark Jobling and Francois Balloux, for accepting to be my examiners. Your feedback has been invaluable for the final product of this thesis and the closing chapter of my life as a PhD student.

To Dr Kaustubh Adhikari, a great mentor (and "co-supervisor"), and the rest of Ruiz-Linares Lab (Macarena, Javier and Victor), this journey has been by far better with them side by side. To Hellenthal group for all the discussions and support; especially to Lucy and Saioa, always kind, positive and supportive.

To CANDELA and all the "Candeleros". Especially to Professors David Balding and Gabriel Bedoya for constant input and advice.

To all Latin American volunteers. It has been exciting for me, as a proud Latin American, to have the opportunity to spend the last 4 years of my life trying to understand more our history and evolution. To the collaborators who gathered reference samples. And to those communities who gently provided them.

To GEE and UGI, especially to Prof Ziheng Yang who acted as my third supervisor and Upgrade chair; the graduate tutors Dr Lazaros Foukas and Dr Julia Day; and Fiona, Wendy, Manu, Nikolas, Elvira… a long list!

To Colciencias for sponsoring most of my PhD and to University of Oxford (Prof David Bennett and Dr Annina Schmidt) for giving me a hand and sponsoring me when the economy of my country gave me a hard time, while making me part of very exciting research on the genetics of pain.

To my family (especially to my mother, Luisa), my wife Nati, and my beloved friends for bringing me joy and mental peace while in the doctoral rollercoaster.

I dedicate this work to my late brothers, Julián and Rolando, who died prematurely because of genetics and gave me a reason to live for genetics.

Abstract

The history of Latin America involved extensive genetic admixture, particularly between Native Americans, Europeans and Africans. Although these continental contributions to the genetic make-up of the region have been explored previously with genetic data, more precise information about sub-continental contributions has proven elusive. Applying new haplotype-based approaches to ~600,000 SNPs in ~7,000 Latin Americans from Brazil, Chile, Colombia, Mexico and Peru, this PhD thesis provides a comprehensive analysis of the sub-continental ancestry and demographic history of Latin America at a resolution not previously achieved. Furthermore, using measurements on sampled individuals' physical appearances, I explore the impact of this fine-scale genetic structure on phenotypic variation across Latin America.

To achieve these aims, I use a novel haplotype-based statistical technique that I compare to previously published haplotype-based and allele-frequency-based methods, using real data and simulations mimicking Latin American admixture. I show that this new approach provides a substantial increase in accuracy, allowing more precise inference about ancestral components at both regional and individual levels. Strikingly, Native American ancestry across Latin America mirrors the geographic locations of present-day Native groups. Furthermore, non-Native ancestries match to precise areas within the Iberian Peninsula and elsewhere, consistent with historical records detailing migrations and highlighting previously unseen ancestry sources. For the first time in single-sampled individuals, I date the timings of these non-Native Post-Columbian genetic contributions, including newly identified recent contributions related to East Asia. Finally, I show how this sub-continental ancestral reconstruction correlates with variation in pigmentation and facial features in Latin Americans, unearthing new associations that could not be found with available techniques.

Overall, I demonstrate how increasing the robustness and accuracy of fine-scale genetic structure analysis allows a comprehensive picture of the historical and biological diversity of Latin America, highlighting the impact of regional genetic variation on human phenotypic diversity.

Impact Statement

This thesis provides the most comprehensive study to-date on the genotypic and phenotypic variation in Latin America. Studies of this kind are essential in order to obtain a complete picture of human biological diversity as the research on the subject has been strongly biased towards European-derived populations.

I apply a novel statistical approach that quantifies fine-scale, within country genetic sub-structure related to each major ancestry component (i.e. Native American European, Sub-Saharan African, others) within these recently admixed individuals. This approach is improved upon a ground-breaking study that characterized the fine-scale genetic structure of British population (Leslie et al. 2015). Here I not only demonstrate the improved robustness with real and simulated data but also develop a strategy to apply this kind of method in recently admixed populations.

I also use these ancestry sub-components to elucidate the impact of regional genetic variation on physical appearance, providing a template for future studies of phenotypic variation and regional genetic diversity. Such a template is vital given the ubiquity of recent admixture in nearly all world-wide human populations.

Last but not least, I hope that disseminating these results to a wider audience could potentially impact the way Latin Americans conceive of themselves, extolling the value of diversity.

Table of Contents

List of figures

List of tables

Acronyms

1KGP 1,000 Genomes project

AIMs Ancestry Informative Markers

AS-PCA Ancestry Specific Principal Component Analysis

BMI Body Mass Index

CANDELA Consortium for the Analysis of the Diversity and Evolution of Latin America

EAS East Asian

ESM East / South Mediterranean

EUR European

F_{ST} Fixation Index

GWAS Genome-Wide Association Study

HMM Hidden Markov Model

IBD Identity-by-descent

IBS Identity-by-State

IBS (1KGP population) Iberian populations in Spain

ICC Intra-class correlation coefficients

IQR Inter-Quartile Range

LD Linkage Disequilibrium

MCMC Markov Chain Monte Carlo

NAM Native American

PD Procrustes Distance

PAR Pseudo-Autosomal Region

PC Principal Component

PCA Principal Component Analysis

SD Standard Deviation

SES Socio-Economic Status

SNP Single Nucleotide Polymorphism

SSA Sub-Saharan African

YRI (1KGP population) Yoruban in Ibadan, Nigeria

1 Introduction

1.1 Overview

Human populations have likely been exchanging goods, ideas and genes since the birth of mankind around 200,000 years ago (Hunter 2014). However, the maritime navigation during the age of exploration between the 15th and the 18th centuries allowed this exchange to increase in frequency and scale (Bethell 1984; Crawford and Campbell 2012; Kamen 2002), facilitating the interaction of populations that had been diverging for tens of thousands of years (Koehl and Long 2018). The gene flow arising from these contacts created a "natural experiment" that offers an unique opportunity to assess how history shaped the genetic makeup of these populations (Creanza and Feldman 2016; Ruiz-Linares 2014). In an increasingly globalized world, this genetic exchange has become the norm and understanding its consequences is essential (Crawford and Campbell 2012; Pickrell and Reich 2014).

The mixed populations that arose from these transatlantic migrations in the last few hundred years have been usually referred to as recently admixed populations (Seldin et al. 2011; Thornton and Bermejo 2014), and provided the first opportunity for human population geneticists to characterize and quantify genetic admixture (Chakraborti 1986). Only in the last decade, with the availability of larger datasets and the improvement of genotyping technologies, new statistical approaches have provided a significant increase in resolution to differentiate less diverged populations (Lawson and Falush 2012; Novembre and Peter 2016) and to study subtler processes of genetic admixture in numerous populations (Hellenthal et al. 2014; Moreno-Estrada et al. 2013). In addition, the characterization of these differences has allowed us to explore the impact of the genetic ancestry on the phenotypic diversity in both disease and non-disease related traits (Goetz et al. 2014; Tishkoff and Verrelli 2003). For these reasons, the study

of genetic admixture provides a perfect setting in which to explore the demographic history and the evolution of different populations throughout the world.

Latin America probably contains the largest recently admixed populations in the world (Adhikari et al. 2016c), encompassing massive migrations of European conquerors and enslaved Africans and their subsequent admixture with the native peoples of the continent. At a smaller scale, but not less important, migrations from other populations not involved in the initial colonization process have also contributed to the genetic diversity and the population structure of Latin Americans (Crawford and Campbell 2012; Salzano and Bortolini 2002). These demographic processes have generated an extensive genetic and phenotypic diversity throughout the region (Salzano and Sans 2014), and the characterization of such diversity has unearthed patterns of mating, population structure and genetic ancestry as well as new insights into the genetic architecture of complex human traits, as evidenced by admixture mapping and Genome-Wide Association Studies (GWASs) (Adhikari et al. 2016c; Bustamante et al. 2011a; Pasaniuc et al. 2011; Price et al. 2007; Wang et al. 2008; Wilkins 2006).

In this thesis, I have implemented new statistical approaches for the analysis of dense genotype data, in an attempt to increase the ability to find and quantify more precisely the populations involved in the admixture processes, and to understand the impact of this fine-scale genetic ancestry on the phenotypic variation in Latin American populations. This study constitutes the most comprehensive analysis on the genetic admixture in Latin America to date and the strategies of analysis developed here can potentially be extended to any recently admixed population.

I start this chapter with a brief overview of the demographic history of Latin America. Then I put into perspective some major findings from genetic analysis and the bearing of this history on phenotypic diversity in the region. Finally, I describe how these findings are being exploited to dissect the genetic architecture of complex human traits and explain the possible impact of these discoveries in other areas of knowledge.

1.2 Demographic history of Latin America

The term “Latin America” commonly refers to the areas of the Americas and the Caribbean where Spanish or Portuguese is the main language. The origin of this term has been debated by historians, some arguing that geographers in the sixteenth century gave this name to the lands colonized by the Spanish and Portuguese kingdoms, while others state it was coined in France in the 1860s to group all the Latin-language-derived (Spanish, Portuguese and French) countries and territories (Meade 2016). For instance, Sánchez-Albornoz, one of the most renowned Latin American history experts, often includes former French colony Haiti in his works (Sánchez-Albornoz 1994). However, I adhere to the former definition as it is useful from the point of view of genetics to separate Iberian America from Non-Iberian America, given the fact that several geographical, demographical and social factors contributed to genetic admixture being a particularly prevalent process across the Iberian colonies during and after the colonization period (Adhikari et al. 2017).

This extensive genetic admixture between the native inhabitants of the continent and the Iberian conquerors started soon after the arrival of the latter in 1492, and continued with the enslaved Sub-Saharan Africans they brought with them. Historical records allow the recognition of the main demographic events such as the collapse of the Native American population and massive migrations of Europeans and Africans (Burkholder and Johnson 2003; Curtin 1969; Sanchez-Albornoz 1974). In addition, the countries and regions established after the independence have also had different histories, including successive settlements in the vast territories of the continent and new massive migrations from other parts of the world (Sánchez-Albornoz 1994), what makes the admixture processes more heterogeneous and difficult to describe (Figure 1.1).

Figure 1.1. Timeline of the demographic history of Latin America.
Information has been extracted from Bethel (1984), Sánchez-Albornoz (1994), Kamen (2002) and Crawford and Campbell (2012).

1.2.1 The initial settlers and their collapse

It is widely accepted that Native Americans are descendants of ancestral North-east Asian peoples who entered around 15,000 years ago through Beringia, the land bridge that connected Asia and North America during the end of the last glaciation (Cavalli-Sforza et al. 1996; Dillehay 2009; Goebel et al. 2008; Meltzer 2009; Reich et al. 2012). The opening of an ice-free corridor that allowed the initial dispersion of people into North-western America has been widely proposed (Pedersen et al. 2016), with these migrants reaching Tierra del Fuego (the southernmost part of the continent) in only 1,500 years (Brandini et al. 2017). However, a growing list of coastal archaeological sites suggest that people arrived by boat and sailed down the Pacific coast before 14,000 years ago, giving a likely explanation for the rapid expansion into South America (Erlandson and Braje 2011; Wade 2017). Moreover, the inverse correlation of genetic diversity with distance from the Bering Strait confirms a north-to-south migration (Ruiz-Linares 2014; Wang et al. 2007). As the populations started settling throughout the continent, they also became genetically differentiated (Bolnick et al. 2016).

According to genetic data from current-day Native American populations, at least three migratory waves from Beringia took place. Most of the populations in the continent are thought to descend from the first population that crossed the land bridge, while the other two gene flow streams are restricted to North America (Reich et al. 2012). An ancient human specimen found in North America (Anzick-1, ~12,600 years ago) shows more resemblance to modern Central and South American populations (Rasmussen et al. 2014), likely explained by the fact that the two latest migratory waves took place later.

Ancient DNA studies have also suggested a fourth migration (Skoglund and Reich 2016). It is likely that central and South American Native populations have received additional contributions from other populations, like later migrations from Siberia in Central America or Austro-Melanesia (Raghavan et al. 2015; Skoglund et al. 2015), but there is little evidence and it is highly likely that most of the indigenous peoples in this region are descendants of the “first Americans”.

Initial site and date densities using archaeological data and radiocarbon estimates evidence low population sizes followed by a rapid increase in sites from 13,000 to 9,000 years ago reaching peaks and abrupt declines likely linked to the

South American mega-faunal extinction, and only starting a steady growth after the predominance of agricultural subsistence (Goldberg et al. 2016). As they expanded through the continent, they found highly heterogeneous environments and faced geographical barriers that caused isolation of groups, developing in the process a range of ways of life. Genetic data have provided additional evidence of the strong serial bottlenecks faced by these populations, since the initial settlements in Beringia (Fagundes et al. 2008) to the Spanish arrival (Lindo et al. 2016; Llamas et al. 2016; O'Fallon and Fehren-Schmitz 2011).

At the time of the conquistadores' arrival around 1492, the distribution and number of Native Americans throughout the continent were uneven probably due to the fact that geography and environmental changes conditioned their dispersal dynamics and social configuration, ranging from hunter-gatherers to complex hierarchical civilizations (Salzano and Bortolini 2002). Several historians have tried to establish the total size of Native populations at the moment of the conquest, and even though the estimates are highly variable, the size was likely to be around tens of millions (Denevan 1992; Sanchez-Albornoz 1974; Thornton 1987).

Figure 1.2 shows a map we elaborated for a literature review where we consider the most relevant estimates and the current political borders, to provide an overview of the magnitude and the variability of indigenous populations throughout the continent at the moment of the initial contact with the Europeans (Adhikari et al. 2017). This variation likely reflects different levels of societal structure, with higher population densities usually coinciding with areas with more social and technological organization (e.g. big populated centres in Mesoamerica and the Andes), and lower densities with simpler structures, like those of hunter gatherer peoples (Adhikari et al. 2017; Bellwood 2004).

Native American populations are thought to have collapsed during the first century of the colonial period, with a reduction of approximately 90% of the population size, meaning that in several colonized areas with relatively small populations the natives were essentially annihilated (Thornton 1987). This catastrophe was primarily caused by violence, famines and infectious diseases. Furthermore, the profound damage to the social structures of the indigenous peoples prevented a rapid population recovery (Sánchez-Albornoz 1994).

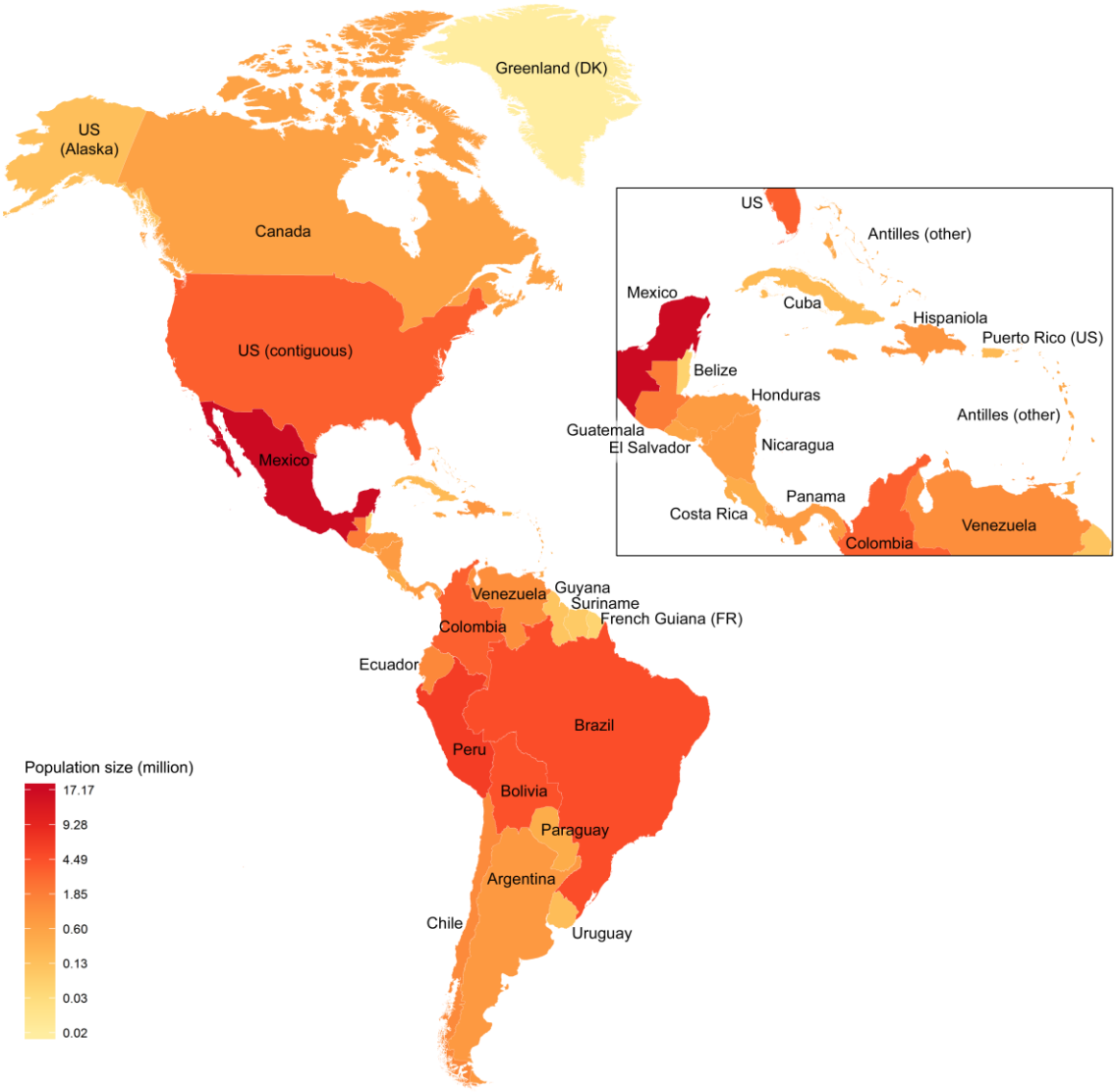

Figure 1.2. Estimated size of the Native American population at the time of Columbus's first landing on the continent, in 1492.
To facilitate comparison with other figures in this article, population size estimates are shown by country, as defined by current borders. The actual population density varied geographically independent of these modern political borders. The population of most of the Antilles has been grouped, as has that of Haiti and the Dominican Republic, which share the island of Hispaniola. The country associated with each American dependency is indicated in parentheses (DK, Denmark; FR, France; US, United States). Exact values and sources are provided in Adhikari et al. (2017), from which this figure was adapted. Generated by J.C. Chacón-Duque and K. Adhikari.

Genetic studies in present-day Native populations have demonstrated that they have lower genetic diversity and higher differentiation between them compared to other continental groups (Wang et al. 2007). Compared to admixed Latin Americans, they usually show high affinity with populations from the same areas, suggesting that the admixture processes took place with the local native communities (Section 1.3.2).

1.2.2 The Conquistadores: The Iberian imprint

In April 1492, in the middle of the maritime race between the Kingdoms of Spain and Portugal for overseas expansion towards the west, the Genoese navigator Cristóforo Colombo (known in the English-speaking world as Christopher Columbus) reached an agreement with the Catholic Monarchs, Queen Isabella I of Castile and King Ferdinand II of Aragon, in order to support his expedition to "discover and acquire islands and mainlands in the Ocean Sea" (Elliott 1984). Columbus landed with his crew later that year on an island in (what is currently known as) the Bahamas and soon after this finding, immigrants mostly from the Spanish kingdom started arriving (Kamen 2002).

During the sixteenth century, these settlers expanded throughout the Caribbean and reached some coastal mainland, including settlements in the Pacific coast (Adhikari et al. 2017). Both kingdoms found themselves fighting for the sovereignty of some of the territories until they agreed to divide the territory according to a decision imposed by the Catholic Church through the "Treaty of Tordesillas", which conferred to Portugal the territories on the west side of an established meridian, including all colonies in the Atlantic Ocean and West Africa, but only a part of the south-east of South America (Elliott 1984; Salzano and Bortolini 2002). This division is still reflected today, with the only Portuguese-speaking country in Latin America being Brazil.

Perhaps the most determinant causes of the extensive genetic admixture in these early stages of the colonization period was the male-biased migration from Europe (i.e. the Iberian Peninsula), favouring the intermixing with Native -and sometimes Sub-Saharan African- women in a patriarchal fashion (Kamen 2002; Lavrin 1992; Morner 1967). This pronounced bias and its widespread pattern amongst several populations in Latin America has been corroborated using genetic data, evidencing the impact of this phenomenon during the foundation of current Latin American populations (Section 1.3.3). Another element favouring higher rates of interethnic matting was the fact that the establishment of the colonies usually coincided with the presence of Native American settlements, precisely because of the nature of the colonization process, finding in these peoples sources of labour and taxation (Salzano and Bortolini 2002; Sánchez-Albornoz 1994).

Latin American populations grew at a rapid pace, and the admixed individuals quickly outnumbered the people of entirely European, Native American or African descent (Sánchez-Albornoz 1994). Even though the migration of Spanish and Portuguese people has never stopped, and other Europeans have also immigrated in different historical periods up to the present (Kent 2016), these contributions have been more restricted geographically and have had a less prominent impact on the make-up of most of the populations (Sections 1.2.4 and 1.3.2).

It has also been of especial interest for historical research (and particularly relevant for the topic of this thesis) to further identify the precise origins of the immigrants. Vast amounts of information about the emigrations that took place out of Spain during the colonial period have been compiled and catalogued (Boyd-Bowman 1964; Boyd-Bowman 1976; Boyd-Bowman 1985), pointing to a predominant southern Spanish origin of the settlers, Seville and Huelva being the most common places of origin. It is estimated that ~37% of the immigrants during the colonial period were Andalusians, followed by people from Extremadura and Castilian provinces, which altogether account for another 46%. Only ~3% of the travellers were from outside Spain. In Section 1.3.2 I explain the main findings in this regard from genetic studies, and in Chapter 5 (Section 5.3.2.3) I describe how our results confirm these historical accounts for the first time.

One controversial topic of research, considers the possibility that considerable numbers of non-Christians who were being persecuted by the Catholic Monarchs at the moment of the conquest arrived to the New World clandestinely, given the fact they were formally forbidden to migrate (Sachar 1994). The records are scant, but Y-chromosome genetic data have suggested a likely contribution of these populations to the colonization (Section 1.3.2). However, caution must be taken with these results, as the genetic structure of the Iberian populations is highly complex and this signal could also be related to earlier events (Botigue et al. 2013; Moorjani et al. 2011).

1.2.3 The Slaves: The involuntary African legacy

The trade of Sub-Saharan African enslaved people was initiated by the Spanish and the Portuguese early in the colonial period and was intensified to mitigate the loss of the labour force due to the collapse of the native population (Curtin 1969;

Thomas 1997). After the treaty of Tordesillas, the Portuguese gained control over the African settlements and around ~1,530 took control of the supply of slaves (Sánchez-Albornoz 1994). Most calculations suggest that more than 5 million Africans arrived in Latin America, with more than 4 million arriving specifically in Brazil (http://www.slavevoyages.org, Figure 1.3).

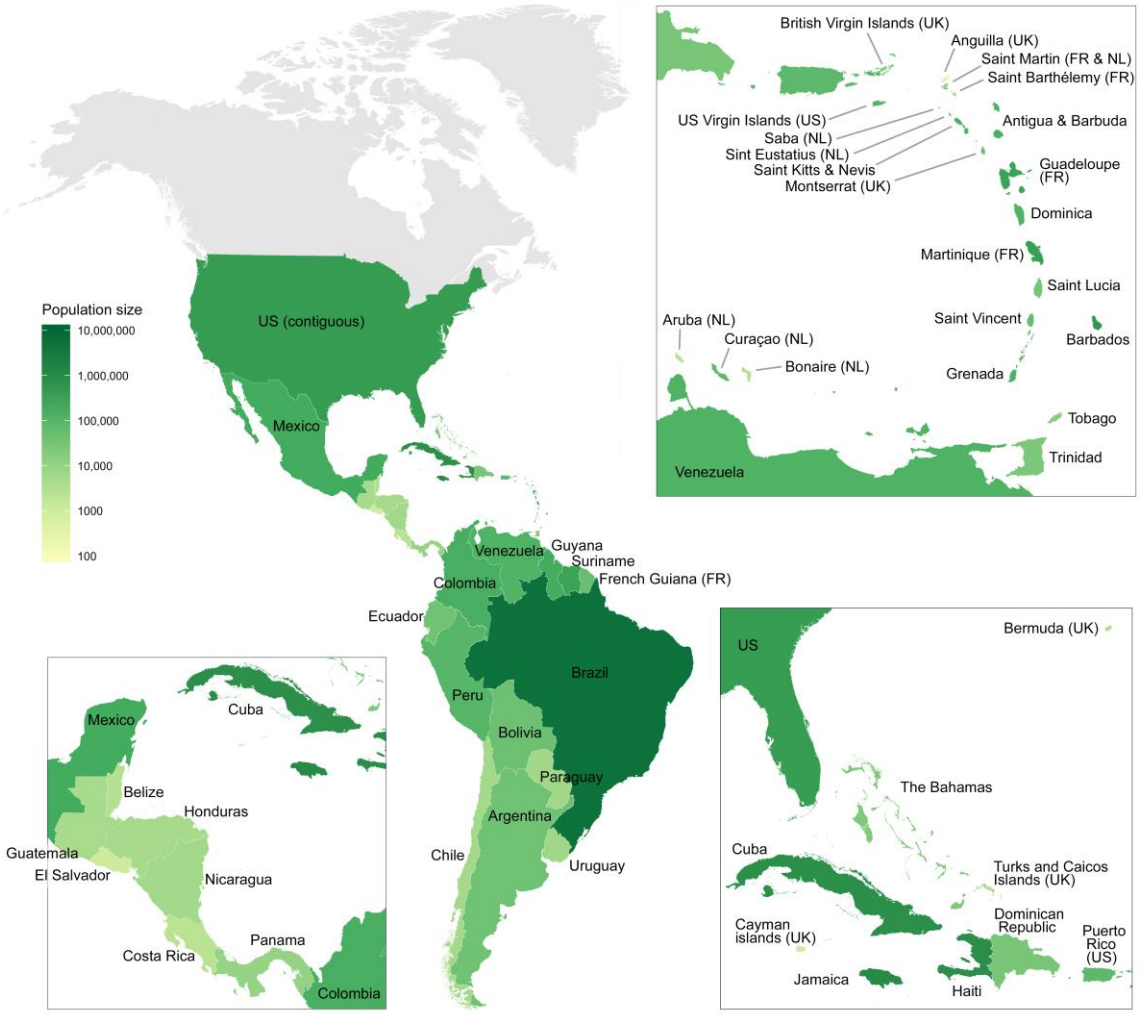

Figure 1.3. Estimated number of Sub-Saharan African slaves transported to the American continent.
To facilitate comparison with other figures, estimates are shown by country, as defined by current borders. Adapted from Adhikari et al. (2017). Generated by JC Chacon-Duque and K Adhikari.

The main sources of the slave trade were located in territories nowadays comprising Senegal and The Gambia on the West Coast and the Gulf of Guinea, and from the 17th century from Angola and Mozambique as well (Sánchez-Albornoz 1994). At the beginning of the colony the Spanish introduced significantly more slaves than the Portuguese, but this changed through time, with Brazil receiving

more Sub-Saharan Africans at the later stages of the colonial period (Curtin 1969). This could explain the increased South / East African ancestry in Brazil observed with genetic data (Section 1.3.2 and Chapter 5, Section 5.3.2.5).

1.2.4 Latin Americans: Up to the present

In the final days of the colonial era and during and after independence in the 19th century, Latin American populations kept increasing due not only to continued immigrations and a recovery of Native American populations, but also to internal expansions of local (often admixed) populations looking for new economic activities (Crawford and Campbell 2012; Parsons 1968; Sánchez-Albornoz 1994).

Although in some countries the independence processes caused a relaxation of the prohibitions regarding interethnic relations and promoted the equality of citizens (Loveman 2014; Wade 2009), in some others further European immigration was encouraged in order to "whiten" the populations (Stepan 1991). The latter approach was quite successful in Southern South America, where some of the biggest European exoduses of the late 19th and the early 20th centuries found their destination (Sánchez-Albornoz 1994). These individuals were mostly of Spanish or Portuguese origin, but also considerable amounts of Germans and Italians made the journey.

Furthermore, migrants from other parts of the world also moved to America during the 19th and the early 20th centuries. In the case of Latin America, considerable numbers of East Asians (mainly Chinese underpaid labourers) have settled along the Pacific coast (Crawford and Campbell 2012; Romero 2010), and their genetic contribution has been detected in several countries (Section 1.3.1). Other migrations have been more restricted, like those from the former Ottoman empire (Fawcett and Posada-Carbo 1997).

The internal expansions have also played a huge role in the demographic dynamics and the diversification of the populations in the region especially through the occurrence of deep and successive founder effects (Koehl and Long 2018). Taking Colombia as an example, its geographical features created the conditions for small groups to settle and grow in isolation for centuries (Carvajal-Carmona et al. 2000). However, the steady growth of these populations and the desire to find better opportunities elsewhere, motivated individuals from these populations to

venture into new lands. These pockets of isolation became suddenly an important source for the settlement of a large portion of Western Colombia (Bedoya et al. 2006).

The following statement is a fragment of a memorial sent to the governor of the province of Antioquia dated on August 27th 1789, in which inhabitants of some villages, Rionegro (where I was born) and Marinilla (where I am from), explain their reasons to pursue the colonization of pristine and prosperous lands (Parsons 1968):

"We, the undersigned vecinos of the ciudad de Rionegro and the Valle de San Jose de Marinilla, come before you in all humility… and declare: We have been led to make this move by our extreme poverty in material goods and the scarcity of lands, either to till as our own or on which to build homes for ourselves and our families. These conditions have been caused by the rapid increase of our people. Thus we have come, penniless, to these mountains of Sonsón, where there is good soil, ample pasture for our stock, salines and rich gold mines, to make our homes and erect a new town. This will bring benefits both to ourselves and to the Royal Treasury… "

This is a reflection of the motivations followed by Latin Americans to expand through the continent without major resistance from the authorities, adding another level of complexity to the study of their genetic backgrounds and their population structure, as genetic drift may have played a major role.

1.3 Genetic history of Latin America

The history of the region has been complicated and in many cases the records are scarce. Population genetics is providing a unique opportunity to contrast, debate and reconstruct past demographic events from a different and (hopefully) less biased perspective. Even though molecular and statistical approaches are constantly improving the way we gather and analyse the data, it is important to consider that most of the samplings to date have not covered the region homogenously and as such the inferences about the genetic history of Latin America may be highly population specific. Bearing this in mind, I describe the most important findings on the genetic history of the region, paying special attention to sub-continental ancestry, which is the central topic of this thesis.

1.3.1 Continental ancestry

A considerable number of surveys of genetic diversity have demonstrated that present-day Latin Americans display a large variation in Sub-Saharan African, European and Native American ancestry proportions both within and between populations. Several reviews have summarized the information available in this respect (Adhikari et al. 2017; Salzano and Bortolini 2002; Salzano and Sans 2014). In Adhikari et al. (2017), we made a careful collection of ancestry estimations all across the American continent, including Latin American countries. We only used information from studies reporting at least 30 Ancestry Informative Markers (AIMs, defined as genetic markers showing high differences in allele frequencies between parental populations (Parra et al. 2004; Pfaff et al. 2001)) and more than 25 samples, in order to filter the most reliable estimates. Figure 1.4 visualizes a summary of these data with averages for each country weighted by the size of every population within a country.

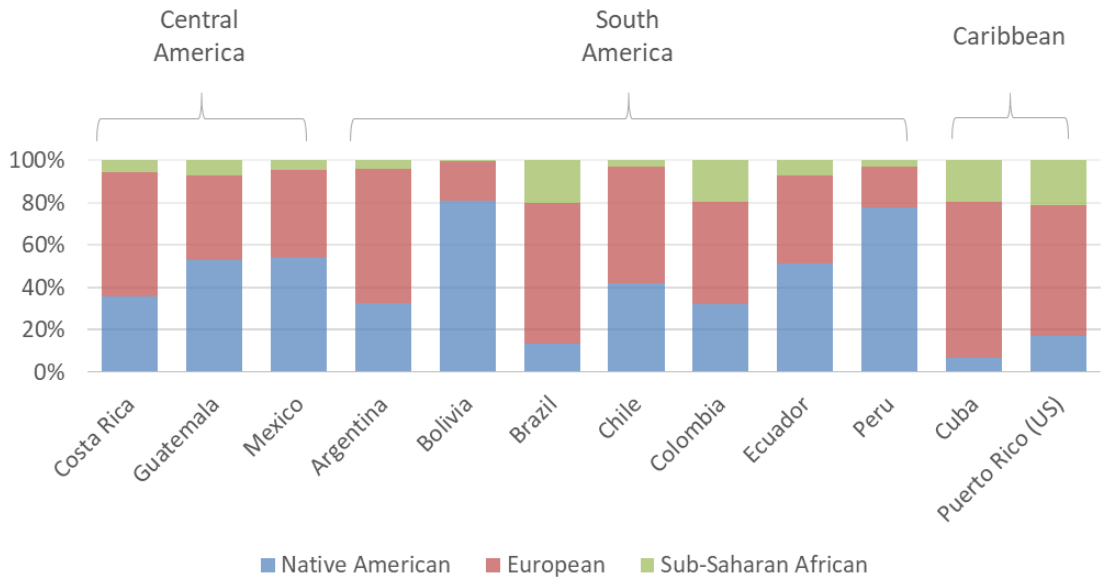

Figure 1.4. Average genetically estimated Native American, European and Sub-Saharan African ancestry for samples from countries in Latin America.
Information for the figure taken from Adhikari et al. (2017). When multiple studies were available for a territory, an average across studies was obtained by weighting based on the size of each population sampled (All this information is displayed and described in detail in the Supplementary material of the review paper). Data collected and figure elaborated by JC Chacon-Duque.

We also compared these averaged genetic ancestry proportions with the frequency of equivalent categories based on self-perceived ancestry as reported in

population censuses (and other surveys) data. We found a strong and significant correlation, suggesting a relationship of this perception with the variation in physical appearance caused by genetics, although other factors may influence it considerably (Adhikari et al. 2017).

The increasing availability of genetic data has allowed us to describe the genetic ancestry of several populations and contrast it with different aspects of the demographic history of their geographic locations. Furthermore, the increase in density of autosomal markers has also allowed the execution of more precise individual ancestry analyses, revealing a striking population structure in Latin America, where these individual estimations usually extend across the whole range of variation of the three main continental ancestries (Adhikari et al. 2016c; Browning et al. 2016; Bryc et al. 2015; Conomos et al. 2016; Eyheramendy et al. 2015; Han et al. 2017; Homburger et al. 2015; Johnson et al. 2011; Kehdy et al. 2015; Moreno-Estrada et al. 2013; Pena et al. 2011; Ruiz-Linares et al. 2014; Silva-Zolezzi et al. 2009; Wang et al. 2008). Figure 1.5 displays individual ancestry estimations calculated using genome-wide data for ~6,300 Latin American individuals from the CANDELA consortium (this is the same dataset used in this thesis, described in Section 1.6), revealing the extent of variation in ancestry at the individual level.

This individual variation has been shown to be correlated with population census size, suggesting that recent events, like population expansion and urbanization, have impacted the genetic make-up of Latin American populations (Ruiz-Linares et al. 2014; Wang et al. 2008). As described in Section 1.2.4, admixture was extensive at the beginning of the colonial settlement but this dynamic changed over time, as populations started splitting and becoming isolated, and in many cases not receiving considerable amounts of immigration up to the present (Bedoya et al. 2006).

East Asian continental ancestry have also been detected, but so far it is restricted to countries which are known to have received considerable immigration of East Asian (mainly Chinese) workers starting in the 19th century, like Costa Rica (Campos-Sanchez et al. 2013) and Peru (Homburger et al. 2015; Sandoval et al. 2013).

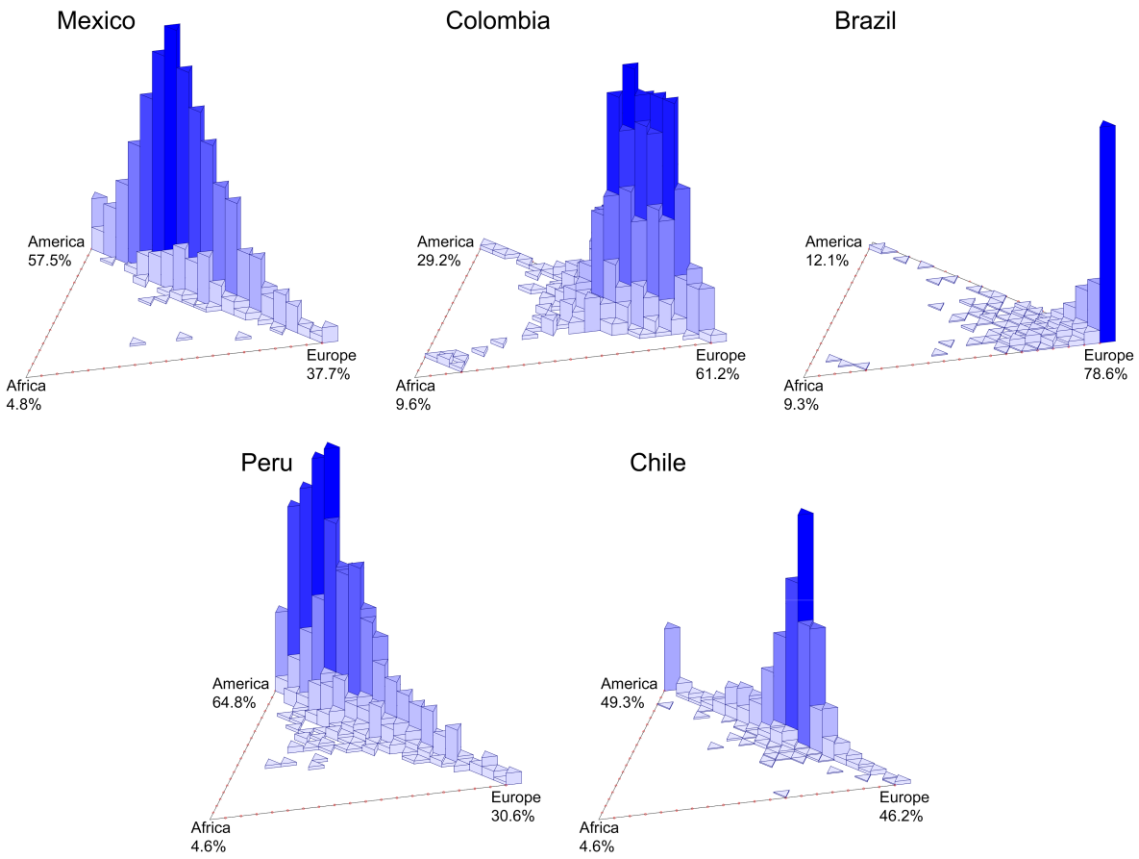

Figure 1.5. Proportion of individual Sub-Saharan African, European and Native American ancestry estimated from 93,328 SNPs typed in 6,357 Latin Americans from five countries (Mexico, Colombia, Brazil, Peru and Chile).
The mean admixture estimates are given in the edges of the triangle plots. Figure modified from Adhikari et al. (2016a and 2016c).

1.3.2 Sub-continental ancestry

Genetic studies have also made it possible to explore patterns of ancestry at sub-continental level, allowing us to narrow down the search for the origins of the ancestors of current Latin Americans. Perhaps one of the most significant findings has been the confirmation that the variation in Native American sub-components of ancestry in admixed Latin Americans matches the regional variation in ancestry detected in present-day Native groups, interpreted as an evidence of “genetic continuity” since pre-Columbian times (Adhikari et al. 2016c). It suggests a scenario where local indigenous populations were somehow assimilated into the populations being created as a product of admixture. This continuity was first demonstrated using mtDNA, showing how the haplotypes of admixed Latin Americans are highly similar to those carried by Native American populations inhabiting the same areas (Carvajal-Carmona et al. 2000; Marrero et al. 2007). The former

study, in which mtDNA was predominantly Native American (see section 1.2.2 for an explanation of the sex bias), showed that the genetic distance between the admixed people from the region of Antioquia (Colombia) and the neighbouring Native American Embera population is not statistically significant, suggesting a genetic continuity of these populations, with the founder women likely coming from the same area (Carvajal-Carmona et al. 2000).

This trend has also been corroborated using genome-wide data, from microsatellites to dense SNPs (Conley et al. 2017; Homburger et al. 2015; Johnson et al. 2011; Moreno-Estrada et al. 2014; Moreno-Estrada et al. 2013; Price et al. 2007; Romero-Hidalgo et al. 2017; Via et al. 2011; Wang et al. 2008). The first attempt to describe this variation in Native American sub-components at the autosomal level was done using microsatellites and revealed increased similarity between the Native American component in different admixed populations and Native groups located in geographic proximity (Wang et al. 2008).

A method called Ancestry-Specific Principal Component Analysis (AS-PCA), has been applied in several Latin American populations and has provided an increase in resolution, allowing the use of Principal Component Analysis (PCA) to investigate differences at the sub-continental level (Browning et al. 2016; Conley et al. 2017; Homburger et al. 2015; Moreno-Estrada et al. 2014; Moreno-Estrada et al. 2013). This approach can be considered partially haplotype-based as it uses phased data for inferring local continental ancestry in order to mask specific ancestries (more details about the differences between allele-frequency-based and haplotype-based approaches are given in Chapter 2). It performs a variant of PCA that allows for missing data, taking as input a masked dataset only containing information for a given continental ancestry. However, PCA does not explicitly quantify proportions of sub-continental ancestry, and the patterns of variation can also be influenced by other factors different to admixture, especially genetic drift or statistical artefacts due to the local ancestry estimation prior to the AS-PC analysis (Browning et al. 2016).

When AS-PCA is performed in the Native American component in populations from Central and South America, it consistently shows that the ancestry of these populations is most closely related to natives sampled in the same areas (Conley et al. 2017; Homburger et al. 2015; Moreno-Estrada et al. 2013). This structure

has been detectable even within a country, Mexico, suggesting that the degree of population structure present in Native American populations is greater than was initially thought (Moreno-Estrada et al. 2014).

Unlike the predominant differentiation of Native American sub-components throughout the region, most of the European ancestry in Latin Americans can be traced to the Iberian Peninsula (Browning et al. 2016; Bryc et al. 2015; Conley et al. 2017; Montinaro et al. 2015), with some Italian and North-western European ancestry detected in Argentina and Brazil (Homburger et al. 2015; Kehdy et al. 2015). Until now, only analyses based on Y-chromosome haplogroups have allowed detection of ancestry from specific regions in the Iberian Peninsula. A study carried on a population sample from Antioquia (North-western Colombia) found that Y-chromosome haplogroups point to a predominant southern Spanish origin, with significant contributions from haplotypes more commonly found in northern Iberian (i.e. Basque Country) and Jewish (including Sephardic) populations (Carvajal-Carmona et al. 2000). However, given the heterogeneous background of Iberian populations (see Section 1.2.2 for historical details), it is not possible to know whether this Semitic heritage was carried by admixed Spanish or by members of these populations (Botigue et al. 2013; Moorjani et al. 2011). Similar to the study conducted in Colombia, a Y-chromosome characterization in Brazil showed that the most frequent Y-chromosome haplogroup has its highest frequencies in Portuguese and Italian populations (Abe-Sandes et al. 2004).

Sub-Saharan African ancestry has also shown some variation at the sub-continental level. Most of the studies carried out to date have found a predominant contribution of non-Bantu populations in North-west and West-central Africa as major sources of ancestry, with smaller contributions from East and South African areas (Bryc et al. 2010; Conley et al. 2017; Fortes-Lima et al. 2017; Kehdy et al. 2015; Mathias et al. 2016; Moreno-Estrada et al. 2013; Tishkoff et al. 2009). Furthermore, some regional variation has been reported, with a higher amount of South and East African ancestry in Brazil (De Mello Auricchio et al. 2007), and more exactly in the southern part of the country (Kehdy et al. 2015), consistent with historical records (Section 1.2.3).

Overall, these findings highlight the high level of population structure of Latin American populations. More specific details about several studies mentioned in

this section can be found in Chapter 5 (Section 5.3.2), where I describe the major findings of this thesis in terms of sub-continental ancestry, demonstrating how our techniques increase the level of resolution and unveil new details about the demographic history of Latin America.

1.3.3 Sex-biased mating

The strong male bias in migration at the beginning of the colonial settlements made the mating between European men and Native women a common feature (Section 1.2.2). This sex-biased mating has been uncovered by studies involving uniparental (mtDNA and Y-chromosome) markers, with several Latin American populations tracing most of their paternal ancestry to Europeans and their maternal ancestry to Native Americans (Alves-Silva et al. 2000; Bedoya et al. 2006; Carvajal-Carmona et al. 2003; Carvajal-Carmona et al. 2000; Green et al. 2000; Grugni et al. 2015; Ruiz-Linares 2014). The common observation across many populations is that the proportion of European ancestry using Y-chromosome markers is consistently larger than the proportion estimated with mtDNA. Conversely Native American and African ancestries are larger when estimated with mtDNA markers (Figure 1.6, (Adhikari et al. 2016c)).

Autosomal data have allowed comparative analysis between X-chromosome and autosomal markers, displaying lower estimates of European ancestry in the X-chromosome compared to the autosomal estimates (Figure 1.6), since women contribute two X-chromosomes to the offspring while men only contribute one (Bryc et al. 2015; Conomos et al. 2016; Homburger et al. 2015; Kehdy et al. 2015; Moreno-Estrada et al. 2013; Wang et al. 2008).

The sex bias in admixture has been reported in several recently admixed populations (Goldberg and Rosenberg 2015; Goldberg et al. 2014; Webster and Wilson Sayres 2016; Wilkins 2006), suggesting that this is a common scenario during colonization.

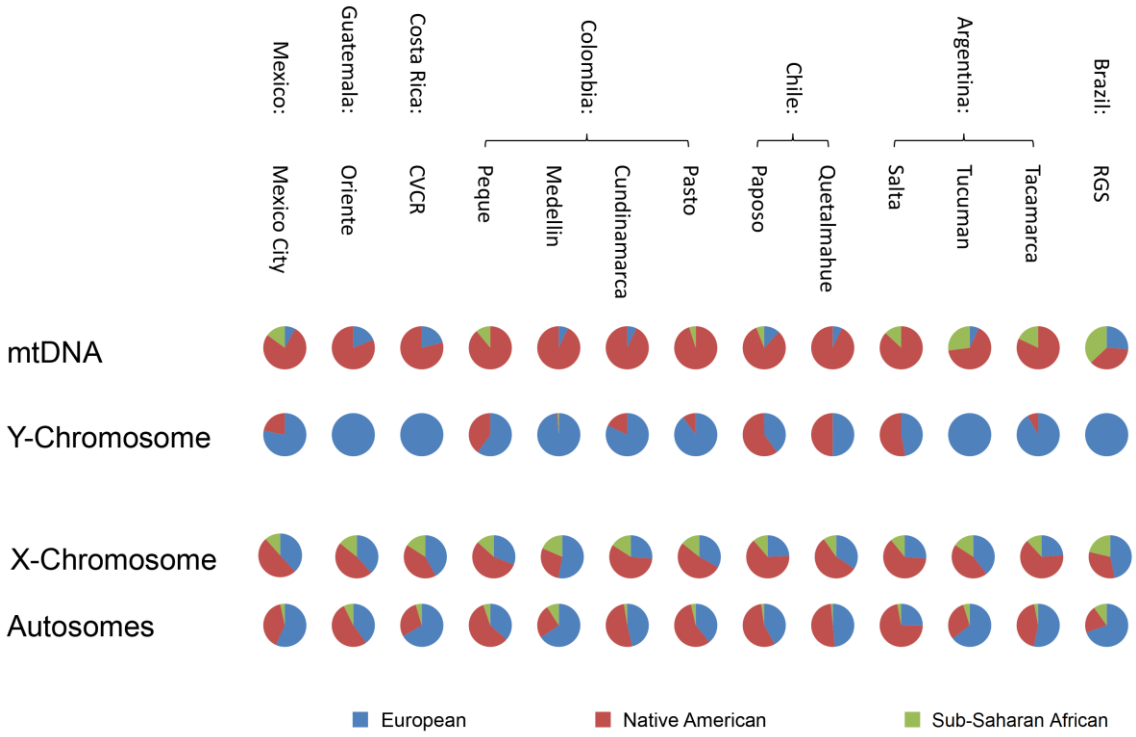

Figure 1.6. Proportion of European, Native American and Sub-Saharan African ancestry estimated with mtDNA, Y-chromosome, X-chromosome and autosomal markers in 13 Latin American population samples.
This figure is modified from Wang et al. (2008) and Ruiz-Linares (2014), and published in Adhikari (2016c). Abbreviations: CVCR (Central Valley of Costa Rica), RGS (Rio Grande do Sul).

1.3.4 Dating the admixture

Statistical modelling of linkage disequilibrium using genetic data (details in Chapter 2), can be used to estimate the time since admixture events (Gravel 2012; Hellenthal et al. 2014; Loh et al. 2013; Patterson et al. 2012; Price et al. 2009). These estimates have been found to be consistent with dates for major demographic events taking place in the areas studied (Adhikari et al. 2017).

Estimates for the Caribbean (~16 generations ago) are older than those calculated in mainland Latin America (~9 to 14 generations), which, according to historical accounts, was populated later (Homburger et al. 2015; Moreno-Estrada et al. 2013; Wang et al. 2008). Moreover, genetic data have also evidenced the complexity of these demographic events, suggesting significant influx of immigrants for long periods (Bedoya et al. 2006) or multiple admixture events involving subsequent flow of European or Africans, depending on the country examined, usually matching with their historical records (Homburger et al. 2015; Kehdy et al. 2015; Moreno-Estrada et al. 2013). Overall, these results are further evidence for

the complex demographic history of the region and its striking population structure (Adhikari et al. 2016c).

1.4 Genetic and phenotypic variation in Latin America

Genetic variation has long been recognized as an important factor underlying phenotypic diversity between human populations (Haldane 1940). The extent of this phenotypic variation has been considerably influenced by key aspects of the evolutionary history of the species, like the dispersal of modern humans throughout the world into varied environmental conditions (sometimes followed by rapid population growth, especially after the rise of agriculture) and the recent mixing of populations from different continental origins (Jobling et al. 2014). The latter aspect has not only created a striking genetic heterogeneity (as described in Section 1.3), but also a high phenotypic diversity as it has already been evidenced in Latin American and other admixed populations (Beleza et al. 2013; Ruiz-Linares 2014; Shriver et al. 2003). Implicitly, this increased phenotypic diversity also extends to disease-related traits (Martin et al. 2017; Rosenberg et al. 2010).

Given the fact that Non-European populations are severely underrepresented in genetic studies (Popejoy and Fullerton 2016; Rosenberg et al. 2010), Latin America becomes an invaluable source to explore new avenues in the genetic architecture of common traits (Wojcik et al. 2017). One of these avenues consists in the characterization of fine-scale genetic structure and its correlation with different traits, as a way to understand more about the impact of human genetic diversity on the genetic architecture of complex phenotypes (Bomba et al. 2017).

However, it is also essential to acknowledge that Latin American populations display a high level of social stratification, and this is often correlated with genetic ancestry (McEvoy and Visscher 2009; Ruiz-Linares et al. 2014; Salvatore et al. 2010). This interaction needs to be considered as it can constitute a confounding factor in association studies linking genetic diversity and disease susceptibility (Burchard et al. 2003; Gonzalez et al. 2005; Risch 2006; Risch et al. 2002; Tang et al. 2005). Some examples are provided in Section 1.4.2.

1.4.1 Non-disease related traits

Several non-pathogenic phenotypes, particularly physical appearance traits, have been widely used in physical anthropology for racial classification, as they show high heritability and consistent differentiation between continental populations (Griffiths 2012; Relethford 2002; Relethford 2009). Furthermore, variation in continental ancestry proportions in Latin Americans (and other admixed populations) is significantly associated with physical appearance traits, being pigmentation the most studied one (Beleza et al. 2013; Hernandez-Pacheco et al. 2017; Parra et al. 2003; Ruiz-Linares et al. 2014; Shriver et al. 2003; Wilson et al. 2011).

Ruiz-Linares et al. (2014), using 30 AIMs typed in 7,342 samples from the CANDELA Consortium, extended this association to a wide range of physical appearance traits, finding significant association between genetic ancestry and height, waist circumference, melanin index (as a quantitative measurement of skin pigmentation), eye colour, hair colour and shape, balding, eye fold, face size and facial features (based on PCA using three-dimensional landmarks). Figure 1.7 shows nose protrusion as an example. This trait has ~84% heritability and displays considerable differentiation between Europeans and Native Americans (Figure 1.7A), with a significant correlation between European ancestry and greater protrusion (Figure 1.7B) (Adhikari et al. 2016b).

Although the amount of phenotypic variance explained by genetic ancestry is usually low, the differences between populations are of great use for identifying specific loci (Ruiz-Linares et al. 2014). GWASs studies carried out in Latin Americans have not only replicated associations previously reported for continental (mainly European) populations but also have identified several novel loci that usually display large differences in frequency between Native Americans, Sub-Saharan Africans and Europeans (Adhikari et al. 2016a; Adhikari et al. 2016b; Adhikari et al. Submitted; Adhikari et al. 2015; Hernandez-Pacheco et al. 2017). Below, I describe the most significant findings for pigmentation and facial features, given the fact these traits are with the ones significant associated with sub-continental ancestry (Chapter 6).

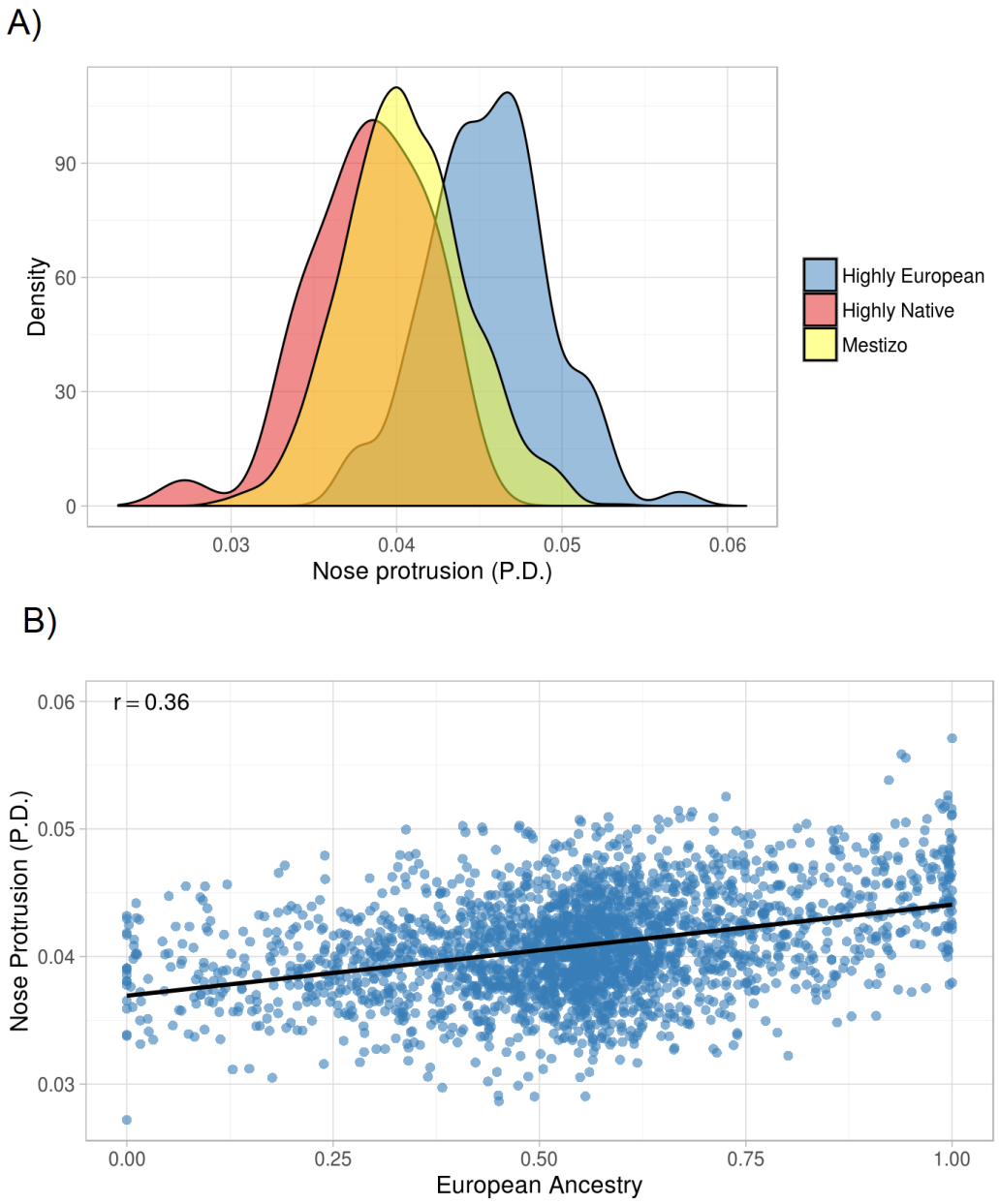

Figure 1.7. Variation of a physical appearance trait in Latin Americans.
A) Density plots for nose protrusion for individuals included in Figure 1.5. To illustrate the phenotypic differentiation between populations contributing to Latin American admixture separate plots are shown for individuals with >95% Native American ancestry (red) or >95% European ancestry (blue). Variation in the rest of the sample is shown on the yellow plot. Nose protrusion was measured as a Procrustes Distance (P.D.) (calculated as detailed in Adhikari et al. (2016b)). **B)** Scatterplot comparing individual nose protrusion with European ancestry and evidencing a significant correlation (r = 0.36; P-value = 2 x 10^{-16}). Adapted from Adhikari (2016c).

Pigmentation is a highly heritable trait that has been extensively studied. Genetic ancestry in Latin Americans explains 19% of the variation in skin pigmentation, displaying a significant correlation between Sub-Saharan African ancestry and higher (darker) skin pigmentation (Ruiz-Linares et al. 2014). In two GWASs we

recently performed in the CANDELA dataset (Adhikari et al. 2016a; Adhikari et al. Submitted), we have replicated associations between skin, hair and/or eye pigmentation and different loci in the genes *SLC24A5, SLC45A2, OCA2, HERC2, TYR, MC1R* and *IRF4*, which have all been reported previously in populations from Europe (Lamason et al. 2005; Liu et al. 2015), South Asia (Stokowski et al. 2007) and/or East Asia (Soejima and Koda 2007). Additionally, we found a significant association between skin pigmentation and the SNP rs2240751, which encodes a missense variant in the gene *MFSD12*, common in East Asians and Native Americans and almost absent in other populations (Adhikari et al. Submitted). Interestingly, *MFSD12* has been recently reported to carry another SNP associated with skin pigmentation (Crawford et al. 2017b), which can indicate an event of convergent evolution, similar to that described for the genes *OCA2* and *MC1R* in Western and Eastern Eurasia (Norton et al. 2007). Another GWAS performed in Puerto Ricans and replicated in African Americans, reported a new variant in the intergenic region between *BEND7* y *PRPF18*, which seems to be mostly present in African-related populations (Hernandez-Pacheco et al. 2017).

Facial features have been far less studied than pigmentation phenotypes and show modest associations with genetic ancestry (only explains 2-5% of the variation for these traits (Ruiz-Linares et al. 2014)), but a considerable amount of associated loci have been found in the last few years. The first loci associated to normal variation in craniofacial morphology, *FGFR1*, was reported in 2005, in a study carried in several populations around the world (Coussens and Daal 2005). In 2012, two GWASs in European populations were published simultaneously, finding the same association between the position of the nasion (the deepest point on the nasal bridge and *PAX3*, a gene previously associated with Waardenburg syndrome, a disease that involves several abnormalities including a broad nasal bridge (Liu et al. 2012; Paternoster et al. 2012). In 2016, we carried out a GWAS for facial features in the CANDELA sample (Adhikari et al. 2016b), not only replicating this finding but also reporting five other gene regions impacting on (mostly) nose shape in the genes *EDAR*, *DCHS2*, *RUNX2*, *GLI3* and *PAX1*. The association with the latter gene has been recently replicated in European populations together with more novel variants (Shaffer et al. 2016). Interestingly, all the significantly associated SNPs in our GWAS display large differences in

allele frequencies between European and East Asian / Native American populations and intermediate values in CANDELA, indicating the increase in statistical power conferred by the admixed populations. In Chapter 6, I show how this differences in allele frequencies can be even detected at the sub-continental level.

The first attempt to compare phenotypic variation with fine-scale population structure (measured by AS-PCA) found a significant association between Native American sub-continental variation and a measurement of lung function in a Mexican cohort (Moreno-Estrada et al. 2014). However, PCA is not an explicit estimator of ancestry (Chapter 2, Section 2.3.1) and its interpretation can be complex. In this thesis I use sub-continental ancestry estimations to look for associations for all the traits that we have previously reported in the published GWASs, aiming to deepen our understanding of the effect of genetic ancestry on physical appearance.

1.4.2 Disease related traits

Associations between continental genetic ancestry and disease-related phenotypes have also been widely found (Goetz et al. 2014; Mountain and Risch 2004; Tishkoff and Verrelli 2003). Perhaps the most studied case is Type 2 Diabetes, which correlates with Native American ancestry (Gardner et al. 1984; Williams et al. 2000). This disease is also a good example because it has been associated to socio-economic status, and even after considering this, a great proportion of its prevalence keeps being explained by ancestry (Campbell et al. 2012; Florez et al. 2009). Other associations with ancestry include cardiovascular disease (Tang et al. 2006), pulmonary disease (Vergara et al. 2013), cancer (Amirikia et al. 2011) and infectious diseases (Chacon-Duque et al. 2014; Ettinger et al. 2009), among others. These findings have suggested that the relationship between prevalence of certain traits and genetic ancestry can be linked to the variation of susceptibility alleles, this being the basic idea underlying admixture mapping (Seldin et al. 2011; Winkler et al. 2010).

GWASs conducted in Latin Americans have also reported novel loci related to Native American ancestry correlated with different disease phenotypes like Type 2 diabetes (DIAGRAM-Consortium et al. 2014; SIGMA-T2D-Consortium 2013),

breast cancer (Fejerman et al. 2014), asthma (Galanter et al. 2014), and autoimmunity (Alarcon-Riquelme et al. 2016; Paternoster et al. 2015).

1.5 Implications of the study of genetic diversity in Latin American populations

The characterization of genetic variation in Latin American populations, and in admixed populations in general, goes far beyond the study of demographic history and its impact on phenotypic variation. The study of these populations can also provide a useful framework to address long-standing questions in different fields of study, including human evolution (Tang et al. 2007), genetic epidemiology (Rosenberg et al. 2010) and forensic genetics (Phillips 2015).

Furthermore, the study of population genetics in admixed populations has been widely explored by social scientists and has a potential impact in society as a whole. Concepts like race, ethnicity and nation have been a subject for discussion, gaining importance in recent times with the developments in genomics (Wade et al. 2014). Race is key concept in this discussions, which is constantly reformulated being both deconstructed (utopian perspective) and reinforced (dystopian perspective) (Tyler 2008). Additionally race (and genetic admixture) are essential factors to consider for public health purposes (Royal et al. 2010).

1.5.1 Human Evolution

The study of genetic variation in Latin America is crucial for understanding the genetic basis of biological attributes differentiated between the ancestral populations, many of those being probably the subject of natural selection (Salzano 2016). It has been proposed that the colonization imposed strong environmental challenges to both Natives and newcomers, especially related to infectious diseases (Cook 1998). By exploring the distribution of local ancestry along the genome, increments of a given ancestry in specific segments probably indicate natural selection, in a similar fashion to admixture mapping (Meyer et al. 2017; Tang et al. 2007).

The impact of natural selection on specific Native American populations has been detected (Crawford et al. 2017a). In this scenario, the characterization of sub-continental ancestry could aid the performance of analysis to understand environmental adaptation in admixed individuals carrying specific regional Native ancestries. Additionally, a better understanding of the impact of fine-grained genetic structure on phenotypic variation in physical traits could help to disentangle long debates on the patterns of dispersal of modern humans as genetic data have sometimes provided contradictory results compared to cranial morphometric data (Manica et al. 2007; Reyes-Centeno et al. 2014).

1.5.2 Genetic epidemiology

As mentioned previously, recently admixed populations are advantageous for genetic association studies given their increased genotypic and phenotypic diversity (Burchard et al. 2003; Rosenberg et al. 2010; Thornton and Bermejo 2014). At the end of the 1990s, the estimation of local continental ancestry (ancestry at the loci level) emerged as a possibility to map disease-related loci showing differences between continental populations by testing for association between traits and the ancestry at loci, allowing the development of the approach known as Admixture Mapping (McKeigue 1998). The LD generated by recent admixture makes it possible to map the entire genome with a small subset of AIMs, making it a cost-effective process for disease-related loci discovery. Modelling studies have suggested that between 2,000 and 5,000 AIMs are sufficient for admixture mapping when analysing a population product of an admixture event taking place 15 generations back (Seldin et al. 2011). This approach has been applied successfully and more recently the availability of dense SNP datasets has made it possible to exploit the advantages of SNP association testing and admixture mapping simultaneously (Qin and Zhu 2012). For instance, an admixture mapping study in women of Latin American origin in the United States found a loci significantly associated with breast cancer that was later replicated by the same researchers through GWAS (Fejerman et al. 2014; Fejerman et al. 2012).

In Section 1.4 I describe how GWAS in Latin Americans have allowed the identification of novel loci associated with several disease and non-disease related loci (Rosenberg et al. 2010). However, there is also challenges to association studies

in admixed populations, mostly related to an imprecise characterization of population structure, which can create both false-positives and false-negatives in the analyses (Wang et al. 2011). The characterization of fine-grained genetic structure could help to overcome these challenges by accounting for additional population structure in association studies focused on Latin American populations (Adhikari et al. 2016c; Conomos et al. 2016). Moreover, it can also potentially provide more precision to pinpoint associated loci and to account more effectively for the effects of genetic structure in the genetic architecture of complex traits, towards a reduction of the missing heritability (Zaitlen et al. 2014).

1.5.3 Forensic genetics

The identification and clustering of individuals into bio-geographical categories, and the prediction of physical features based on DNA information are some of the central goals of forensic genetics (Kayser 2015; Phillips et al. 2014). These goals could be benefited by the increase in resolution displayed by sub-continental ancestry estimations, and could eventually aid the selection of Ancestry Informative Markers (AIMs) that maximize the informativeness (Phillips 2015). As discussed in section 1.5.1, sub-continental ancestry can potentially serve to understand subtler levels of genetic differentiation with considerable effects on physical appearance traits.

However, social, ethical and legal concerns have been raised about the creation of databases with genetic profiles and the prediction of phenotypes using genetic data, as there is a lack of systematic studies measuring the “forensic utility” of these approaches (Williams and Wienroth 2017). It becomes imperative to ensure the scientific validity and reproducibility of the methodologies applied, given the fact that forensic genetics is playing a fundamental role in (*i*) the detection and conviction of criminal suspects, (*ii*) the quality of expert evidence in criminal trials and (*iii*) the development of new forms of biological surveillance.

1.5.4 Other implications

The use of race-related concepts in human genetics has been described as “problematic at best and harmful at worst” (Yudell et al. 2016). Latin America is not the exception as the conception of Latin American populations as the product of genetic admixture between Europeans, Sub-Saharan Africans and Native Americans has been interpreted in a social context in two opposite ways (Wade et al. 2014). It has been seen as a problem in the sense that has supposedly created a “racial degeneration” and as an opportunity because many countries after the independence have sought to build national identities around the “mestizaje” (genetic admixture).

Another implication arises from the relationship between genetic ancestry and population health (which needs to be addressed carefully considering other factors such as social disparities) as it might have important consequences in decision-making processes regarding public policy (Royal et al. 2010).

1.6 Consortium for the Analysis of the Diversity and Evolution of Latin America - CANDELA

This thesis has been conceived as part of the Consortium for the Analysis of the Diversity and Evolution of Latin America (www.ucl.ac.uk/candela), henceforth Denoted CANDELA. This consortium, led by A Ruiz-Linares, is an international initiative studying the biological diversity of Latin Americans and its social context.

The CANDELA dataset consists of 730,525 SNPs (Illumina Omni Express bead chip, see Chapter 3, Section 3.2 for details) from 6,852 individuals ascertained in five Latin American countries (Brazil N=676, Chile N=1,891, Colombia N=1,713, Mexico N=1,288 and Peru N=1,284; Figure 1.8). This sample has been described in detail in Ruiz-Linares et al. (2014). Briefly, adult individuals of both sexes were ascertained at one main recruitment site per country (Porto Alegre in Brazil, Arica in Chile, Medellín in Colombia, Mexico City in Mexico and Lima in Peru). A structured interview recorded the birthplace of volunteers and their ancestors (up to grandparents), as well as information on the language(s) spoken by them.

Additionally, a wide range of physical appearance traits were also collected, by physical observation of the volunteers and/or by examining facial photographs.

Most of these traits have already been included in a series of GWASs we published in the last few years (Adhikari et al. 2016a; Adhikari et al. 2016b; Adhikari et al. Submitted; Adhikari et al. 2015). From these publications we selected 28 traits for the analyses in this thesis, which are described in detail in Chapter 6.

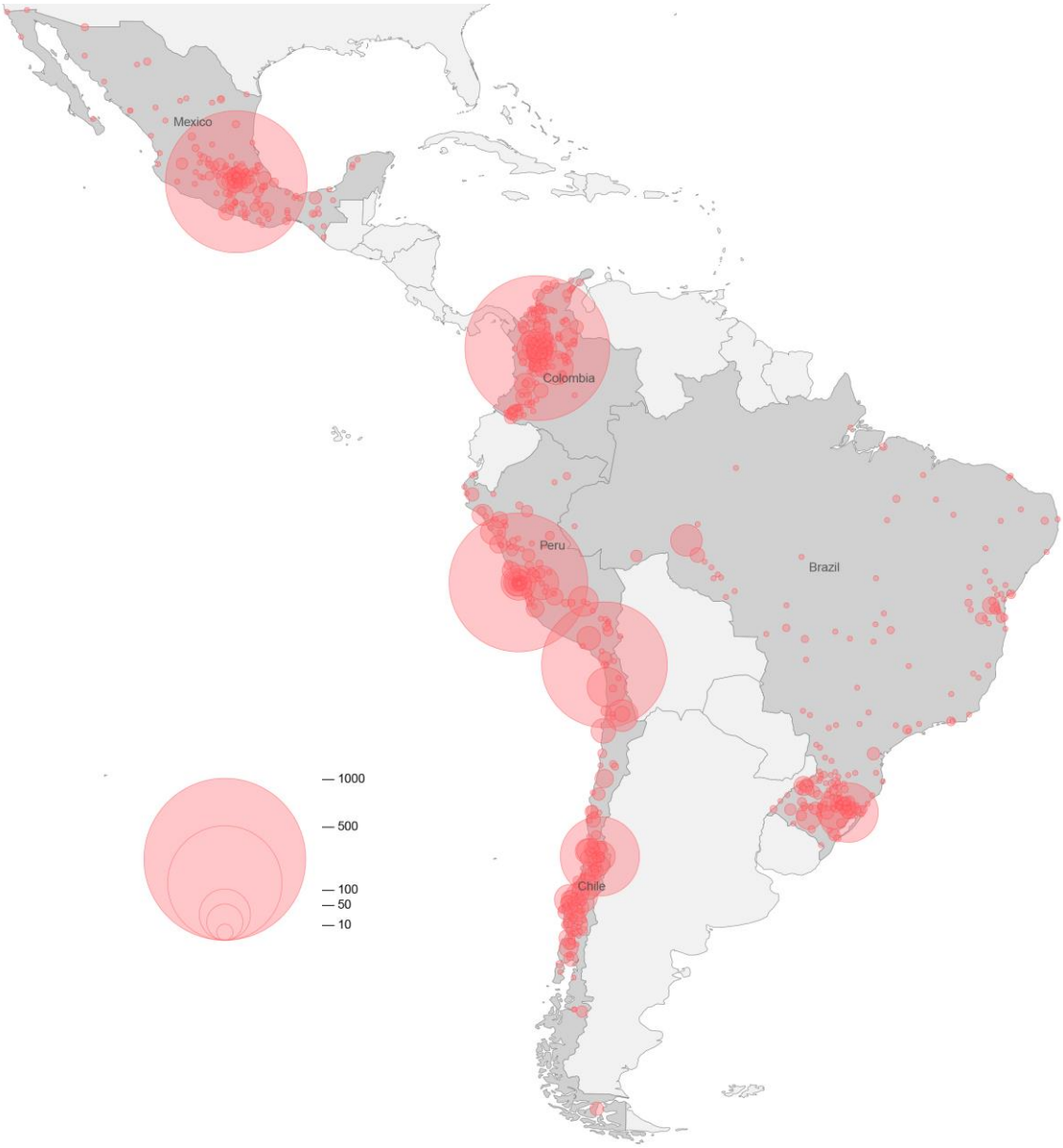

Figure 1.8. Birthplace location of CANDELA volunteers.
The circle size is proportional to the amount of people sampled on the same geographic location. Elaborated by K Adhikari based on script produced by Nicolas Ray.

As shown in Figure 1.8, although the sampling has a wide coverage, some regions within the countries are severely underrepresented. Also, given the experimental design of the Consortium, the sampling favoured the inclusion of individuals with considerable levels of European and Native American ancestry (over those with Sub-Saharan African ancestry) because the methods developed for

the characterization of the morphological traits were better suited for hybrid populations (Ruiz-Linares and Adhikari, personal communication). These biases need to be considered when interpreting the inferences about the demographic events and I mention this in the text when relevant.

1.7 Thesis aims and structure

This thesis aims to perform a comprehensive analysis of the sub-continental ancestry and demographic history of Latin America and to explore the impact of this fine-scale genetic structure on the phenotypic variation in the region.

For estimating the sub-continental ancestry and characterizing the fine-scale population structure I used a set of haplotype-based methods (see Chapter 2 for details), including a new method developed by G. Hellenthal (Chacón-Duque et al. 2018), with increased resolution over previous approaches.

In Chapter 3 I infer clusters of reference populations and characterize their fine-scale population structure using the haplotype-based software fineSTRUCTURE and supplement this with accessory analyses, aiming to find a reasonable classification for the reference groups/clusters to be used in the sub-continental ancestry inference.

In Chapter 4 I perform a series of simulations to mimic the admixture in Latin America in order to assess the robustness and accuracy of the methods we use to estimate sub-continental ancestry (NNLS and SOURCEFIND) and the dates of admixture events (GLOBETROTTER).

In Chapter 5 I explore patterns of sub-continental ancestry in more than 6,500 Latin American individuals across five countries (Mexico, Colombia, Peru, Chile and Brazil), and interpret these results according to the history of the region. Additionally, I estimate the timings and sources involved in the main genetic admixture events.

Finally, in Chapter 6 I evaluate the impact of sub-continental ancestry on a range of physical features measured in the CANDELA sample.

2 Methods: Approaches to understand the history of recently admixed populations

2.1 Overview

In this chapter I give an overview of different statistical approaches that have been designed and/or can be used to infer genetic distance and relatedness, population structure (focusing in admixed populations) and genetic ancestry using SNPs, providing detailed information on those relevant for the development of this thesis. I use several of these approaches to characterize the fine-grained population structure in a panel of reference populations representing the ancestors of Latin Americans (Chapter 3), to demonstrate the robustness of these approaches for estimating sub-continental ancestry sources and times since admixture (Chapter 4) and to measure the sub-continental ancestry in the CANDELA sample, including the times since and sources involved in the admixture events (Chapter 5). Specific details about the implementation of these methods in the performed analyses are explained in every chapter. Additionally, I look for associations between sub-continental ancestry proportions and physical appearance traits (Chapter 6) and all the analyses performed for this purpose are described in this chapter.

I start by describing measurements of genetic distance and relatedness between populations and individuals, then move on to methods for estimating population structure and, finally, I explain the approaches for estimating both ancestry proportions (at continental and sub-continental scales) and the sources, and times since admixture. All the sections of this chapter are divided according to two main approximations that throughout the manuscript I refer to as (*i*) allele-frequency-based and (*ii*) haplotype-based methods. The difference between these two lies in the way the information contained in dense SNP datasets is utilized.

The former approaches rely on differences in allele frequencies. However, these models do not use all the information contained in dense SNP datasets as they are not designed to deal with linkage disequilibrium (LD), making it necessary to thin the data by removing linked markers. In contrast, haplotype-based approaches take advantage of LD by modelling recombination, making a more comprehensive use of the data and providing additional information about relatedness beyond allele frequency patterns. These approaches have allowed a considerable increase in the resolution to cluster and differentiate individuals and populations as I describe below.

2.2 Genetic distance and relatedness

One of the main goals of population genetics is to understand the amount of differentiation between populations and between individuals. F_{ST} has been the most used approach to quantify the genetic distance between populations (Bhatia et al. 2013; Holsinger and Weir 2009; Jakobsson et al. 2013). At the individual level, approaches to quantify genetic relatedness have been developed based on Identity-by-descent (IBD) measurements, which have been substantially improved by the availability of dense SNP datasets (Thompson 2013; Weir et al. 2006).

Most of these methods do not explicitly account for other evolutionary forces. Even though some of them (including most haplotype-based models) have allowed the incorporation of several parameters like mutation, effective population size and recombination, they still do not account for genetic admixture, an essential process to consider when studying genetic differentiation in recently admixed populations. Given the limitation of genetic distance measurements in these populations, efforts have been made to understand the mathematical properties of F_{ST} when comparing the estimates in admixed populations to those in their parental source populations (Boca and Rosenberg 2011).

2.2.1 Allele-frequency-based methods

2.2.1.1 Genetic distance between populations

A member of the family of measurements known as fixation indices or *F*-statistics (Malécot 1948; Wright 1951), F_{ST} was originally designed for measuring the genetic variation between sub-populations (S) compared to the genetic variation of the total population (T) (Wright 1951), and has been formulated in different ways that measure some aspect of population differentiation (Jakobsson et al. 2013). For instance, F_{ST} can be used to compare two populations and establish their genetic distance:

$$F_{ST} = \frac{Var(p)}{p(1-p)}$$

Where *p* is the mean allele frequency in two populations and *Var(p)* the variance of the allele frequency between two populations (Jobling et al. 2014).

In general, F_{ST} is equivalent to the proportion of genetic diversity that can be explained by the differences in allele frequencies among populations (Edge and Rosenberg 2015; Holsinger and Weir 2009). Following this interpretation, numerous studies have presented results of population differentiation among human groups using F_{ST} estimates, usually ranging from ~0.05 (Rosenberg et al. 2002) to ~0.15 (Barbujani et al. 1997), depending on the kind of genetic markers (Holsinger and Weir 2009; Jakobsson et al. 2013) and the estimator used (Bhatia et al. 2013), while populations within the same continent normally display F_{ST} values below 1% (Novembre and Peter 2016).

The increase of molecular data has provided the possibility to propose new metrics based on the same concept (known as f-statistics (Reich et al. 2009)) and to understand better the basic properties and the limitations of the different F_{ST} estimators using different types of data (e.g. chip genotyping vs. sequence data; Bhatia et al. 2013) and low sample sizes (Willing et al. 2012).

The analyses performed by Willing et al. (2012) corroborated that the most commonly used F_{ST} estimator (Weir and Cockerham 1984) is considerably affected by small sample sizes (Excoffier 2008). Considering that most of the reference populations included in this thesis have extremely low sample sizes (90 out of 117 reference populations contain <15 individuals; Chapter 3, Table 3.1), I found

problematic to interpret F_{ST} results and decided not to report them, as they are not directly associated with the aims of this thesis.

2.2.1.2 Genetic relatedness between individuals

Allelic identity approaches can be used as estimators of relatedness between individuals. The increasing availability of dense SNP datasets allowed the estimation of genome-wide marker-based estimates of relatedness, with several uses in population genetics and genetic epidemiology (Thompson 2013). These measurements are particularly useful for quality controls, by allowing the detection of pedigree errors, cryptic relatedness and experimental errors (Purcell et al. 2007).

PLINK v1.9 implements a Hidden Markov Model (HMM) to detect IBD sharing between pairs of individuals in genome-wide data, an approach first described by (Milligan 2003). By a method-of-moments approach, the probabilities of sharing 0, 1 or 2 alleles identical by descent for any two individuals in the matrix are estimated. It is assumed that these individuals come from the same homogenous and panmictic population (Purcell et al. 2007).

In this model, the number of alleles shared IBS is denoted as *I* and the number of alleles shared IBD as *Z* (in both cases the possible states are 0, 1 or 2), and the prior probability of IBS sharing is:

$$P(I = i) = \sum_{z=0}^{z=i} P(I = i|Z = z)P(Z = z).$$

For each SNP, $P(I|Z)$ is specified and obtained according to the allele frequency for *P(Z=0), P(Z=1)* and *P(Z=2),* and the proportion of alleles shared by IBD between every pair of individuals is calculated using the formula:

$$\hat{\pi} = P(Z = 1)2 + P(Z = 2)$$

More details about additional steps that the model requires to account for biases due to finite samples, are explained in detail in Purcell et al. (2007).

One of the main goals when implementing IBD estimates in population genetic studies is to account for relatedness. The thresholds applied in most studies vary from 0.1 to 0.125, intending to account for close relatives (IBD for first cousins is

about 12.5%), though choosing an appropriate threshold can be influenced by issues such as SNP ascertainment. In Chapter 3 (Section 3.3) I discuss how I applied this as part of the quality controls of our merged CANDELA + Reference populations' dataset.

2.2.2 Haplotype-based methods

Statistical approaches modelling LD patterns between closely located markers in dense SNP datasets not only take advantage of higher amounts of data (i.e. by not needing to prune to decrease LD between SNPs), but also provide additional molecular information (Lawson and Falush 2012). LD patterns reveal traces left across the genome by past demographic and evolutionary events, including geographic subdivision and subsequent population differentiation (Slatkin 2008).

Most of the approaches relating genetic variation to LD are based on the Coalescent theory (Kingman 1982), and more precisely, on a generalization developed for including recombination (Hudson 1990). However, although this framework has been useful for simulated scenarios, it is still too complex at the computational level for statistical inference. An alternative approximation that captures the essential properties of the coalescent process has gained popularity, given that it models patterns of LD effectively by relating these patterns to the recombination process and reduces the computational burden considerably (Li and Stephens 2003). In this section I explain the basics of SHAPEIT2 (Delaneau et al. 2013) and CHROMOPAINTER (Lawson et al. 2012), two programs based on this approximation made by Li and Stephens (2003). I used SHAPEIT2 to infer the haplotype "phase" in the dataset and CHROMOPAINTER to infer haplotype similarity patterns, which can be also seen as estimators of genetic differentiation. Other tools developed based on this approximation include MULTIMIX (Churchhouse and Marchini 2013) and IMPUTE2 (Howie et al. 2011), amongst others.

One advantage of applying these approaches to our dataset is that low sample sizes do not have a major effect as the comparisons can be done at the individual level. However, there are also some challenges for these approaches that need to be considered, mainly associated with the accuracy of recombination infor-

mation introduced in the model and the ascertainment bias introduced by selecting SNPs based on physical distance or LD patterns and in specific populations (Novembre and Ramachandran 2011).

2.2.2.1 Li and Stephens model

In their seminal work, Li and Stephens (2003) developed a Hidden Markov Model (HMM) for interpreting and analysing patterns of LD across multiple loci, with several properties that allow us to make inferences from the data considering all markers simultaneously, capturing major features of the coalescent (e.g. that some individuals are more related than others, and relatedness patterns vary along the genome). This model provided an unprecedented framework for the development of more efficient and sophisticated haplotype-based methods (Li and Stephens 2003). Briefly, every haplotype of a single individual is represented as a mosaic of the haplotypes that are present in the reference panel. This "copying" along the genome indicates the ancestral relationships shared by every two samples (the one in the reference set, and the one being assessed).

2.2.2.2 Phasing: SHAPEIT2

Parsimony approaches were the first approximations to statistically phase genotypic data but only models implementing coalescent approximations allowed accuracy and computational tractability in large datasets (Jobling et al. 2014). SHAPEIT2 implements a version of the HMM developed by Li and Stephens (2003), modelling local haplotype sharing between individuals taking into account mutation and recombination.

To reduce the computational burden every individual is compared with each other by segments (windows) and their haplotypes are modelled choosing a subset of K haplotypes in local overlapping windows of length W Mb in every step of a Markov chain Monte Carlo (MCMC) process. In order to define this subset SHAPEIT2 applies the IMPUTE2 "surrogate family" phasing approach (Howie et al. 2011), where K haplotypes with the smallest distance to the current sampled haplotype are chosen as "surrogate family members" because they (ideally) share

recent ancestry with the study individual. These “informative” haplotypes are expected to capture the majority of the likelihood of the distribution as the model is built under coalescent theory assumptions (Pompanon et al. 2012).

SHAPEIT2 has been shown to be robust to diverse ancestries and can take advantage of haplotype sharing between populations to improve performance (Delaneau et al. 2013). What is more, it has been demonstrated that SHAPEIT2 phasing ignoring pedigree information is also very accurate, showing how the sharing of long-range haplotypes between related samples can help the phasing and that the approach works properly with a wide spectrum of relatedness (O'Connell et al. 2014).

In this thesis all samples were phased together without reference haplotypes, taking advantage of both haplotype diversity and sharing patterns between the specific populations here included, some of which might be under-represented in phasing reference panels. The details of the procedure are given in Chapter 3 (Section 3.5).

2.2.2.3 Inferring haplotype similarity patterns: CHROMOPAINTER

Taking into consideration that recombination breaks up chromosomes progressively in each transmission of genetic material from parents to offspring, haplotype segments shared among individuals become shorter over time since they shared a common ancestor. Therefore, the sharing of longer haplotype fragments typically reflects more recent common ancestry between any two haplotypes.

The integration of this layer of information allows the inference of “haplotype similarity patterns”, which are obtained from a “co-ancestry matrix” that contains estimates of the proportion of the genome of each individual in the matrix that is most closely related to every other individual in the same matrix. The usage of these profiles substantially increases the resolution to differentiate populations and individuals, according to simulated and real datasets (Lawson et al. 2012). Additionally, these profiles can also be summarised at the population level.

CHROMOPAINTER, the software I implement in this thesis for inferring the haplotype similarity (informally, “chromosome painting”) across individuals, estimates the proportion of DNA in a given set of individuals (denoted recipients) that is

most closely related to a set of other individuals (denoted donors). Donors and recipients can be the same set of individuals, or completely different. These donor-recipient relationships switch along the genome reflecting ancestral recombination events, and the software represents such changes as a block-by-block mosaic of the genomes sampled. Like SHAPEIT2, it uses a version of the HMM of Li and Stephens (2003), and the way the transition probabilities are inferred is schematized below, following the explanations in (Lawson et al. 2012):

Every phased haplotype contains L total SNPs ordered according to position within each chromosome. A haplotype $h_* = \{h_{*1}, \dots, h_{*L}\}$ is constituted by the observed allele at each site l, and is reconstructed based on J donor haplotypes $h_1, \dots, h_J$. Linkage Disequilibrium is introduced in the model as a population-scaled vector of genetic distances $\vec{\rho} = \{\rho_1, \dots, \rho_{L-1}\}$ (where $\rho_l = N_e g_l$, with N_e analogous to effective population size and g_l the genetic distance in Morgans between l and $l+1$), as well as a mutation parameter θ, accounting for mismatches between recipients and donors, what has been defined by the program developers as “imperfect” copying.

Finally, a vector of copying probabilities $\vec{f} = \{f_1, \dots, f_J\}$ is created, where each f_j corresponds to the probability of copying from every donor haplotype h_j at any SNP. The conditional probability $\Pr(h_*|h_1, \dots, h_j; \vec{\rho}, \vec{f}, \theta)$ is structured as a HMM and thus, the hidden state sequence vector is defined as $\vec{Y} = \{Y_1, \dots, Y_L\}$ with Y_l as the donor haplotype that h_* copies from at site l. The switches between Y_l and Y_{l+1} occur as a Poisson process with rate ρ_l, with the following transition probabilities:

$$\Pr(Y_{l+1} = y_{l+1}|Y_l = y_l) = \begin{cases} \exp(-\rho_l) + (1 - \exp(-\rho_l))f_{y_{l+1}} & if\ y_{l+1} = y_l; \\ (1 - \exp(-\rho_l))f_{y_{l+1}} & otherwise, \end{cases}$$

And at the same time it allows “imperfect” copying:

$$\Pr(h_{*l} = a|Y_l = y) = \begin{cases} 1.0 - \theta & h_{yl} = a; \\ \theta & h_{yl} \neq a. \end{cases}$$

The final output is defined as the “co-ancestry matrix” with donors listed in columns and recipients in rows, containing either the number of haplotype chunks (*chunkcounts.out*) or the total genome length (*chunklengths.out*) in cM that each recipient copies from every donor. The way these values are estimated from the

transition probabilities is fully explained in Lawson et al. (2012). Figure 2.1 shows a graphical example of the coancestry matrix generated using the chunk lengths generated using data from 1,000 Genomes Project (1KGP) and some Native American reference samples (see Chapter 3 for details).

In this thesis, I use only reference population individuals as donors, as the aim is to characterize the fine-grained genetic make-up and to quantify ancestry at the sub-continental level on the admixed populations that are used as recipients.

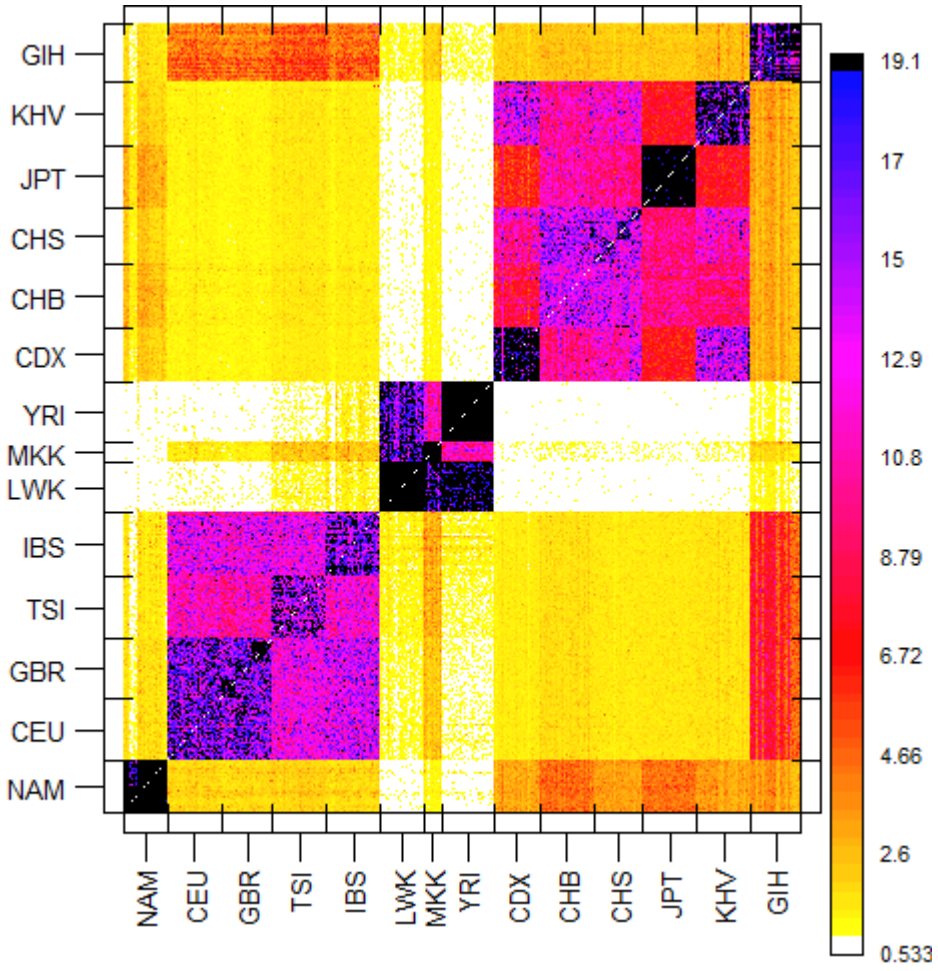

Figure 2.1. Heatmap from coancestry matrix obtained with CHROMOPAINTER. Using 1KGP (Phase 1) samples + some Native American population references. All samples were used as donors and recipients.

2.3 Methods for estimating population structure

The detection and apportionment of genetic structure in human populations is an essential step to reconstruct history (Rosenberg et al. 2002). Although genetic distance measurements have been useful to characterize population structure,

assumptions have to be made about the genetic homogeneity of the members of a population (Foster and Sharp 2002). However, the availability of genome-wide data has allowed us to explore population structure at the individual level and to develop clustering methods (Jobling et al. 2014).

2.3.1 Allele-frequency based methods

Although technically considered another measurement of genetic distance and lacking an explicit clustering model, Principal Component Analysis (PCA) has been used to explore and summarize population structure since just a handful of molecular markers were available (Menozzi et al. 1978; Reich et al. 2008). Briefly, PCA is a multivariate technique that reduces high-dimensional related variables into a smaller set of linear uncorrelated variables that seek to explain most of the variation in the data (Jackson 2003). These uncorrelated variables can be plotted (bi-dimensionally or even tri-dimensionally) and the variation reflected in the samples will be projected in gradients, with the most differentiated individuals for that specific variable lying on the extremes of the axis of variation.

All kinds of matrices can be used for PCA, including “synthetic maps” (Menozzi et al. 1978), genetic distance measurements (Li et al. 2008), and allelic profiles of individuals (Patterson et al. 2006). The latter approach has been developed for genome-wide unlinked data and additionally provides a framework for assessing the statistical significance of the components, and was initially presented as a less computationally expensive option when handling genome-wide datasets compared to model-based approaches like STRUCTURE (Pritchard et al. 2000).

The main issue with PCA relies on the fact that there are no objective criteria for clustering, but some solutions have been proposed, such as Discriminant Analysis of Principal Components (DAPC) (Jombart et al. 2010). This approach performs a Discriminant Analysis (DA) in the PCA results, trying simultaneously to maximize differences between populations and to minimize differences within. Additionally, it performs a K-means clustering analysis (Lee et al. 2009).

Furthermore, results from PCAs can be difficult to relate to geographic patterns and specific migration events. Although it has been proposed that PCs can be used to understand underlying demographic processes (McVean 2009), it has

also been demonstrated that mathematical artifacts can arise from spatial data, suggesting that the historical inferences made from these results need to be taken with caution (Novembre and Stephens 2008). The number of markers included is extremely correlated with the ability of PCA to detect structure at more subtle levels (i.e. fine-structure) (Novembre and Peter 2016).

Model-based methods for clustering have also been developed for allele frequency data, with STRUCTURE (Pritchard et al. 2000) being the most prominent. It is a Bayesian clustering approach based on allele frequencies designed for assigning individuals into populations and inferring admixture proportions based on the K populations inferred. In its initial release the model assumed the populations to be in Hardy Weinberg and Linkage Equilibrium, and later it also allowed for linkage between loci in admixed populations, taking advantage of LD generated by admixture (Falush et al. 2003).

Due to the lack of computational tractability of STRUCTURE when handling massive numbers of SNPs and samples, more efficient models have been proposed. ADMIXTURE (Alexander et al. 2009) uses the same likelihood model of STRUCTURE but maximizes this likelihood using sequential quadratic programming (SQP) rather than inferring parameters using Markov Chain Monte Carlo (MCMC) as in STRUCTURE, which in theory is more efficient and can handle more SNPs with less computational burden.

For these model based approaches, inferring the most accurate K can be problematic. The most popular solution to this problem is a cross-validation procedure that measures the consistency between different runs at a particular K (Alexander and Lange 2011).

I have implemented ADMIXTURE analyses in this thesis. The results can be found in Chapter 3 (Section 3.8).

2.3.2 Haplotype-based methods

As described in section 2.2, haplotype-based methods can have a considerably increased the level of resolution for detecting genetic differentiation. The authors of CHROMOPAINTER developed fineSTRUCTURE (Lawson et al. 2012), a MCMC clustering model based on haplotype similarity patterns. Using the co-

ancestry matrix (chunkcounts.out) generated by CHROMOPAINTER, the model aims to partition the dataset into *K* groups with indistinguishable haplotype similarity profiles.

The main caveat of this model is that it does not directly incorporate admixture, hence it is not suitable for characterizing recently admixed populations. This is the reason why I only apply fineSTRUCTURE to the reference populations' data, as the detection of pre-Columbian admixture should not bias the interpretation of these results. Specific details about this analysis are in Chapter 3.

2.4 Methods for estimating ancestry proportions

As discussed in Chapter 1, recently admixed populations descending from groups that have high levels of genetic differentiation provide a unique scenario to quantify the genetic contributions from these "parental" populations. Although some of the approaches for inferring population structure (previous section) have been widely used to suggest admixture (e.g. PCA), they are not designed for estimating ancestry proportions explicitly. In this section I explain some approximations to this problem, and provide a quick overview on their limitations/advantages.

2.4.1 Allele-frequency-based methods

Early approximations using allele frequencies in the parental populations to estimate ancestry proportions are based on the assumptions that (*i*) there is no error in the choice of parental groups or in their allele frequencies (implicitly, it assumes that the sampled frequencies represent the true frequencies of the population) and (*ii*) no changes in allele frequencies have occurred independently of the gene flow (Cavalli-Sforza and Bodmer 1971; Chakraborti 1986; Salzano and Bortolini 2002). Some of these approaches are included in programs like ADMIX.PAS (Parra et al. 1998), ADMIXMAP (Hoggart et al. 2004) and STRUCTURE (Pritchard et al. 2000).

The improvement of allele-frequency based approaches was undoubtedly powered by the advances in molecular biology, particularly the development of tools to characterize various types of molecular markers experimentally (such as

SNPs, microsatellites and Indels) which provided the possibility of establishing panels of markers with alleles displaying large frequency differences between populations, defined as Ancestry Informative Markers (AIMs) (Shriver et al. 2003). Later, the availability of genome-wide data has allowed the refinement of these methods.

The most widely-used software for ancestry estimation in dense SNP data is ADMIXTURE. In addition to the unsupervised analysis where the software clusters subsets of individuals' genomes into K partitions, it is also possible to fix the potential sources of admixture and obtain the percentages of ancestry associated to these specific sources.

Often, results from unsupervised ADMIXTURE analyses are interpreted as ancestry proportions that each individual carries from K ancestral populations presumed to have existed. However, results need to be taken with caution as some other forces, especially genetic drift, can generate clusters that do not resemble admixture events between putative ancestral populations. For instance, admixed populations with high levels of genetic drift can be incorrectly assigned to their own (presumably unadmixed) cluster (Chapters 3 and 5 discuss this issue in more detail). For supervised ADMIXTURE analyses, on the other hand, surrogates for ancestral source populations are fixed and ADMIXTURE infers the proportion of DNA carried by each individual that is most closely related to these sources. This inferred proportion is often interpreted as the proportion of DNA inherited from these ancestral sources. In addition to the limitations mentioned above, it is also necessary to consider that, in cases where the sources provided have low amounts of genetic differentiation, the software may not have enough resolution to assign the ancestry proportions correctly (see Chapter 5 for details and further discussion).

2.4.2 Haplotype-based methods

The increase in resolution conferred by haplotype-based approaches can also be exploited for ancestry estimation. These approaches have allowed the quantification of ancestry proportions at the sub-continental level. In this thesis I present a new model-based approach for sub-continental ancestry estimation developed by G. Hellenthal, SOURCEFIND (Chacon-Duque et al. 2018). It shows increased

resolution over the Non-Negative Least Square (NNLS) approach proposed by Hellenthal et al. (2014), used to infer ancestry proportions in different populations, including the United Kingdom (Leslie et al. 2015) and Latin America (Montinaro et al. 2015).

As a starting point, individual haplotype similarity profiles are summarized in terms of the reference population individuals (used as donors in the CHROMOPAINTER analysis), preferably grouped according to the clustering provided by fineSTRUCTURE (see Chapter 3 for details). The individual donor values are summed according to these groups and the new value is defined as a "copying vector". To cope with differences in reference clusters' sample sizes and to account for incomplete lineage sorting, each CANDELA individual's copying vector (used as recipients in the CHROMOPAINTER analyses) is modelled as a weighted mixture of the surrogates' copying vectors (Hellenthal et al. 2014; Leslie et al. 2015).

Let $l^r \equiv \{l_1^r, \ldots, l_D^r\}$ be the copying vector describing the total genome length (in cM) that individual (or group) r copies from each of the $d \in [1, \ldots, D]$ donor reference groups as inferred by CHROMOPAINTER (note that copying vectors can also be averaged across recipients to perform the analysis in groups). Here for any r, $\sum_{d=1}^{D} l_d^r = C$, where C is equal to the total genome length of DNA (in cM), and we further define $f_d^r \equiv \frac{l_d^r}{C}$. Henceforth we let *r* denote an admixed individual, and *s* denote a surrogate group. In the latter case, l_d^s represents an average across all individuals from surrogate group *s*.

We assume that:

$$Pr(l^r | l^1, \ldots, l^S, C, \beta^r) = Multinomial\left(C; \sum_{s=1}^{S} [\beta_s^r f_1^s], \ldots, \sum_{s=1}^{S} [\beta_s^r f_D^s]\right)$$

Where $\beta^r \equiv \{\beta_1^r, \ldots, \beta_S^r\}$ are the mixture coefficients we aim to infer and every $s \in [1, \ldots, S]$ represents a "surrogate" group used to describe the ancestry of group *r*. In practice, often all the donor reference groups are used as surrogates, so that $S = D$. However, in our case the surrogates are a subset of the donors so that $S < D$.

We take a Bayesian approach to inferring β^r, further assuming the following:

$$Pr(\beta^r|\lambda) = \text{Dirichlet}\,(\lambda_1, \ldots, \lambda_S),$$

$$Pr(\lambda) = \text{Uniform}(0,10).$$

For each recipient *r*, we wish to sample the mixing coefficients $\{\beta_1^r, \ldots, \beta_S^r\}$ based on their posterior probabilities conditional on $l \equiv \{l^r, l^1, \ldots l^S\}$. We do so using the following Markov Chain Monte Carlo (MCMC) technique. We start with an initial value of λ(0) = 0.5 and sample our initial values of $\beta^r(0) \equiv \{\beta_1^r(0), \ldots, \beta_S^r(0)\}$ from the prior distribution Dirichlet $(\lambda(0), \ldots, \lambda(0))$. Then for $m \in [1, \ldots, M]$:

Update $\beta^r(m) \equiv \{\beta_1^r(m), \ldots, \beta_S^r(m)\}$ using a Metropolis-Hastings (M-H) step:

i. Randomly sample $Y \sim Unif\,(0,0.1)$.

ii. Randomly sample a surrogate s_x and set $\beta_{s_x}^r(m) = \beta_{s_x}^r(m-1) + Y/5$. For numerical stability, if $\beta_{s_x}^r(m) > 1 - 1e^{-7}$, set $\beta_{s_x}^r(m) = 1 - 1e^{-7}$.

Repeat this for 4 additional randomly sampled (with replacement) surrogates s_x.

iii. Randomly sample a surrogate s_x and set $\beta_{s_x}^r(m) = \beta_{s_x}^r(m-1) - Y/5$. For numerical stability, if $\beta_{s_x}^r(m) < 1 - 1e^{-7}$, set $\beta_{s_x}^r(m) = 1e^{-7}$.

Repeat this for 4 additional randomly sampled (with replacement) surrogates s_x

iv. For all other surrogates $s \in [1, \ldots, S]$, excluding the randomly sampled set above, set $\beta_s^r(m) = \beta_s^r(m-1)$.

v. Re-scale $\sum_{s=1}^{S} \beta_s^r\,(m) = 1.0$.

vi. Accept $\beta^r(m)$ with probability $\min(\propto, 1.0)$, where:

$$\alpha = \frac{Pr\left(l^r \middle| l^1, \ldots, l^S, C, \beta^r(m)\right) Pr(\beta^r(m)|\lambda(m-1))}{Pr\left(l^r \middle| l^1, \ldots, l^S, C, \beta^r(m-1)\right) Pr(\beta^r(m-1)|\lambda(m-1))}.$$

Update each $\lambda_s(m)$ for $s = 1, \ldots, S$ using a M-H step:

i. Propose a new $\lambda_s(m)$ from a Normal $(\lambda_s(m-1), sd = 0.2)$.

ii. Automatically reject if $\lambda_s(m) \notin [0,10]$.

iii. Otherwise accept $\lambda_s(m)$ with probability $\min(\propto, 1.0))$, where:

$$\alpha = \frac{Pr(\beta^r(m)|\lambda(m))}{Pr(\beta^r(m)|\lambda(m-1))}.$$

For large *M*, this algorithm will converge to the true posterior distribution of the β^r's (Gamerman 1997). We refer to the final estimates of $\beta_1^r, \ldots, \beta_S^r$, weighted-averaged across posterior samples using the log-posterior as weights, as our inferred proportions of ancestry for group *r* conditional on this set of *S* surrogates. This approach differs from the mixture model procedure previously described (Hellenthal et al. 2014; Hofmanova et al. 2016; Leslie et al. 2015; Montinaro et al. 2015; van Dorp et al. 2015) in that it assumes that l^r is multinomial distributed and solves for β^r using a Bayesian approach rather than a non-negative least squares optimization. This model is similar to one developed by G. Hellenthal and applied to ancient DNA data (Broushaki et al. 2016), but alters the way that λ is estimated and uses a more efficient (in practice) MCMC proposal procedure. The accuracy and robustness of the ancestry estimations obtained by SOURCEFIND and NNLS were evaluated using real data and simulations mimicking Latin American admixture (Chapter 4, Section 4.2).

Currently G. Hellenthal is also developing an alternative, more computationally efficient version of SOURCEFIND that uses the same likelihood function, but which removes Lambda and replaces the prior on the β^r values with a truncated Poisson (mean=3) prior on the number of contributing surrogates *S'*. At each MCMC iteration, this alternative SOURCEFIND allows only a maximum of *S'* surrogates to have $\beta_s^r > 0$ and for the β_s^r values of each of these *S'* surrogates to be 0.01,…,1 in increments of 0.01. The proposed move at each MCMC iteration is as follows. The β_s^r value of a randomly chosen surrogate group is either completely (with probability 0.1) or partially (with probability 0.9) distributed across the other currently included surrogates. (This set of other included surrogates contains up to *S'* members, with new randomly chosen surrogates added if the total number of surrogates is less than *S'*.) With probability 0.5, the β_s^r value is added to that of a single other surrogate; otherwise it is distributed randomly across the other surrogates. This proposal is then accepted or rejected using a Metropolis-Hastings step. Results under this approach ran much more quickly and gave qualitatively similar conclusions in applications to simulated and non-simulated data, as described in Chapter 4 (Section 4.3) and Chapter 5 (Section 5.3.2.7). The R code has been made publicly available with the publication of the bioRxiv preprint (Chacón-Duque et al. 2018).

2.5 Estimation of number of generations since admixture

Recombination can be seen as a time-related process, as it tends to “break” haplotype segments into smaller pieces as generations pass, reducing LD. This information can be used for dating admixture, as these LD patterns will reflect haplotype segments tracing their ancestry back to the populations involved in the admixture process. Theoretically the decay of LD with time follows a negative exponential distribution, and this property has been used to generate models aiming to fit decay curves to an exponential distribution in order to infer approximate times since admixture (Hellenthal et al. 2014).

Models such as ROLLOFF (Patterson et al. 2012) and ALDER (Loh et al. 2013) quantify the exponential decay of LD generated by admixture as a function of genetic distance, and propose statistical tests to fit the data within specific scenarios. MALDER (Pickrell et al. 2014) additionally allows the inference of complex admixture process involving multiple dates of genetic exchange. The main limitation of these approaches is the definition of the populations (i.e. sources) that originally contributed to the admixture process (or processes), as the reference populations may have diverged considerably with respect to the sources, or they could not descend from exactly the same population.

GLOBETROTTER (Hellenthal et al. 2014) tries to overcome this limitation by modelling the source populations as mixtures of the reference donor populations used in the analyses. Haplotype similarity patterns (as obtained with CHROMOPAINTER) can be modelled as weighted mixtures of the sampled donor populations, using the NNLS approach (Leslie et al. 2015). This modelling allows us to represent the original source populations as mixtures of the sampled reference groups, inferring the source rather than fixing it. The target population or individual (e.g. CANDELA) is then represented as a mixture of the profiles of the surrogate sources estimated by the software.

The size and the distribution of the segments matching to every source are estimated using another CHROMOPAINTER output (*samples.out*) that contains the haplotype matching of every SNP for 10 different samples of the hidden state (i.e. which donor is copied at each SNP) taken from the HMM. This information is then used to produce coancestry curves for each pair of donor populations, plotting genetic distance on the *X* axis against a relative probability that measures how

often a pair of haplotype segments separated by a given amount of genetic distance correspond to different donors. These probabilities are calculated using information for every pair of SNPs located from 1cM to 50cM from each other. In the case of a single admixture event over a narrow time period, the decay is expected to be exponential. In the case of multiple admixture events, the decay is expected to be equal to a sum of exponential distributions, one curve per admixture event. GLOBETROTTER determines whether the LD decay curves among all pairings of surrogates can be fitted using a single exponential distribution or whether they are significantly better fitted using the sum of two exponential distributions. Detailed information on the method can be found in Hellenthal et al. (2014).

I show that using GLOBETROTTER in the analysis of recently admixed populations is advantageous because inferences can be performed at the individual level (Chapter 4). This is particularly useful in our sample as the ancestry proportions at the continental level vary enormously, reflecting the difficulty in defining a homogeneous a population made up of recent admixed individuals.

3 Establishing reference panels to represent ancestral sources

3.1 Overview

A major issue on the reconstruction of demographic history and estimation of ancestry using genetic data is the scarce availability of samples that accurately represent the sources of ancestors of current-day individuals/populations. Only recently have dense DNA data from ancient human remains (ancient DNA) become available, and there are still plenty of technical constraints that do not allow isolation of high quality ancient DNA from several regions across the world, especially in tropical and sub-tropical areas where the environmental conditions do not favour the preservation of these remains. Given this scenario, it is often necessary to use contemporary samples as “surrogates” for the original parental populations. This immediately poses a challenge on the interpretation of the results, as these surrogate populations may have changed substantially compared to their ancestors, and further the populations derived from admixture among the original ancestral groups may have undergone drastic changes.

It is thus essential to establish systematic and objective protocols for the clustering of reference populations based on their genetic profiles, as the accuracy and reproducibility of this clustering is essential to make the analyses robust and reproducible and to provide a clearer interpretation of the results. This is especially important in the case of recently admixed populations, where it is necessary to distinguish such recent mixing from patterns of admixture already present in their parental populations.

In this chapter, I implement clustering using the haplotype-based software fineSTRUCTURE and supplement this with accessory analyses, aiming to find a reasonable classification for the reference groups/clusters we will be using in the sub-continental ancestry inference. As described in Chapter 2 (Section 2.3.2)

haplotype-based methods not only provide a model-based framework for clustering, but also show an increase in power detecting subtle differences that cannot be achieved with allele-frequency-based approaches (Lawson et al. 2012). This is the first important step towards providing an accurate characterization of the sub-continental ancestry patterns in CANDELA and to achieve an objective interpretation of the results.

3.2 Reference dataset

To characterize the sub-continental ancestry in the CANDELA individuals I collated a reference population dataset from different regions across the world having potentially contributed to admixture in Latin America. I combined publicly available data with data from newly genotyped samples obtained for this thesis. As described in Table 3.1, altogether I collated data for 2,359 individuals from 117 reference populations (38 Native American, 42 European, 15 East/South Mediterranean, 15 Sub-Saharan African and 7 East Asian) distributed geographically as indicated in Figure 3.1. The preparation of the dataset was done with the support of K. Adhikari.

Of these, 430 individuals from 42 population samples (comprising 27 Native American, 7 European and 8 East/South Mediterranean), were newly genotyped on the Illumina HumanOmniExpress chip which contains 730,525 SNPs, including markers in all autosomal chromosomes, X and Y chromosomes, and the Pseudoautosomal region (PAR) (Table 3.2). An additional group of 1,230 SNPs was not assigned to any of the chromosomes.

Only SNPs from autosomal chromosomes were used for the analyses presented on this thesis for several reasons. Firstly, the Y chromosome does not recombine and cannot be used for inferring haplotype similarity patterns (Chapter 2, Section 2.2.2.3). Secondly, the estimation of haplotype similarity requires equivalent information in all samples regardless of sex, which invalidates the inclusion of the X chromosome. This chromosome could be analysed separately for comparison purposes, but its total number of SNPs is too low compared to the number of SNPs in all autosomes, likely producing noisy estimations.

Table 3.1. 117 reference population samples

code	Sample label	Group*	N (Pre-QC)	N (Post-QC)	Country of origin (.sample)	Data source**
1	Pima	NAM	2	2	Mexico.1	1
2	Nahua	NAM	25	25	Mexico.2	This study
3	Mixe	NAM	2	2	Mexico.3	1
4	Mixe.B	NAM	16	16	Mexico.4	This study
5	Mixtec	NAM	2	2	Mexico.5	1
6	Mixtec.B	NAM	10	10	Mexico.6	This study
7	Zapotec	NAM	2	2	Mexico.7	1
8	Zapotec.B	NAM	12	12	Mexico.8	This study
9	Mayan	NAM	2	2	Mexico.9	1
10	Kaqchikel	NAM	8	8	Guatemala	This study
11	Cabecar	NAM	5	5	Costa.Rica.1	This study
12	Guaymi	NAM	4	4	Costa.Rica.2	This study
13	Embera	NAM	21	21	Colombia.1	This study
14	Waunana	NAM	5	5	Colombia.2	This study
15	Wayuu	NAM	3	3	Colombia.3	This study
16	Kogi	NAM	6	6	Colombia.4	This study
17	Zenu	NAM	7	7	Colombia.5	This study
18	Piapoco	NAM	2	2	Colombia.6	1
19	Ticuna	NAM	4	4	Colombia.7	This study
20	Inga	NAM	3	3	Colombia.8	This study
21	Karitiana	NAM	3	3	Brazil.1	1
22	Surui	NAM	2	2	Brazil.2	1
23	Xavante	NAM	4	4	Brazil.3	This study
24	Andoa	NAM	20	20	Peru.1	This study
25	Aymara.A	NAM	13	13	Bolivia.1	This study
26	Aymara.B	NAM	4	4	Chile.1	This study
27	Quechua	NAM	3	3	Peru.2	1
28	Quechua.B	NAM	14	14	Bolivia.2	This study
29	Uros	NAM	8	8	Peru.3	This study
30	Colla	NAM	25	23	Argentina.1	2
31	Wichi	NAM	25	19	Argentina.3	2
32	Wichi.B	NAM	4	4	Argentina.4	This study
33	Toba	NAM	4	4	Argentina.5	This study
34	Ache	NAM	5	5	Paraguay	This study
35	Guarani	NAM	5	5	Argentina.6	This study
36	Chane	NAM	2	2	Argentina.7	This study
37	Mapuche	NAM	9	9	Argentina.2	This study
38	Huilliche	NAM	10	10	Chile.2	This study
39	PRT.A	EUR	18	18	Portugal.1	This study
40	PRT.B	EUR	31	31	Portugal.2	This study
41	IBS-Galicia	EUR	12	8	Spain.1	3
42	SP-CAN	EUR	14	14	Spain.2	This study

43	IBS-Canarias	EUR	3	2	Spain.3	3
44	SP-AND	EUR	15	15	Spain.4	This study
45	IBS-Andalucia	EUR	4	4	Spain.5	3
46	IBS-Extremadura	EUR	12	8	Spain.6	3
47	IBS	EUR	7	7	Spain.7	3
48	SP-CSP	EUR	15	15	Spain.8	This study
49	IBS-Cast.Leon	EUR	18	12	Spain.9	3
50	IBS-Cast.Mancha	EUR	9	6	Spain.10	3
51	IBS-Murcia	EUR	12	8	Spain.11	3
52	IBS-Valencia	EUR	21	14	Spain.12	3
53	IBS-Aragon	EUR	6	6	Spain.13	3
54	SP-CTL	EUR	7	7	Spain.14	This study
55	IBS-Cataluna	EUR	15	10	Spain.15	3
56	IBS-Baleares	EUR	12	8	Spain.16	3
57	IBS-Cantabria	EUR	9	6	Spain.17	3
58	SP-BAS	EUR	14	14	Spain.18	This study
59	IBS-Pais.Vasco	EUR	12	8	Spain.19	3
60	Basque	EUR	2	2	France.1	1
61	French	EUR	3	3	France.2	1
62	Bergamo	EUR	2	2	Italy.1	1
63	Sardinian	EUR	3	3	Italy.2	1
64	TSI	EUR	107	106	Italy.3	3
65	Tuscan	EUR	2	2	Italy.4	1
66	CEU	EUR	99	91	NW.Europe	3
67	GBR-Kent	EUR	38	31	UK.1	3
68	GBR-Cornwall	EUR	32	29	UK.2	3
69	GBR-Corn-Devon	EUR	1	1	UK.3	3
70	GBR-Scotland	EUR	4	3	UK.4	3
71	Orcadian	EUR	2	2	UK.5	1
72	GBR-Orkney	EUR	26	21	UK.6	3
73	Bulgarian	EUR	2	2	Bulgaria	1
74	Hungarian	EUR	2	2	Hungary	1
75	Greek	EUR	2	2	Greece.1	1
76	Crete	EUR	2	2	Greece.2	1
77	Georgian	EUR	2	2	Georgia	1
78	FIN	EUR	99	99	Finland	3
79	Estonian	EUR	2	2	Estonia	1
80	Russian	EUR	2	2	Russia	1
81	MRC	ESM	14	11	Morocco.1	This study
82	Moroccan_Jew[#]	ESM	7	7	Morocco.2	This study
83	TUN	ESM	14	14	Tunisia.1	This study
84	Tunisian_Jew[#]	ESM	6	6	Tunisia.2	This study
85	LIB	ESM	15	14	Libya.1	This study
86	Libyan_Jew[#]	ESM	7	7	Libya.2	This study
87	JRD	ESM	15	15	Jordan.1	This study
88	Sephardi_Jew[#]	ESM	7	7	Turkey.1	This study

89	Turkish	ESM	2	2	Turkey.2	1
90	BedouinB	ESM	2	2	Israel.1	1
91	Druze	ESM	2	2	Israel.2	1
92	Iraqi_Jew	ESM	2	2	Iraq	1
93	Jordanian	ESM	3	3	Jordan.2	1
94	Palestinian	ESM	3	3	Palestine	1
95	Yemenite_Jew	ESM	2	2	Yemen	1
96	YRI	SSA	108	101	Nigeria.1	3
97	ESN	SSA	99	95	Nigeria.2	3
98	GWD	SSA	113	111	Gambia	3
99	MSL	SSA	85	69	Sierra.Leone	3
100	LWK	SSA	99	73	Kenya	3
101	Anuak	SSA	21	3	Ethiopia	4
102	South_Sudanese	SSA	21	8	South.Sudan	4
103	GuiGhanaKgal	SSA	15	14	Botswana	5
104	Juhoansi	SSA	18	15	Namibia.1	5
105	Karretjie	SSA	20	3	South.Africa.1	5
106	Khomani	SSA	39	4	South.Africa.2	5
107	Khwe	SSA	17	14	Namibia.2	5
108	SEBantu	SSA	20	19	South.Africa.3	5
109	SWBantu	SSA	12	9	Namibia.3	5
110	Xun	SSA	19	19	Angola	5
111	KHV	EAS	99	95	Vietnam	3
112	CDX	EAS	93	82	China.1	3
113	CHS-Hu_Nan	EAS	102	66	China.2	3
114	CHS-Fu_Jian	EAS	48	31	China.3	3
115	CHB	EAS	103	101	China.4	3
116	Korean	EAS	2	2	Korea	1
117	JPT	EAS	104	104	Japan	3
	TOTAL		2359	2058		

(N=Sample Size)

*NAM: Native American, EUR: European, ESM: East/South Mediterranean, SSA: Sub-Saharan African and EAS: East Asian. # Samples obtained from The National Laboratory for the Genetics of Israeli Populations (http://yoran.tau.ac.il/nlgip/).

**References: 1: (Mallick et al. 2016), 2: (Eichstaedt et al. 2014), 3: (1000 Genomes Project et al. 2015), 4: (Pagani et al. 2012), 5: (Schlebusch et al. 2012). Genotypes at SNPs shared between published datasets were reported to have been obtained by full genome sequencing (1) or genotyping on the following platforms: Illumina OmniExpress (2), Illumina Omni2.5M (3,5) and Illumina Omni1M (4). Table adapted from Chacón-Duque et al. (2018).

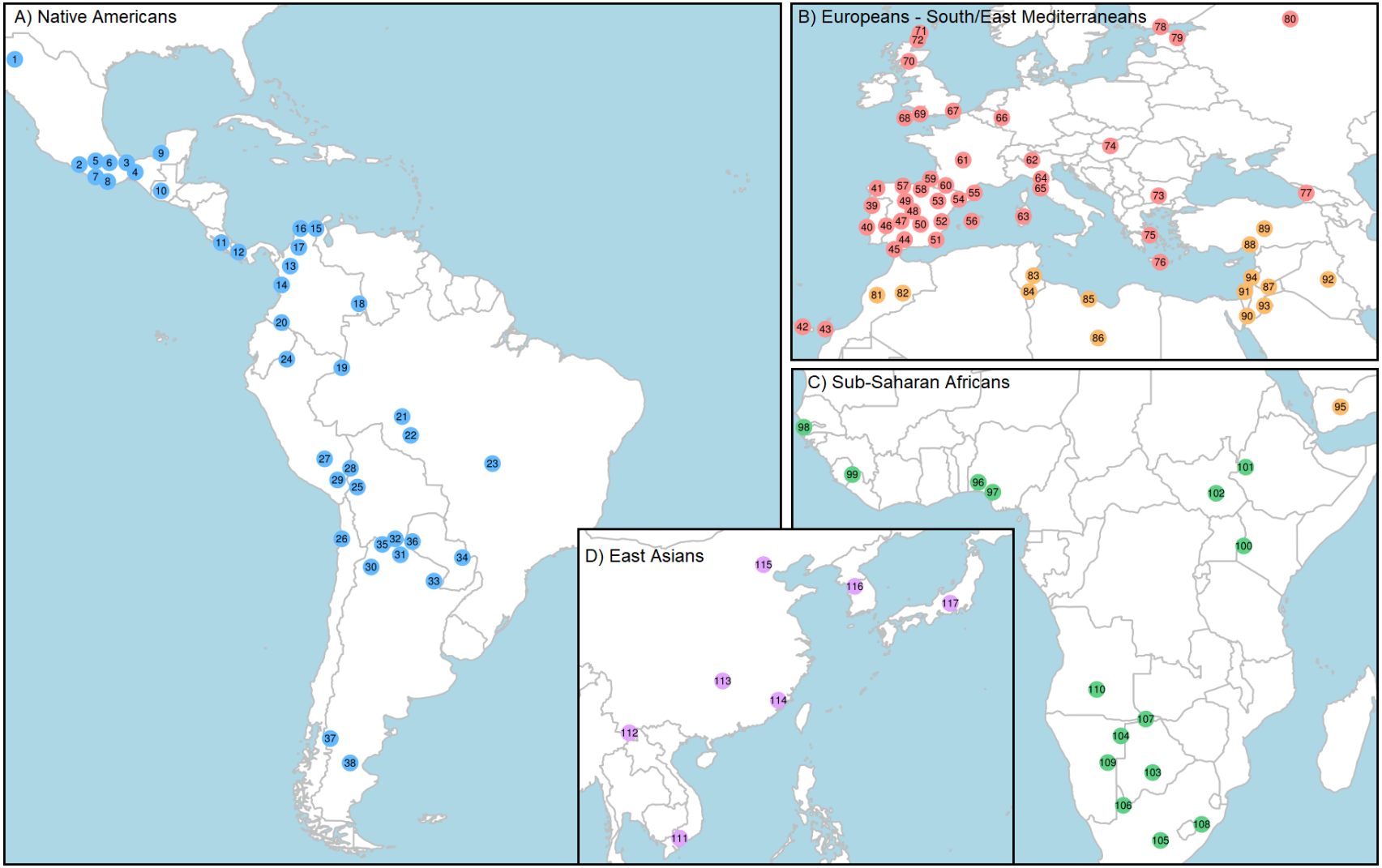

Figure 3.1. Approximate geographic location of the 117 reference populations.
Numbers correspond to those in Table 3.1. Adapted from Chacón-Duque et al. (2018). Generated by JC Chacón-Duque and K Adhikari.

Table 3.2. Number of markers per chromosome contained in the Illumina Human OmniExpress chip

Chr	N	Chr	N	Chr	N	Chr	N	Chr	N
1	59,487	**6**	48,510	**11**	36,831	**16**	22,893	**21**	10,292
2	57,949	**7**	38,317	**12**	35,722	**17**	20,372	**22**	10,531
3	47,430	**8**	37,202	**13**	27,963	**18**	21,800	**X**	18,055
4	40,606	**9**	32,974	**14**	23,436	**19**	15,211	**Y**	1,409
5	42,272	**10**	39,258	**15**	21,776	**20**	18,526	**PAR**	473

N = Number of SNPs

I supported the logistics for the samples' reception and prepared the DNA samples (dilutions, quantifications and experimental quality controls) for chip genotyping. The genotyping was done together with the CANDELA samples in UCL Genomics, a collaborative research facility at UCL.

As described in detail in Adhikari et al. (2016b), we followed the suggested protocol on the Illumina GenomeStudio genotype calls (Guo et al. 2014). All the metrics generated from the GenCall algorithm were analysed and SNPs with low GenTrain score (<0.7), low cluster separation score (<0.3) or high heterozygosity values (>0.5) were excluded. In order to account for batch and plate effects during the genotyping, we included repeated samples in a large number of plates (at random positions) to check for consistency across plates (Pompanon et al. 2005). We used several of these "control" samples and in all cases the genotypes were consistent across all plates (consistency rate ≥0.9999). Furthermore, we looked at SNP call rates and allele frequencies across batches to check for any inconsistencies and did PCA of samples annotated by batches to see if any pattern was detected. Finally, to check a possible effect of the DNA quality, a set of control samples was re-genotyped on different plates after 2x and 4x dilutions and the genotype consistency was ≥0.9996.

3.3 Exploratory analyses and quality controls in the combined reference populations + CANDELA dataset

Prior to merging the datasets (including CANDELA), quality controls were applied, primarily using PLINK v1.9. Since many analyses were done jointly with CANDELA, I describe QC done on these samples where appropriate. First, SNPs

and individuals with >5% missing data, and SNPs with minor allele frequency (MAF) <1% were excluded, following standard protocols for Genome-Wide data (Anderson et al. 2010). A considerable amount of missing data in an individual can evidence low quality DNA, while markers with missing data can be indicative of low quality for genotyping and problems with the genotype-calling algorithms. Similarly, in the case of MAF-based exclusions, it is suggested that small amounts of heterozygote and rare-homozygote clusters for a given marker can affect the performance of genotype-calling algorithms (Anderson et al. 2010).

An X chromosome concordance check was done by comparing sex assignments in the dataset with X chromosome inbreeding coefficients with PLINK v1.9 (option --check-sex). The following quality controls were also applied to all the datasets, but the Native American samples excluded were kept in the merged data and were used for the phasing as to keep the highest possible number of Native American segments, but were not used in any of the subsequent analyses.

Offspring were removed from the trios collected in 1KGP. Cryptic relatedness was assessed by estimating PI_HAT (the IBD test implemented in PLINK v1.9; see Chapter 2, Section 2.2.1.2) in each dataset after doing LD-based pruning to account for the fact that this test is not LD-sensitive. Case-control association studies usually consider that two samples are unrelated when the maximum relatedness is less than that of a second-degree relative (equivalent to a PI_HAT <0.25). Anderson et al. (2010) proposed a threshold of 0.1875, which is half-way between second and third-degree relatives, considering that other factors like LD, population structure and genotyping errors can affect the estimation.

Moreover, it also is essential to take into consideration the ascertainment bias introduced while designing the chips. In this dataset, this is especially important for Native American populations, which are underrepresented in most chip designs including Illumina Omni platforms, and whose high levels of genetic drift can also inflate PI_HAT estimates. While in European populations we observe that the median PI_HAT is close to zero, In Native Americans we observe median values close to the values for second and third-degree relatives (data not shown). Assuming that the median PI_HAT can be seen as a baseline value, we propose to establish a different PI_HAT threshold for every population, removing individuals with PI_HAT>0.1 above the median PI_HAT for that group. The 10% addition

to the baseline allows us to be conservative after accounting for some variability in the estimates. For instance, for the population Kogi the median PI_HAT is 0.21 (Table 3.3), which means that only individuals with PI_HAT values >0.31 were excluded for this population (in this case, Kogi1 and Kogi4).

Table 3.3. PI_HAT estimates for the reference population Kogi

ID*	Kogi1	Kogi2	Kogi3	Kogi4	Kogi5	Kogi6
Kogi1	0	0.5624	0.3224	0.223	0.2165	0.2143
Kogi2	0.5624	0	0.2958	0.2412	0.2095	0.2132
Kogi3	0.3224	0.2958	0	0.1987	0.2011	0.2075
Kogi4	0.223	0.2412	0.1987	0	0.3333	0.2269
Kogi5	0.2165	0.2095	0.2011	0.3333	0	0.2222
Kogi6	0.2143	0.2132	0.2075	0.2269	0.2222	0

*These are not the real sample IDs.

In addition, to control for possible recent admixture that could potentially confound the clustering, an unsupervised ADMIXTURE analysis was performed (using the LD-pruned data described above) to identify and exclude Native Americans, Sub-Saharan Africans, East Asians and Europeans with less than 95% of their own continental ancestry. In the case of East/South Mediterranean individuals, ADMIXTURE consistently inferred a mixture of European and Sub-Saharan African ancestry. The estimated proportions of both components were found to be homogeneous across individuals within populations, probably indicating the detection of an old admixture event (Jobling et al. 2014) which might not affect the detection of fine-scale genetic structure. Based on this assumption, I excluded four individuals with admixture proportions deviating markedly from those observed in the population sample, suggestive of recent admixture (three Moroccans with Sub-Saharan African ancestry >40% and one Libyan with Sub-Saharan African ancestry of 79%; both of these populations have an estimated average Sub-Saharan African ancestry of ~20±3%). Additionally, for the CANDELA dataset, individuals sampled in a given country but born outside it were relocated when coming from one of the five countries included in this study or otherwise removed.

Finally, prior to the merging, I evaluated and corrected flipped strands using the reference build hg19/b37 and removed palindromic SNPs (SNPs were their al-

leles are both purines or pyrimidines), because the strand flipping issues are undetectable for such SNPs, as their alleles are complimentary and potentially create merging and alignment problems when compared to other datasets.

After Quality Control, the merged CANDELA + reference population dataset comprised genotypes for 546,780 autosomal SNPs in 8,647 individuals (including 6,589 Latin Americans and 2,058 individuals from the reference population samples).

3.4 Selection of reference samples from CANDELA

Given the lack of representation of some Italian and German populations (important sources of migration to Latin America, see Chapter 1, Section 1.2.4) in our reference samples, and the low overall numbers of Native Americans, we decided to include as reference samples individuals from CANDELA with considerably high levels of European or Native American ancestry.

Following the same ADMIXTURE analyses described in the previous section, we found 52 individuals with >99% European ancestry in the Brazilian sample, 37 reporting full German and 15 full Italian ancestry through records of native language spoken by their grandparents. Clustering analyses (Sections 3.7 and 3.8) corroborated their resemblance to the respective countries or regions.

In addition, there were 1 Colombian, 22 Mexicans, 65 Chileans and 17 Peruvians with >95% Native American ancestry. There is additional information for 30 of these individuals on the specific Native American populations they or their ancestors belong to (especially in Mexico), and for the others detailed information on their geographic places of origin is provided. PCA and haplotype-based clustering analyses are generally consistent with this information (Sections 3.7 and 3.8). In Chapter 5 (Section 5.3.2.7), I illustrate how our inferred ancestry results change (though largely remain consistent) if I instead remove these CANDELA individuals from the reference set.

3.5 Phasing

Phasing of the whole merged dataset was performed with SHAPEIT2 using default parameters. Genetic distances used were obtained from the HapMap Phase II genetic map build GRCh37 (International HapMap et al. 2007). Missing SNPs (following QC, only individuals with <5% missingness remained) were imputed during the phasing process.

3.6 Inference of haplotype similarity profiles between individuals

I set up CHROMOPAINTER to provide estimates of the proportion of DNA in every reference population individual (recipients) that is most closely related to each other reference population individual (donors). The software automatically excludes the recipient individual being painted from the donors, hence reconstructing haplotype similarity profiles for every individual in terms of the others. This procedure creates a squared coancestry matrix, required as input for fineSTRUCTURE, containing the number of haplotype segments for which each individual is inferred to share most recent ancestry with each other individual (for details, see Chapter 2). This set of reference populations included the CANDELA individuals selected in the previous section to be included in the reference.

The recombination scaling constant N_e and mutation parameter θ used by CHROMOPAINTER were jointly estimated for every individual in a subset of chromosomes (1, 6, 13 and 22) with ten Expectation-Maximization steps, starting from default values defined by the software. The weight-averaged N_e and θ values across chromosomes (weighted by each chromosome's SNP count) were then used for subsequent CHROMOPAINTER runs on all autosomes (N_e = 290.83 and θ = 0.00038). N_e is an analogous parameter to effective population size, and θ is an estimator of population mutation rate, similar to the Watterson estimator (Watterson 1975). The genetic distances were interpolated for every SNP based on the HapMap Phase II genetic map build GRCh37.

3.7 Definition of clusters of reference population individuals

I have implemented a set of analyses based on previous methodologies developed by Leslie et al. (2015), Hellenthal et al. (2014), and Lawson et al. (2012), which aim to generate a clustering of genetically homogeneous groups to be used as surrogates for the ancestral populations involved in the genetic make-up of Latin America. These analyses take advantage of the increased resolution provided by haplotype-based ancestry inference and also facilitate the interpretation of the sub-continental ancestry estimation on the admixed individuals.

I first used the software fineSTRUCTURE to explore the fine-grained genetic structure of the reference populations. Further analyses selected a subset of fineSTRUCTURE clusters to be used as surrogates for the ancestral populations when analysing admixed CANDELA individuals. Essentially, this selection process excludes individuals that are inconsistently assigned to different clusters through the iterations of the MCMC procedure, clusters that do not contribute significantly to the admixed Latin American populations, and clusters with complex demographic histories whose contributions to the admixture can be difficult to interpret. A general picture of the procedure is described below, and the Appendix has a description of this selection process broken down by cluster.

3.7.1 fineSTRUCTURE analysis

I used fineSTRUCTURE to evaluate genetic structure in the reference data, independent of population sampling labels and using haplotype similarity patterns. Using the procedure described in the fineSTRUCTURE instructions, I estimated an adjustment factor *c* of 0.236, which accounts for (incorrect) assumption that the amount of DNA matching among individuals is independent. Two MCMC runs were performed, each using 2,000,000 iterations (sampling every 10,000). Following Leslie et al. (2015), for each run the sample with maximum posterior probability was selected and an additional 100,000 hill-climbing moves were then performed to search for merges or splits that further improve the overall model likelihood (Lawson et al. 2012). After this procedure, fineSTRUCTURE classified individuals into 129 clusters.

In general, the clusters closely match geographic, linguistic and/or historical reported relationships between populations (Appendix), and the resolution is higher to that provided by PCA (Section 3.8). Some populations, especially Native American and African groups, are usually divided into several clusters. This is stratification within populations can be caused by the detection of different levels of genetic drift in the same population. Genetic drift causes an increase on the amount of haplotype similarity among individuals within the same group as the likelihood of finding a common recent ancestor in other population; the more drifted an individual is, the higher their self-copying.

We also tried a procedure that builds a “tree” by merging pairs of genetically similar clusters (one pair at a time until only two remain) under a greedy algorithm described in Lawson et al. (2012), that was successfully applied to study the fine structure of the populations in Great Britain (Leslie et al. 2015). However, for this specific analysis cutting the tree at different levels does not seem to be the best choice, as the distances on the tree branches relate to changes in the posterior probability of the fineSTRUCTURE model and are not directly related to time or measurements of genetic distance (Leslie et al. 2015), and several factors, like big differences in samples size between populations can also complicate the interpretation (Lawson et al. 2012). Perhaps for these reasons, even though most of the 129 clusters are considerably close in the hierarchy of the tree to populations that are also close geographically, some of them are positioned within the tree with other clusters when there is no evidence of clear relationships (Figure 3.2, e.g. Mayan clustering next to Mapuche).

In order to reduce the number of clusters potentially representing sources of ancestry in Latin America, to avoid problems related to colinearity between different surrogate sources when estimating ancestry (as reference populations with close haplotype similarity profiles are often indistinguishable in the regression models, hence increasing the uncertainty of the estimations), and to support the interpretation of results, I performed the further refinements described in the next section.

3.7.2 Additional steps to refine the clustering

I carried out additional analyses in order to evaluate the robustness of the clusters, in a series of steps that culminated in a re-classification of these 129 clusters

into 117 “donor clusters”, a subset of which were used as 56 “surrogate clusters” for inferring sub-continental ancestry in CANDELA individuals as described in Chapter 2 (Section 2.4.2). In general, I made an effort to maintain a wide range of European and Native American groups, these being the two highest contributing continental groups to the genetic make-up of our sample. This re-classification process was as follows:

First, I checked the consistency of the assignments of every individual into a given cluster. Contrasting the results of the two fineSTRUCTURE runs, I removed individuals that were assigned to a different cluster more than 10% of the time across samples in the last 1,000,000 iterations of the two runs, and five clusters where all individuals were inconsistent across these samples. I also extracted twelve individuals assigned to their own unique clusters, and ten small clusters made of either a small number of individuals from distant populations or from populations present in other clusters with greater numbers.

Next, I used the remaining clusters (i.e. those not set aside above) to perform an initial estimation of sub-continental ancestry in the CANDELA samples using a modification of the Non-Negative Least Squares (NNLS) regression approach described in Leslie et al. (2015) and Hellenthal et al. (2014). Based on these analysis, I excluded additional individuals from 17 clusters that contributed to no CANDELA samples. Furthermore, based on the tree inferred by fineSTRUCTURE and on Total Variation Distance (TVD) (e.g. as used in Leslie et al. (2015)), I merged 29 remaining clusters that were difficult to distinguish from one another into 13 groups. After these steps, there were 69 clusters remaining intact from the original 129.

I next took all individuals that had been excluded as described above and reclassified them into 48 clusters based on population label information. This gave me the 117 “donor clusters” that we use throughout. The Appendix lists how individuals from the 129 fineSTRUCTURE clusters were classified into the 117 donor clusters used.

I then performed a few additional steps to define the final 56 “surrogate clusters”, starting from the 69 “intact” clusters described above, using the modified NNLS regression approach. In particular I checked if closely related clusters could potentially contribute to colinearity issues in subsequent analyses or if they had

complex ancestry profiles that could eventually complicate the interpretation of the results. To perform this regression analysis, the proportions of DNA that each individual from the 69 clusters matches to each donor as estimated by CHROMOPAINTER were summed across donors within each of the 117 donor groups defined above.

For each individual from the 69 clusters, this produces a vector of 117 variables that we call a “copying vector” (see Chapter 2), with each variable the proportion of DNA that this individual copies from (i.e. matches to) all individuals contained in that donor group. For each of the 69 clusters, I averaged these copying vectors across all individuals assigned to that cluster, creating a unique copying vector for each of the 69 clusters. Then for each of these 69 clusters, I performed a NNLS regression with the copying vector of that cluster as the response and the copying vectors for all 68 other clusters as predictors.

From these analyses, seven clusters (whose individuals belong to the Native American populations *Uros*, *Kogi*, *Karitiana*, *Surui*, *Ticuna* and *Mixe*) with considerable levels of genetic drift and no contributions to the CANDELA samples were excluded from the surrogate clusters and were also removed from the donors for subsequent analyses. Given their considerable amounts of genetic drift (as indicated by high values of self-copying), we initially tried to use them only as recipients (painted against all the other populations except their own, in the same way the admixed individuals are painted) and surrogates, considering the possibility that by removing their self-copying from their haplotype similarity profiles, their contributions to the Latin American populations could be inferred successfully. However, this approach did not work (data not shown) and I decided to keep them out of the ancestry inference.

Also, six clusters showing complex signals in NNLS analyses were excluded based on the following criteria: (*i*) the cluster contributed to the ancestry profiles of several surrogate groups of interest and (*ii*) the cluster showed ancestry from more than two continental groups. For instance, in the case of (*i*) we excluded *Sardinia* as it was contributing high amounts (~15%) to the ancestry of *Portugal/WestSpain*, *Catalonia* and *Italy*. The Sardinian population, a well-known genetic isolate, could be acting as a surrogate for the ancestors of different populations, possibly through a preserved Neolithic farmers-like ancestry (Chiang et al.

2016; Haak et al. 2015; Olivieri et al. 2017). The best example for (*ii*) is *Turkey*, which was inferred to have more than 5% ancestry from an East Asian source, as well as 5% from a European one. Turkey is strategically located between Europe, the Middle East and Asia, and previous studies on the genetic structure of Turkey have demonstrated an overlap between Turks and Middle Eastern with considerable affinity with European and South / Central Asian populations (Hodoğlugil and Mahley 2012).

These additional analyses resulted in these 69 clusters being reduced to a final list of 56 “surrogate clusters” that are made of 1,444 individuals from the reference panel (Table 3.1). Table 3.4 details the individual makeup of the 56 surrogate clusters, in terms of the original population sample labels, and Figure 3.2 shows a tree relating them based on the distances calculated by fineSTRUCTURE.

Table 3.4. Individual makeup of the 56 clusters defined by fineSTRUCTURE and used for ancestry analysis in CANDELA

Cluster	Size	Includes
1	10	Ethiopia(3/3), South.Sudan(7/8)
2	73	Kenya(73/73)
3	6	Namibia.3(6/9)
4	18	South.Africa.3(18/19)
5	51	Gambia(51/111)
6	68	Sierra.Leone(68/69)
7	99	Nigeria.1(99/101)
8	9	Jordan.1(7/15), Yemen(2/2)
9	7	Jordan.1(1/15), Jordan.2(3/3), Palestine(3/3)
10	7	Morocco.2(7/7)
11	8	Libya.2(1/7), Turkey.1(7/7)
12	12	Tunisia.2(6/6), Libya.2(6/7)
13	28	Tunisia.1(14/14), Libya.1(14/14)
14	11	Morocco.1(11/11)
15	48	Spain.2(1/14), Spain.4(13/15), Spain.5(3/4), Spain.6(4/8), Spain.7(4/7), Spain.9(9/12), Spain.10(3/6), Spain.11(5/8), Spain.12(5/14), Spain.14(1/7)
16	18	Spain.8(1/15), Spain.10(2/6), Spain.12(5/14), Spain.13(5/6), Spain.17(5/6)
17	29	Spain.7(3/7), Spain.12(2/14), Spain.13(1/6), Spain.14(6/7), Spain.15(10/10), Spain.16(7/8)
18	18	Spain.2(13/14), Spain.3(2/2), Spain.6(1/8), Spain.11(2/8)
19	53	Portugal.1(18/18),Portugal.2(31/31), Spain.1(4/8)
20	24	Spain.18(14/14), Spain.19(8/8), France.1(2/2)
21	19	Italy.5*(15/15), Italy.1(2/2), Bulgaria(2/2)
22	31	Italy.3(31/106)

23	101	NW.Europe(68/91), UK.1(31/31), UK.2(1/29), UK.3(1/1)
24	31	Germany*(31/37)
25	2	Russia(2/2)
26	9	Finland(7/99), Estonia(2/2)
27	92	Finland (92/99)
28	72	China.1(72/82)
29	91	Vietnam(91/95)
30	64	China.4(64/101)
31	103	Japan(103/104)
32	56	Chile.1(1/3), Bolivia.2(8/12), Chile.3*(47/65)
33	16	Bolivia.1(8/12), Peru.4*(6/17), Peru.2(2/3)
34	10	Argentina.1(10/19)
35	9	Peru.4*(8/17), Peru.2(1/3)
36	3	Colombia.1(2/16), Colombia.2(1/3)
37	3	Costa.Rica.2(3/3)
38	4	Costa.Rica.1(4/4)
39	4	Colombia.5(4/4)
40	2	Colombia.3(2/2)
41	14	Colombia.1(12/16), Colombia.2(2/3)
42	3	Peru.1(1/13), Peru.4*(2/17)
43	5	Chile.3*(1/65), Argentina.2(2/2), Chile.2(2/2)
44	7	Mexico.9(2/2), Guatemala(5/5)
45	4	Paraguay(4/4)
46	2	Colombia.6(2/2)
47	6	Peru.1(6/13)
48	5	Argentina.6(3/5), Argentina.7(2/2)
49	2	Mexico.1(2/2)
50	9	Mexico.2(2/20), Mexico.10*(7/22)
51	7	Mexico.6(7/8)
52	6	Mexico.8(6/8)
53	16	Mexico.10*(13/22), Mexico.6(1/8), Mexico.8(2/8)
54	19	Mexico.2(18/20), Mexico.10*(1/22)
55	2	Mexico.3(2/2)
56	18	Argentina.3(13/13), Argentina.4(2/2), Argentina.5(3/3)

A tree relating these clusters is shown in Figure 3.2.

*Additional populations were extracted from CANDELA data. Italy.5: Brazilians of Italian descent, Germany: Brazilians of German descent, Chile.3: Native Americans in Chile, Mexico.10: Native Americans in Mexico, Peru.4: Native Americans in Peru. Details of the selection process can be found in Methods. Table adapted from Chacón-Duque et al. (2018).

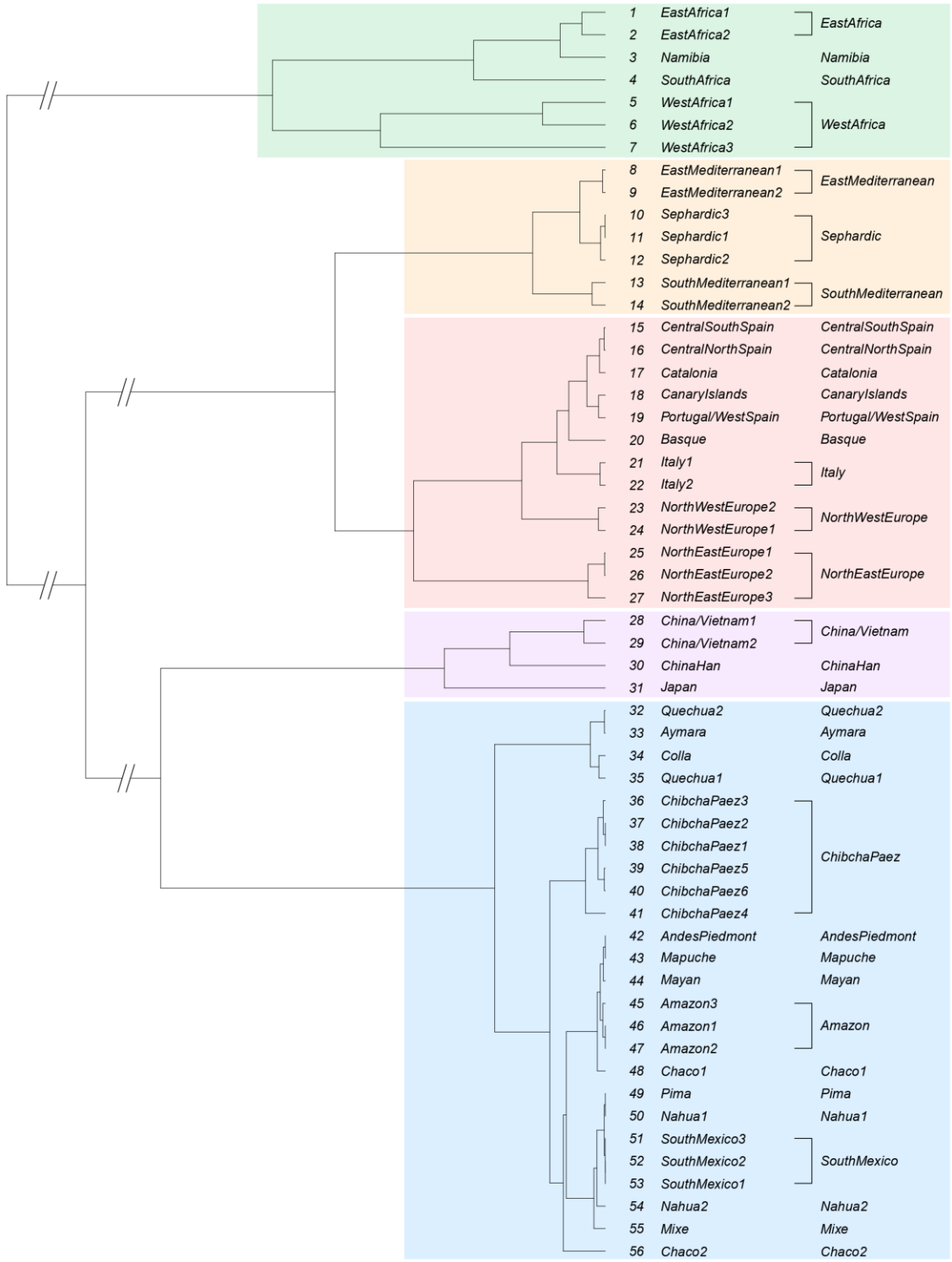

Figure 3.2. Tree topology relating the final 56 clusters which were retained for ancestry analysis of the CANDELA individuals.
Brackets on the right highlight the 35 groups of clusters that were defined for the graphical representations. Table 3.4 provides detailed information of every cluster. Adapted from Chacón-Duque et al. (2018). Generated by JC Chacón-Duque and K Adhikari. The scale on the tree corresponds to the posterior probabilities of the MCMC clustering model. They do not directly reflect scales of time or genetic distance.

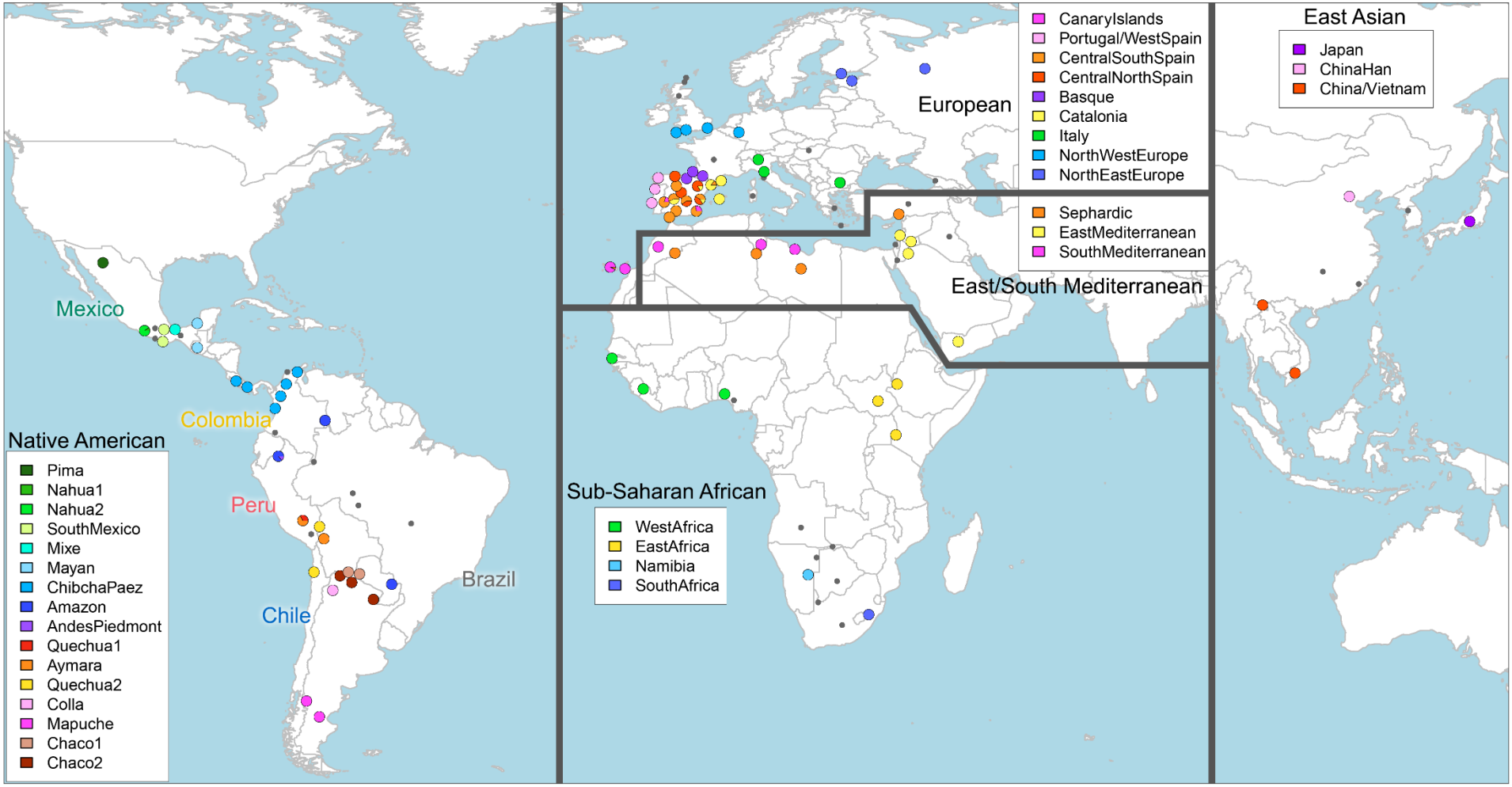

Figure 3.3. Geographic location of the 35 groups of clusters as defined in Figure 3.2.
Reference populations obtained from CANDELA are not included in this map. Grey dots represent reference populations not included in the surrogate groups. Figure adapted from Chacón-Duque et al. (2018). Generated by JC Chacón-Duque and K Adhikari.

All the haplotype-based sub-continental ancestry analyses reported in the remaining chapters of this thesis have been performed using these 56 surrogate clusters. However, for data visualization purposes, an alternative classification encompassing 35 groups of surrogate clusters has been defined by merging subsets of these 56 clusters as shown in Figures 3.2. The map in figure 3.3 shows the geographic location of every of the 35 groups excluding the CANDELA samples used as references.

3.8 Frequency-allele-based approaches for clustering

Standard approaches were also implemented in order to check the consistency of our results and to establish the extent of resolution increase achieved. Unsupervised ADMIXTURE analysis (Section 3.8.1) and PCA (Section 3.8.2) are thoroughly described below.

F_{ST} was also estimated using Weir and Cockerham estimation as implemented in PLINK v1.9. The results were clearly affected by the fact that some samples have very small sample sizes making the results hard to explain (data not shown).

3.8.1 ADMIXTURE analysis

Unsupervised ADMIXTURE analyses were performed on the whole dataset after pruning for LD as described above, retaining a total of 150,858 SNPs. Results for the reference populations organized by the groups of clusters described in figure 3.2 from K = 2 to 10 are displayed in Figure 3.4.

Major continental groups are clearly defined at K=3, with East Asia arising at K=4. At this point a few patterns arise, as some populations are not entirely described by the four continental groups. The northern most Native American groups in the sample (located in Mexico) show a consistent yet marginal amount of the East Asian-like component, likely related to the fact that these populations are closer to Asia according to the migration routes used by the initial settlers. We did not include Siberian or Eskimo groups, but I suspect that this small contribution is likely to represent an ancestral component related to these groups.

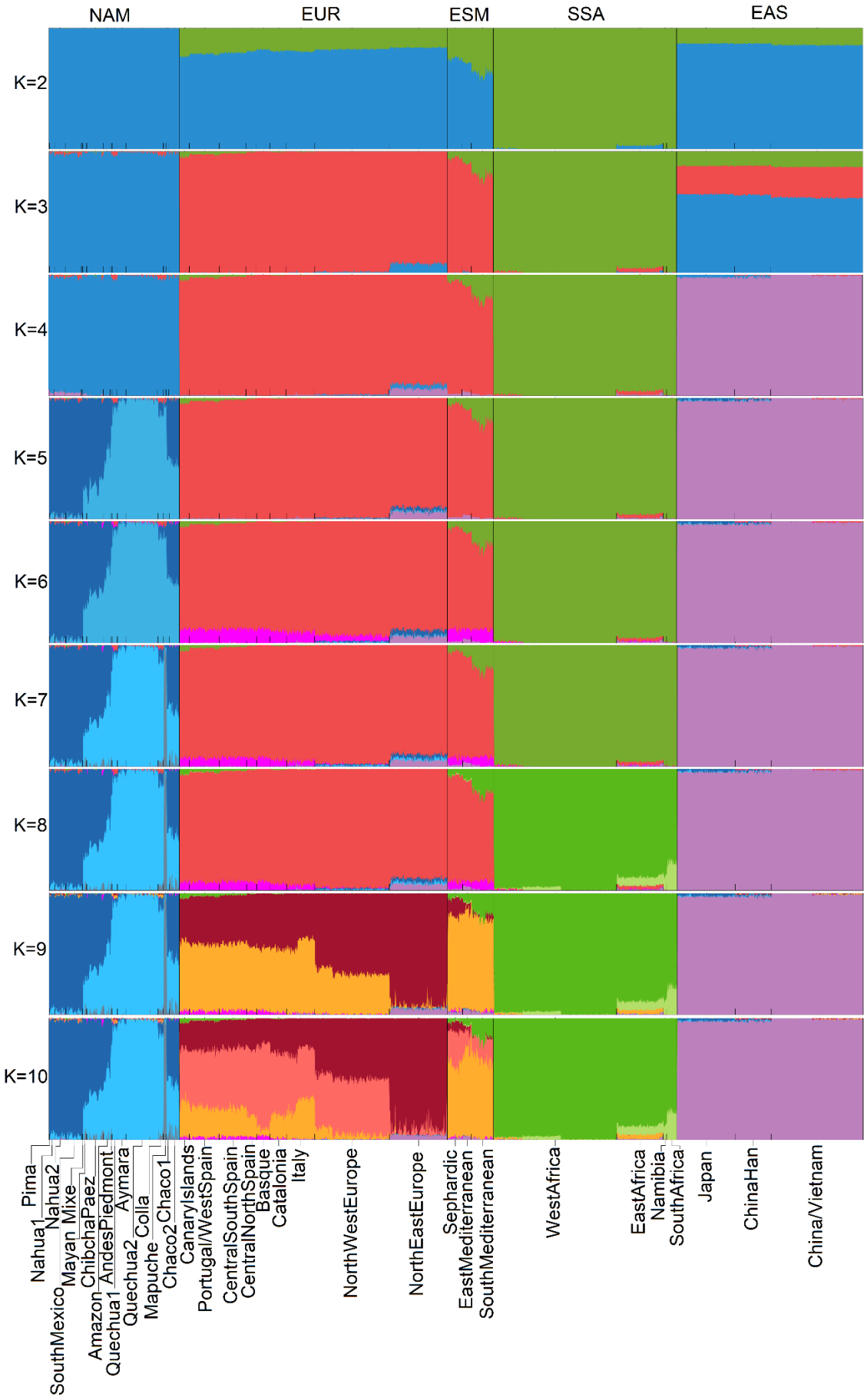

Figure 3.4. Unsupervised ADMIXTURE analysis. Generated by JC Chacón-Duque and K Adhikari.

North Eastern Europeans seem to have a considerable amount of East Asian / Native American ancestry, plausibly related to a North Asian ancestral component e.g. contributed via ancient admixture. East/South Mediterranean individuals show variable degrees of European / Sub-Saharan African ancestry between groups but homogeneous amounts within groups, as described in Section 3.3.

Table 3.5. Unsupervised ADMIXTURE results at different Ks for the Native American groups of clusters which will be presented in the haplotype-based ancestry analyses

	Unsupervised ADMIXTURE*					
	K=3	K=5		K=7		
Cluster group	Native	Native North	Native South	Native North	Native Central Andes	Native SouthChile
Pima	99.8	95.8	2.3	94.4	3.9	0.0
Nahua1	98.7	94.4	4.4	93.5	5.2	0.2
Nahua2	97.3	95.7	1.9	94.8	2.8	0.2
SouthMexico	97.7	95.9	2.4	95.3	2.9	0.3
Mixe	99.0	99.9	0.1	99.2	0.8	0.0
Mayan	99.2	74.1	25.3	72.1	24.1	3.4
ChibchaPaez	99.1	71.9	27.4	62.2	35.2	1.6
Amazon	99.9	46.5	53.5	46.0	52.6	1.3
AndesPiedmont	98.2	21.1	76.8	20.1	72.8	5.2
Quechua1	96.3	4.7	91.3	3.6	92.2	0.0
Aymara	99.5	2.4	97.3	2.5	97.0	0.1
Quechua2	99.2	0.3	99.1	0.2	99.0	0.1
Colla	96.3	10.2	86.0	10.0	82.0	4.5
Mapuche	96.7	2.5	93.7	1.1	2.2	96.7
Chaco1	99.9	49.6	50.3	50.6	47.1	2.2
Chaco2	98.9	52.9	46.0	53.3	40.8	5.2

*Values correspond to the mean ancestry percentages of ADMIXTURE components at those K's and only the components that are related to that specific continental ancestry are displayed.

At *K*=5 the first sub-continental split emerges, with a gradient in Native American populations visible from Mesoamerica to the Andres, reaching its maximum levels

at *Mixe* and *Quechua2* respectively (Table 3.5). *Mayan* and *ChibchaPaez* groups have similar amounts of both ancestries, probably suggesting a more recent common ancestral origin between these two populations; the same happens with *Amazon*, *Chaco1* and *Chaco2*. An additional split between Native American populations also appears, this time separating *Mapuche* from everything else.

Table 3.6. Unsupervised ADMIXTURE results at different Ks for the European and Mediterranean groups of clusters which will be presented in the haplotype-based ancestry analyses

	Unsupervised ADMIXTURE*					
	K=3	K=9		K=10		
Cluster group	Europe	Europe North	Mediterranean	Europe North	Europe Basque	Mediterranean
CanaryIslands	95.2	35.1	57.4	21.4	43.0	29.8
Portugal/ WestSpain	97.3	39.9	55.0	24.5	46.9	24.7
CentralSouth Spain	98.2	39.7	55.5	23.4	49.5	23.6
CentralNorth Spain	99.3	43.3	53.5	24.9	54.8	17.9
Basque	100.0	44.9	51.9	21.1	69.2	6.6
Catalonia	99.3	45.4	52.4	28.2	51.3	19.1
Italy	99.5	40.4	58.3	26.7	43.1	29.7
NorthWestEurope	99.4	65.4	34.2	49.0	45.6	5.1
NorthEastEurope	92.6	91.9	2.4	90.2	5.6	0.3
Sephardic	94.3	14.1	78.2	7.4	27.4	60.2
East Mediterranean	89.4	5.8	81.7	3.8	14.8	72.2
South Mediterranean	79.6	2.5	76.2	0.7	16.0	65.2

*Values correspond to the mean ancestry percentages of ADMIXTURE components at those K's and only the components that are related to that specific continental ancestry are displayed.

A new Latin American component appears at K=6, almost exclusive to Colombia and reaching its highest levels in a population that has long been defined as a genetic isolate, as described extensively in Chapter 5. However, from this point it becomes clear that the ADMIXTURE analysis may not have enough power to

distinguish between recent admixture and genetic drift (van Dorp et al. 2015). African populations become differentiated at K=8. An ancestry gradient can also be seen in European / Mediterranean populations (Table 3.6) at K=9 with a further split at K=10, largely related to the Basque population.

One interesting observation is the high affinity of *CanaryIslands* (conquered by the Kingdom of Castile in 1402) with southern and western Iberian populations. Although some studies have suggested a considerable amount of Guanche (the initial settlers of the island, Berber-like) ancestry in this population (Fregel et al. 2009; Maca-Meyer et al. 2003; Rodríguez-Varela et al. 2017), distinguishing between different Mediterranean contributions is difficult. Given that in the analysis with fineSTRUCTURE they also look similar to other populations, I use them in the rest of this thesis as an Iberian-like source.

3.8.2 Principal Component Analysis

PCs also show similar patterns of differentiation to those found with ADMIXTURE. From PC1 to PC3 continental patterns are defined (Figure 3.5), while PC4 separates the different Sub-Saharan African populations (not shown).

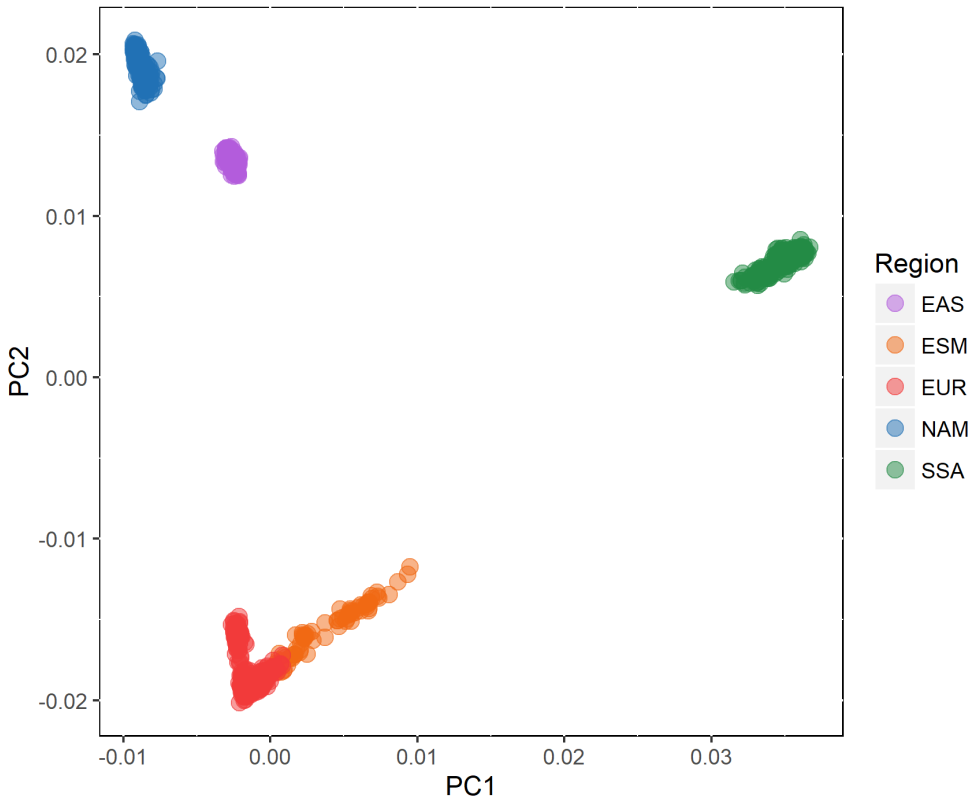

Figure 3.5. Principal component analyses coloured by regions (PC1 vs PC2)

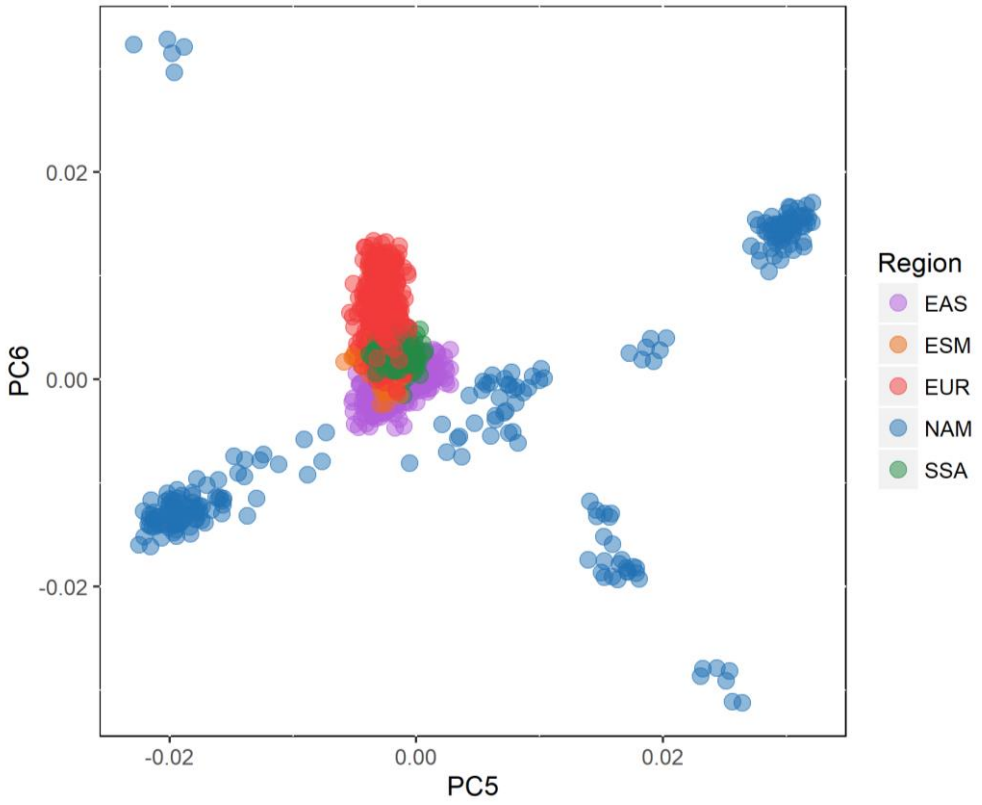

Figure 3.6. Principal component analyses coloured by regions. (PC5 vs. PC6)

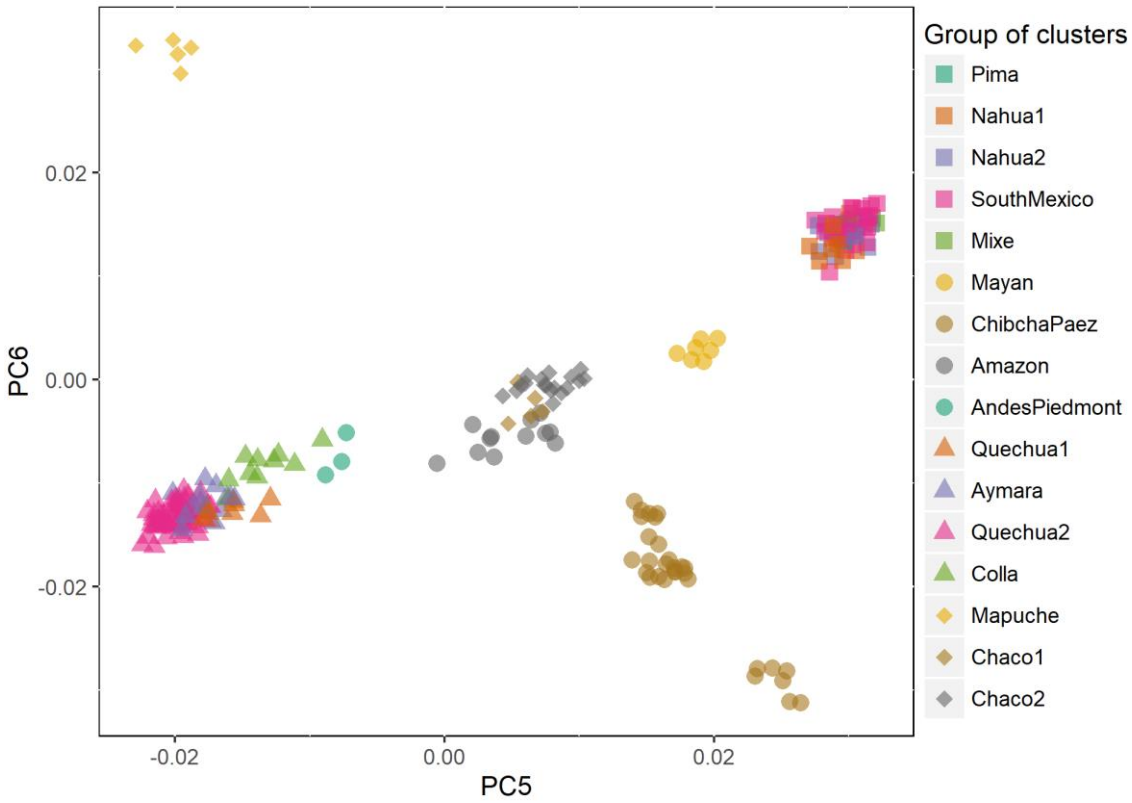

Figure 3.7. Principal component analyses coloured by Native American clusters. (PC5 vs. PC6)

PC5 and PC6 mainly discriminate Native American groups (Figures 3.6 and 3.7). The patterns of differentiation of Native American ancestry seem to be consistent with geography in a similar way as detected by ADMIXTURE. PC5 goes from Mesoamerica to the Andes, while PC6 seem to separate *Mapuche* from all the other populations. PC6 also seems to partially differentiate some European populations.

PC7 and PC8 separates gradients between both Native American and European populations (Figures 3.8 to 3.10) making it complex to interpret from this point.

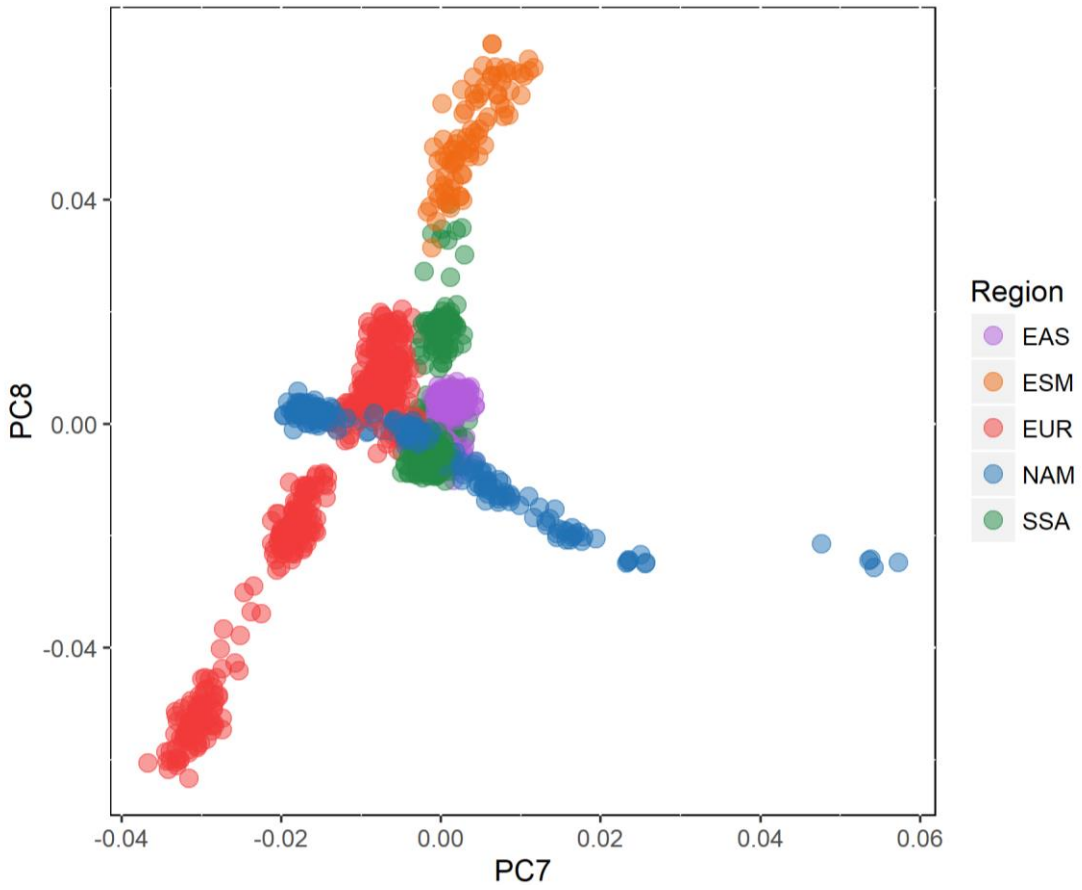

Figure 3.8. Principal component analyses coloured by region. (PC7 vs PC8)

Mapuche seems to be the Native American group that is being separated again from the rest. In the case of European / Mediterranean populations, the gradient again roughly coincides with the one found with ADMIXTURE. However the resolution does not seem to go beyond broad groupings, i.e. North East Europe, North West Europe, Southern Europe (Iberian Peninsula + Italy) and East/South Mediterranean.

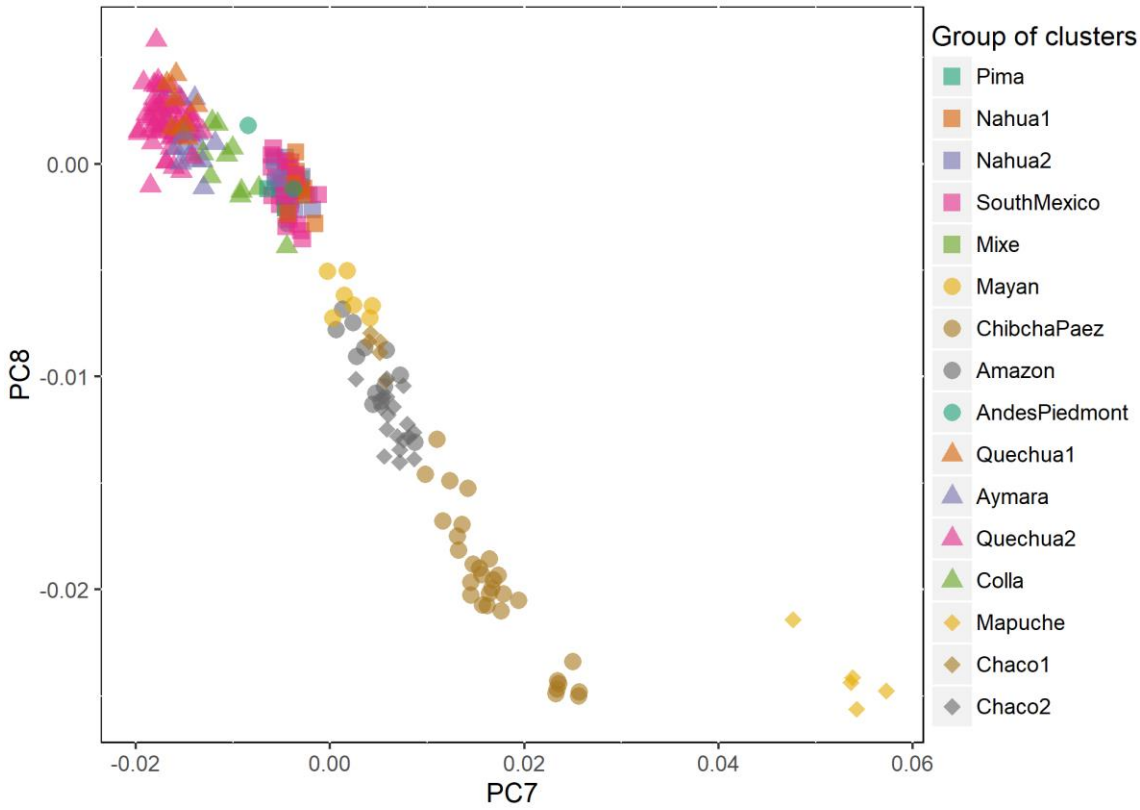

Figure 3.9. Principal component analyses coloured by Native American clusters (PC7 vs PC8)

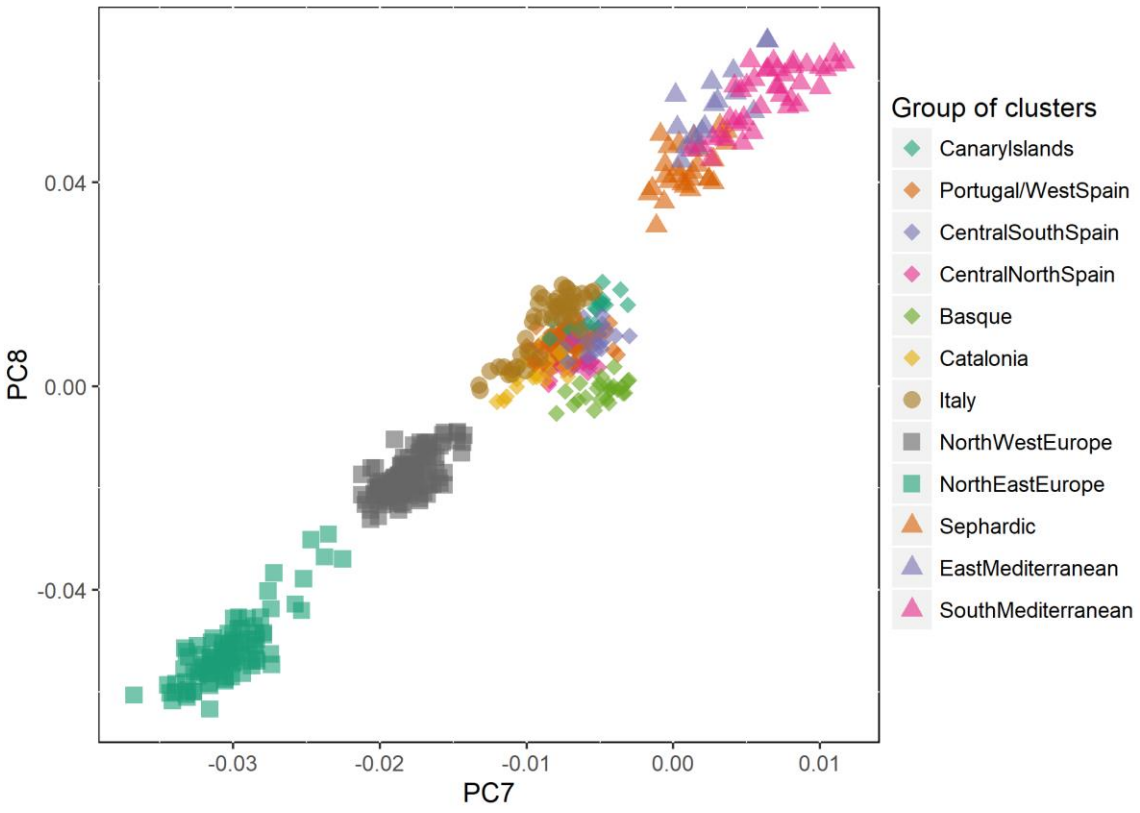

Figure 3.10. Principal component analyses coloured by Native American clusters (PC7 vs PC8)

3.9 Discussion and limitations

The collected dataset aims to represent the sources of admixture of Latin American populations, but there are several limitations that need to be considered. First of all, the selection of the populations is limited by the availability of public data with genotypes for the same Illumina platform or data source that allows a large overlap of SNPs. A large number of SNPs is necessary to take advantage of the power conferred by linkage disequilibrium patterns in the implemented models.

We tried to overcome this by genotyping different reference samples we thought would act as good proxies for ancestral populations in Latin America, and that we could obtain via our research collaborations within funding and time constrains. We further tried to mitigate this by including some CANDELA samples as references, which allowed us to have representatives of populations otherwise absent, such as Germany and Italy (that are main sources of ancestry for the current-day Brazilian populations). Secondly, the samples sizes per population are highly variable (ranging from 1 to 111), which can impact analyses in various ways as described in this section.

Finally, as mentioned before, even if we have sampled individuals from the same geographic locations as the original source populations, it cannot be guaranteed that they will be accurate surrogates. For instance, in the case of the Native Americans, a lot of the original ancestral populations may not exist anymore, as a consequence of the dramatic population collapse occurred during colonial times (Chapter 1, Section 1.2.1). Moreover, the admixed and the source populations could have changed substantially through time after their contact, given the strong bottlenecks and - consequently - the genetic drift the former faced (Koehl and Long 2018), and the additional gene flow they could have all received. All of these limitations need to be considered when interpreting the results generated using these datasets.

The results of this chapter show the increased resolution for detecting population structure provided by haplotype-based methods, confirming what has been suggested in several studies (Busby et al. 2016; Hellenthal et al. 2014; Kerminen et al. 2017; Lawson et al. 2012; Leslie et al. 2015; Markus et al. 2014; Montinaro et al. 2015; van Dorp et al. 2015). In particular the allele-frequency-based clustering from ADMIXTURE can only distinguish broad patterns of population structure

compared to the haplotype-based model fineSTRUCTURE. Additionally, ADMIXTURE analysis can be hard to interpret, as different factors such as admixture, genetic drift and sample size can affect the way the components are segregated at every K (Lawson et al. 2017). Similarly, a problem with PCA is that the same PC can explain variation for several continental ancestries simultaneously. This makes it challenging to relate such variation to specific demographic processes and thus complicates interpretation of results for admixed individuals that contain different continental ancestries differentiated along the same PC.

Some of the findings using allele-frequency-based approaches can be redefined using haplotype-based methods (van Dorp et al. 2015). For instance, the marginal East Asian-like ancestry detected with ADMIXTURE at K=4 in the Mexican Native Americans – likely to be ancient – can be confounded with recent East Asian ancestry in admixed Mexicans, while, in contrast, it is not detected when using haplotype-based approaches (see Chapter 5).

The clustering presented here generally matches historical, geographical and linguistic sources as well as previous genetic studies (Botigue et al. 2013; Reich et al. 2012). By exploring different metrics of population differentiation (Tree distance, TVD) and different iterations of the model, we can be more certain about the consistency of the clustering and the relative differences between the clusters. However, although results are highly consistent within clusters, it is necessary to keep in mind that the fineSTRUCTURE tree distances are not directly related to time and genetic distance measurements, and should be treated with caution when interpreting relationships between clusters (Leslie et al. 2015). Big differences on the sample sizes can also have an effect on the order of the clustering (Leslie et al. 2012).

Due to this issues, after some exploratory analysis I decided not to use the approach that cuts down the levels of the tree until just two remain. The most high-profile study to date using this approach (Leslie et al. 2015) aimed to deconstruct the structure of the British at different levels to show levels of relatedness and the interconnectivity between regions of the country. By contrast, in this thesis I aim to reconstruct a high-resolution sub-continental ancestry, and for that purpose I needed to keep the clusters at the maximum possible level of separation. Indeed

additional work using merged clusters showed less precision in simulations (Chapter 4, Section 4.2).

With the steps I performed after the initial fineSTRUCTURE analysis, I have been able to establish a consistent set of surrogates based on their haplotype similarity patterns, which I will use to infer the contribution of specific population groups to the genetic make-up of Latin Americans.

3.10 Summary

Applying fineSTRUCTURE – a clustering model based on haplotype similarity patterns – to a wide set of reference populations, I have been able to establish a consistent set of genetically homogeneous clusters to be used as surrogates for the original ancestors of Latin American populations. Comparisons with widely used allele-frequency-based approaches provide evidences of an increase in the resolution of fine-scale population structure using haplotype profiles, and provide further support for the interpretation of the results in the remaining chapters.

In the next chapter I assess the accuracy and robustness of the sub-continental ancestry estimations using the clustering established here, and demonstrate that these analyses are allowing the detection of sub-continental ancestry at a level never achieved previously.

4 Assessment of NNLS, SOURCEFIND and GLOBETROTTER performance through simulations

4.1 Overview

In the previous chapter I have established a consistent set of reference populations to use as surrogates for the ancestral populations that contributed to the make-up of current-day admixed Latin American populations. In this chapter I perform a series of simulations modelling the admixture in Latin America in order to assess the robustness and accuracy of the methods we use to estimate sub-continental ancestry (NNLS and SOURCEFIND) as well as the estimated dates of admixture (GLOBETROTTER) in this setting. While previous work has demonstrated the increase in resolution of some of these haplotype-based approaches over traditional frequency-allele-based methods (Lawson et al. 2012, Hellenthal et al. 2014, Leslie et al. 2015), these simulations provide the first formal assessments of the accuracy of (*i*) SOURCEFIND relative to NNLS, (*ii*) SOURCEFIND and NNLS in capturing sub-continental admixture in single individuals and (*iii*) GLOBETROTTER for dating admixture in single individuals.

Furthermore, since the precision of sub-continental ancestry estimates is affected by the relatedness of surrogate clusters, and their level of genetic drift, these simulations also allowed the exploration of which sub-continental ancestries cannot be reliably distinguished. Subsets of some of the 56 surrogate clusters were used to generate simulated admixed individuals following the procedures described in Leslie et al. (2015), Hellenthal et al. (2014), Moorjani et al. (2011) and Price et al. (2009).

4.2 Simulations to assess accuracy of sub-continental ancestry estimates

For each of the four sets of simulations described in Sections 4.2.1 to 4.2.4, I generated 100 simulated individuals as mixtures of three surrogate clusters (as defined in Chapter 3) intermixing 15 generations ago. Simulations were performed as described in Price et al. (2009) and assume a model of instantaneous admixture followed by random mating. Briefly, each simulated haploid genome consists of a mosaic of blocks, each block of size M (in Morgans) sampled from an exponential distribution (with a rate parameter of 15 in this case, to simulate an admixture event 15 generations ago). For each block, the SNP data exactly matched that of a randomly sampled haplotype from one of the surrogate clusters, with the probabilities for selecting a haplotype from each of the three surrogate clusters specified by the admixture proportions being simulated as indicated in Sections 4.2.1 to 4.2.4. This random selection process was repeated independently for each block. Two haploid genomes were randomly combined to generate each simulated diploid individual.

From the clusters selected for the simulations, I used less than half the individuals in each cluster (usually ~30%) to simulate admixed individuals, in order to keep the remaining as surrogates for the sub-continental ancestry estimation. Additionally, for every set of simulations an independent CHROMOPAINTER analysis was performed, excluding for the donors the individuals used as templates for the simulations.

All SOURCEFIND analyses were performed with 20 independent runs using 200,000 iterations for each run, as described in Chapter 2. As with the real data analysis (Chapter 5), for each run I extracted results for the sampled iteration with the highest posterior probability, and I then took a weighted average of these maximum a posteriori results across the 20 runs, using this probability as a weight. Non-negative-least-squares (NNLS) was run using GLOBETROTTER as described in Chapter 2. I note that accuracy of both NNLS and SOURCEFIND depends in part on the number of individuals used in each surrogate cluster, so that removing ~30% of the individuals from each simulating group when performing inference may decrease accuracy.

The precision of the sub-continental ancestry assignments varies depending on which clusters were used to simulate, with contributions from drifted clusters, like Native Americans, typically easier to differentiate. Distinguishing among contributions from less genetically differentiated clusters, such as among European and East/South Mediterranean groups, generally is more challenging, so that these simulations also allowed exploration of which groups cannot be reliably distinguished.

Additionally, for NNLS, I also performed analyses using different numbers of surrogate clusters to explore its effect on the estimation (data not shown). I used the classification of 35 groups of clusters defined in Chapter 3 (Section 3.7, Figure 3.2) and a broad classification covering the five main continental regions included in the reference dataset (Native American, African, European, East Asian, East/South Mediterranean). I found that the best estimates in the simulated and real data were obtained with the 56 surrogate clusters and decided to consistently use them in all the inferences and only summarize the results based on groups of clusters when needed for display or comparison purposes.

Next I describe the details and results for each of the four simulation scenarios. Each set was simulated with different admixture percentages and sources, as described at the beginning of every Section (in parenthesis is indicated the total sample size of each cluster included).

4.2.1 European sub-continental ancestries can be estimated accurately

Here I simulated individuals as mixtures of 16 individuals from *CentralSouthSpain* (N=48), 32 from *NorthWestEurope2* (N=101) and 5 from *SouthMexico1* (N=16), each contributing (on average) 40%, 30% and 30% respectively (Figure 4.1).

When using NNLS as described in e.g. Leslie et al. (2015), ancestry from *SouthMexico1* is inferred with high accuracy, showing little marginal uncertainty and little misassignment even to *Nahua1* (Figure 4.2), a striking result considering that these two surrogate clusters are closely related as shown in the fineSTRUCTURE tree (Chapter 3, Figure 3.2). The accuracy obtained with SOURCEFIND is even

higher, having a nearly perfect match to the true simulated proportions and sources (Figure 4.3).

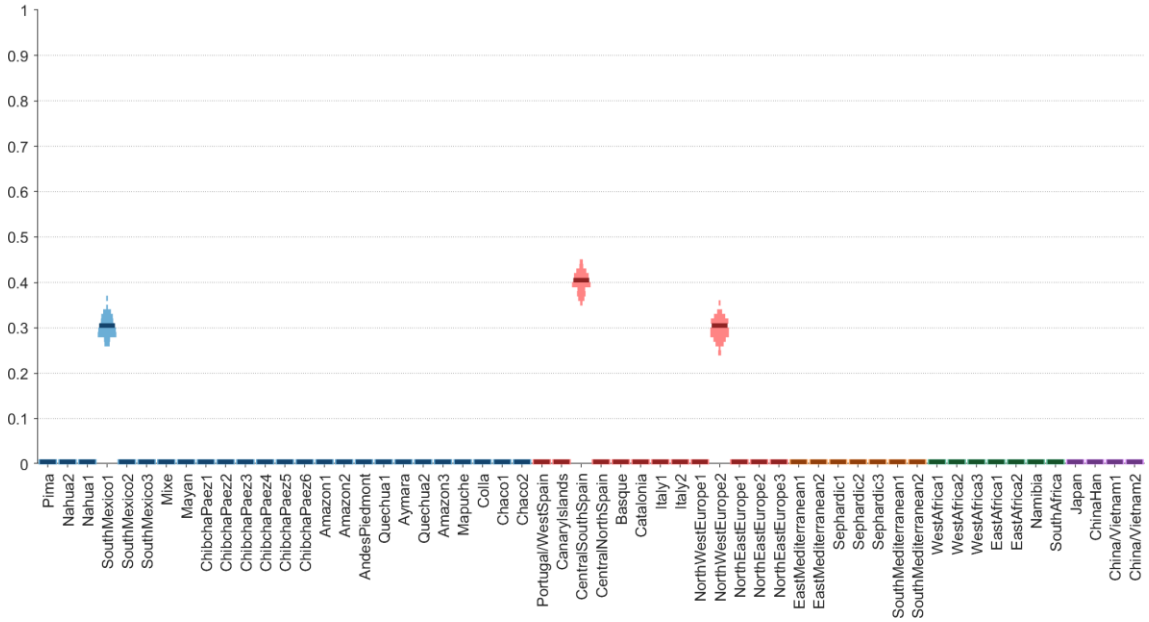

Figure 4.1. Pyramid chart showing the distribution of simulated ancestry proportions from each surrogate cluster across the 100 simulated individuals.
Colours correspond to major geographic regions: NAM (blue), EUR (red), ESM (dark orange), SSA (green), and EAS (purple). Black horizontal lines show the mean proportions of ancestry from each source group. Taken from Chacón-Duque et al. (2018). Generated by JC Chacón-Duque and K Adhikari.

In the case of *CentralSouthSpain*, NNLS shows high levels of misassignment to other Iberian surrogates. The highest misassigned values are to *CentralNorthSpain,* which is the group most genetically similar to *CentralSouthSpain*, with additional misassignments to *Portugal/WestSpain*, *Basque* and others. There are additional inferred contributions from East/South Mediterranean populations, up to a maximum of approximately 5%. In contrast, SOURCEFIND estimations are highly accurate, with only very minor inferred incorrect contributions related to *Italy1*, which may relate to genuine simulated ancestry from *CentralSouthSpain* or *NorthWestEurope2* given their intermediate location between these two simulated source groups. Importantly, there are no mis-inferred contributions from East/South Mediterranean populations when using SOURCEFIND.

The estimation of *NorthWestEurope2* ancestry is typically more accurate, with some noise associated to *NorthWestEurope1* (max ~10%), considerably more in

NNLS relative to SOURCEFIND, causing the overall ancestry of the real source (*NorthWestEurope2*) to decrease slightly.

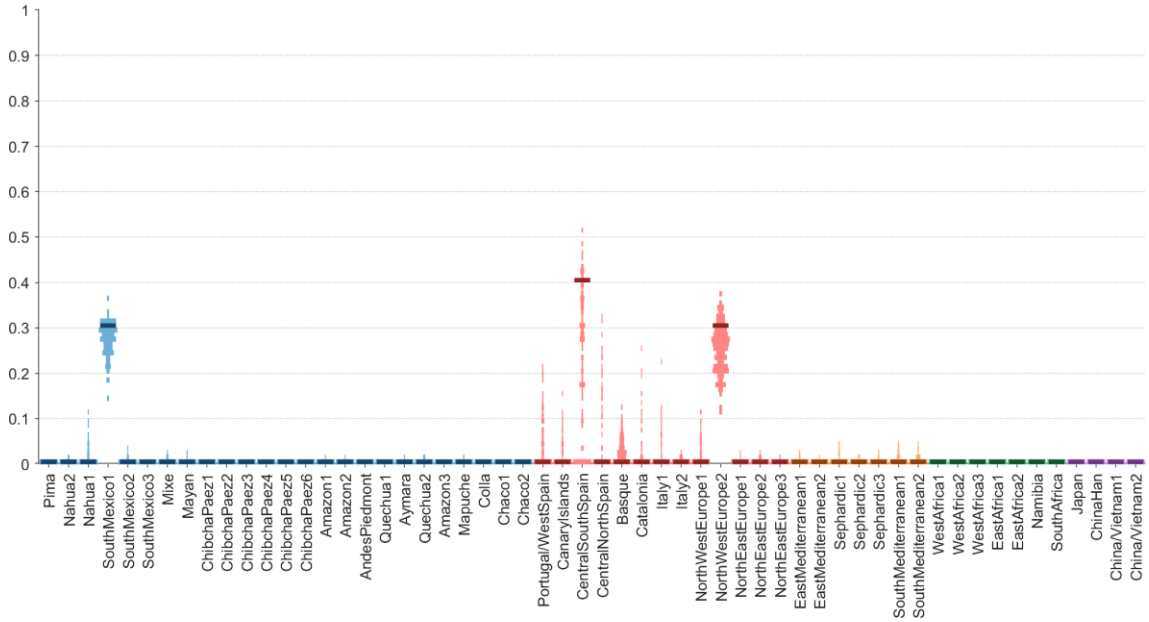

Figure 4.2. Pyramid chart showing the distribution of ancestry proportions assigned to each surrogate cluster across the 100 simulated individuals, as inferred by NNLS.
Other details in Figure 4.1 legend. Taken from Chacón-Duque et al. (2018). Generated by JC Chacón-Duque and K Adhikari.

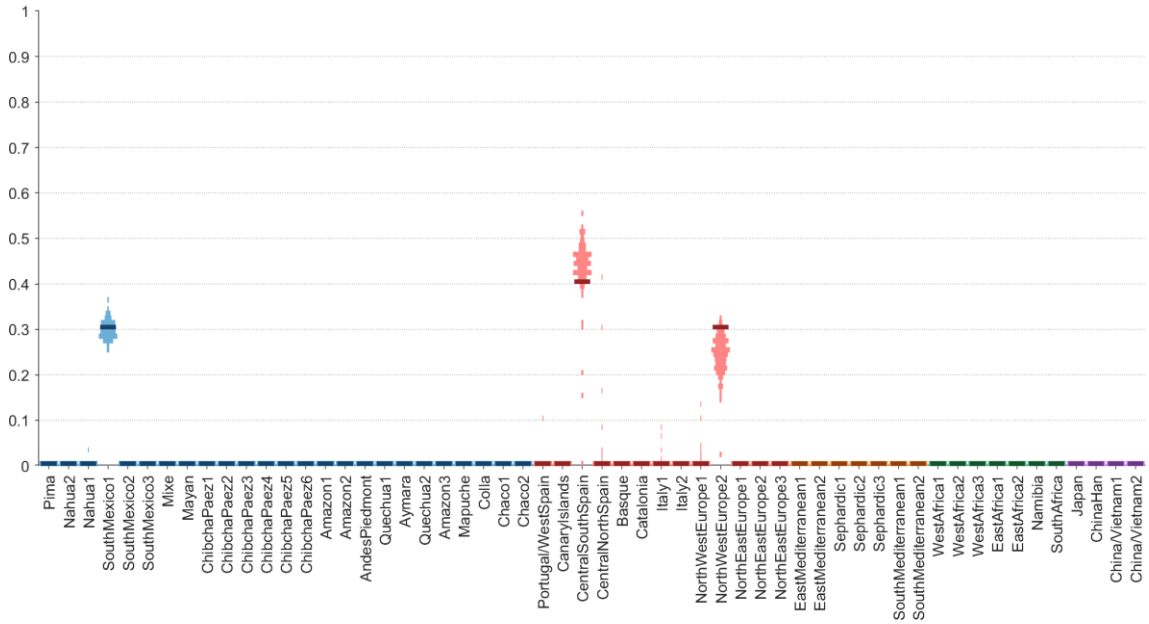

Figure 4.3. Pyramid chart showing the distribution of ancestry proportions assigned to each surrogate cluster across the 100 simulated individuals, as inferred by SOURCEFIND.
Other details in Figure 4.1 legend. Taken from Chacón-Duque et al. (2018). Generated by JC Chacón-Duque and K Adhikari.

Overall, this simulation demonstrates the increased resolution of SOURCEFIND compared to NNLS for resolving ancestral origins among Iberian populations, and in particular greatly decreases the noise of mis-specified contributions related to East/South Mediterranean groups.

4.2.2 Southern European clusters can be distinguished

For the simulations I used 16 individuals from *Portugal/WestSpain* (N=53), 7 from *Italy1* (N=19) and 6 from *Aymara* (N=16), and set up the percentages to be 40%, 30% and 30% respectively (Figure 4.4).

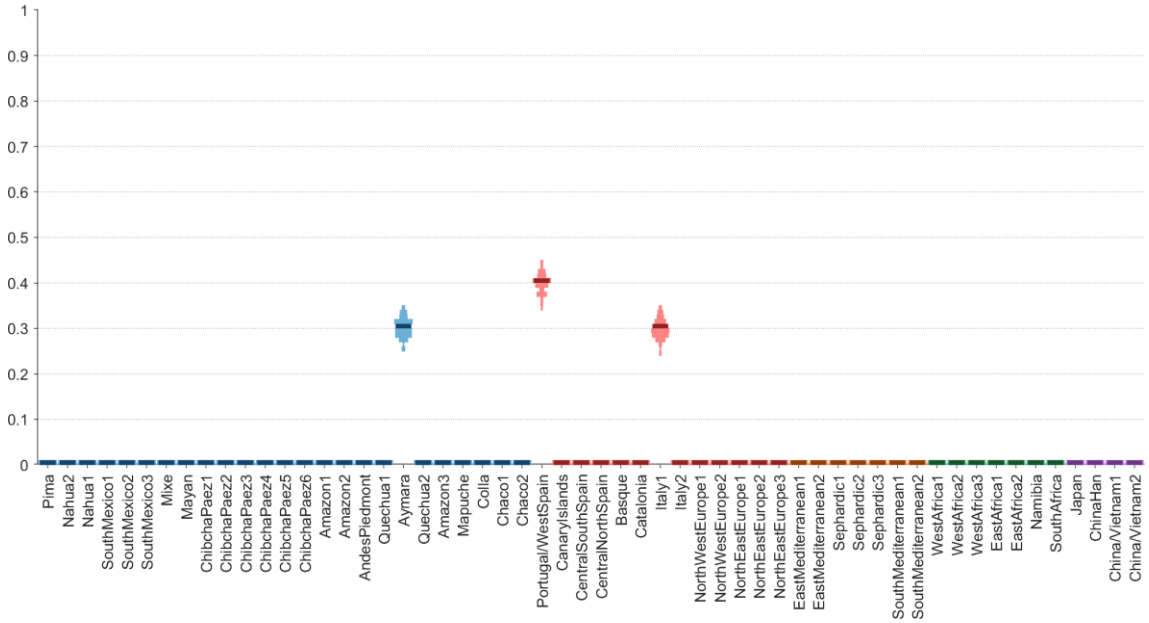

Figure 4.4. Pyramid chart showing the distribution of simulated ancestry proportions from each surrogate cluster across the 100 simulated individuals.
Other details in Figure 4.1 legend. Taken from Chacón-Duque et al. (2018). Generated by JC Chacón-Duque and K Adhikari.

NNLS shows clear difficulty in discriminating *Aymara* from *Quechua2*, consistent with their close genetic relatedness and the small sample size of the former, which makes the inference more challenging (Figure 4.5). When *Quechua2* is simulated (Section 4.2.3), it seems that the resolution to distinguish between the two populations is higher. In the case of SOURCEFIND there seems to be no difficulty resolving these ancestries in either simulation scenario (Figure 4.6).

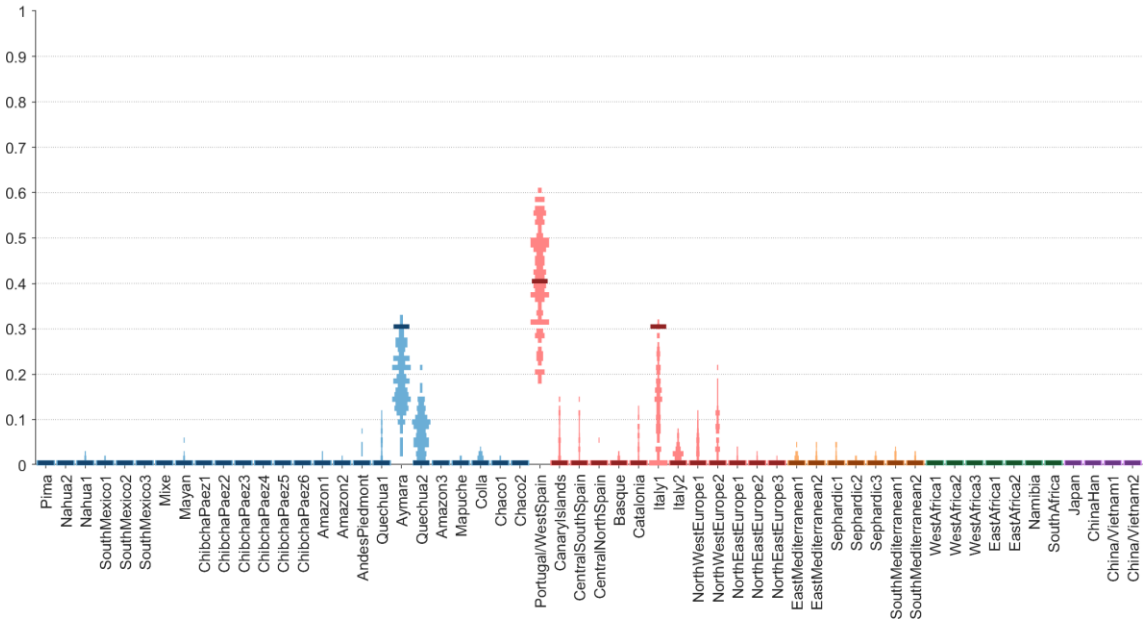

Figure 4.5. Pyramid chart showing the distribution of ancestry proportions assigned to each surrogate cluster across the 100 simulated individuals, as inferred by NNLS.
Other details in Figure 4.1 legend. Taken from Chacón-Duque et al. (2018). Generated by JC Chacón-Duque and K Adhikari.

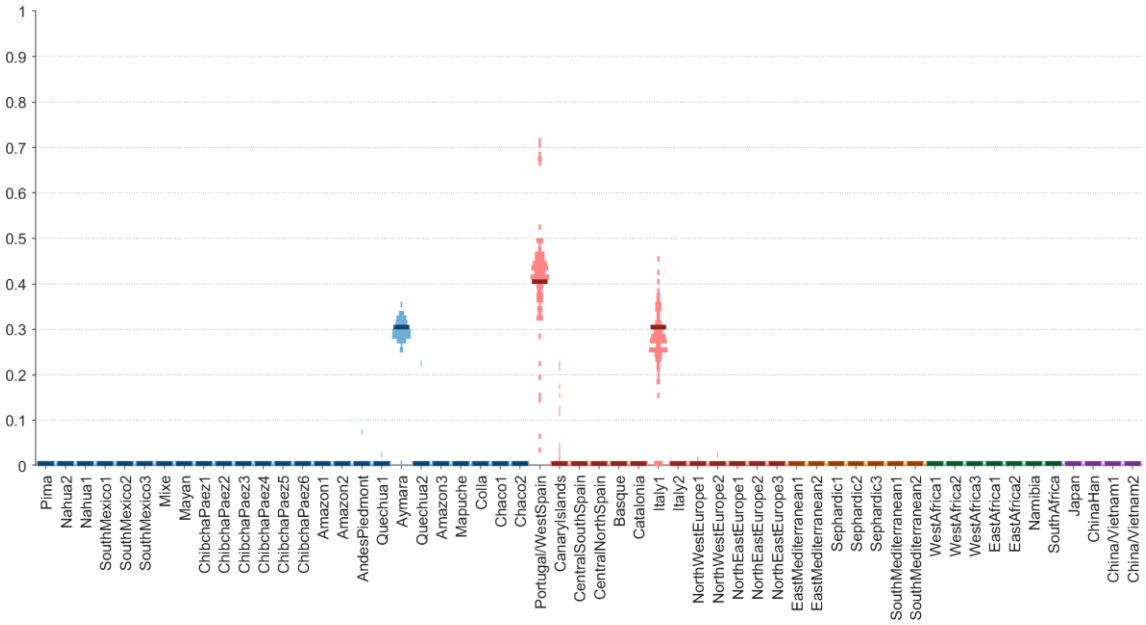

Figure 4.6. Pyramid chart showing the distribution of ancestry proportions assigned to each surrogate cluster across the 100 simulated individuals, as inferred by SOURCEFIND.
Other details in Figure 4.1 legend. Taken from Chacón-Duque et al. (2018). Generated by JC Chacón-Duque and K Adhikari.

Portugal/WestSpain correctly dominates the inferred ancestry from Europe, though shows a tendency to be overestimated while ancestry from *Italy1* is underestimated. Although again SOURCEFIND inferences are better than those of

NNLS, this same tendency is still observed. Again there are marginal East/South Mediterranean contributions incorrectly inferred by NNLS, as well as several additional contributions throughout Europe, though these mis-specifications are avoided by SOURCEFIND, suggesting that these signals are associated to noise related to the NNLS method.

4.2.3 Iberian ancestries can be estimated accurately

For this simulation I used 15 individuals from *Quechua2* (N=56), 16 from *CentralSouthSpain* (N=48) and 22 from *WestAfrica3* (N=99), and set up the percentages to be 40%, 40% and 20% respectively (Figure 4.7).

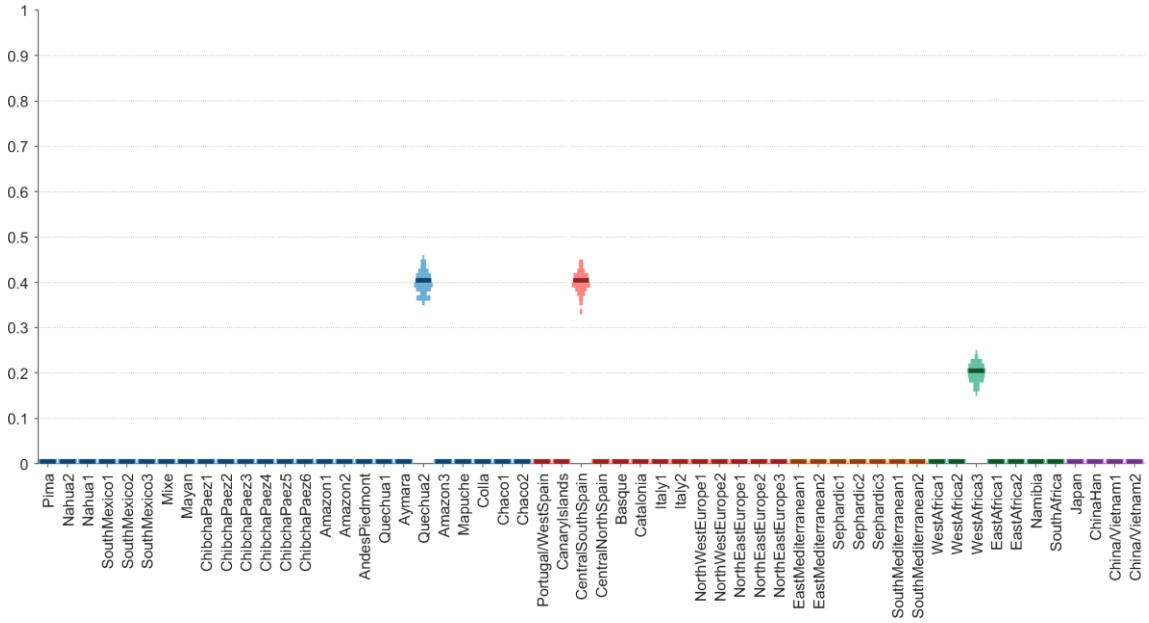

Figure 4.7. Pyramid chart showing the distribution of simulated ancestry proportions from each surrogate cluster across the 100 simulated individuals.
Other details in Figure 4.1 legend. Taken from Chacón-Duque et al. (2018). Generated by JC Chacón-Duque and K Adhikari.

Estimation for *WestAfrica3* is highly accurate in both NNLS and SOURCEFIND, likely due to the large sample size of this reference group and the fact that African haplotypes are easier to classify given their relative amount of genetic differentiation from other reference groups. *Quechua2* is also well differentiated compared to the previous simulation, suggesting that small sample sizes, or perhaps differences in true sources of ancestry, could be a limiting factor for sub-continental ancestry estimation. As before, SOURCEFIND results demonstrate more accuracy than NNLS (Figures 4.8 and 4.9).

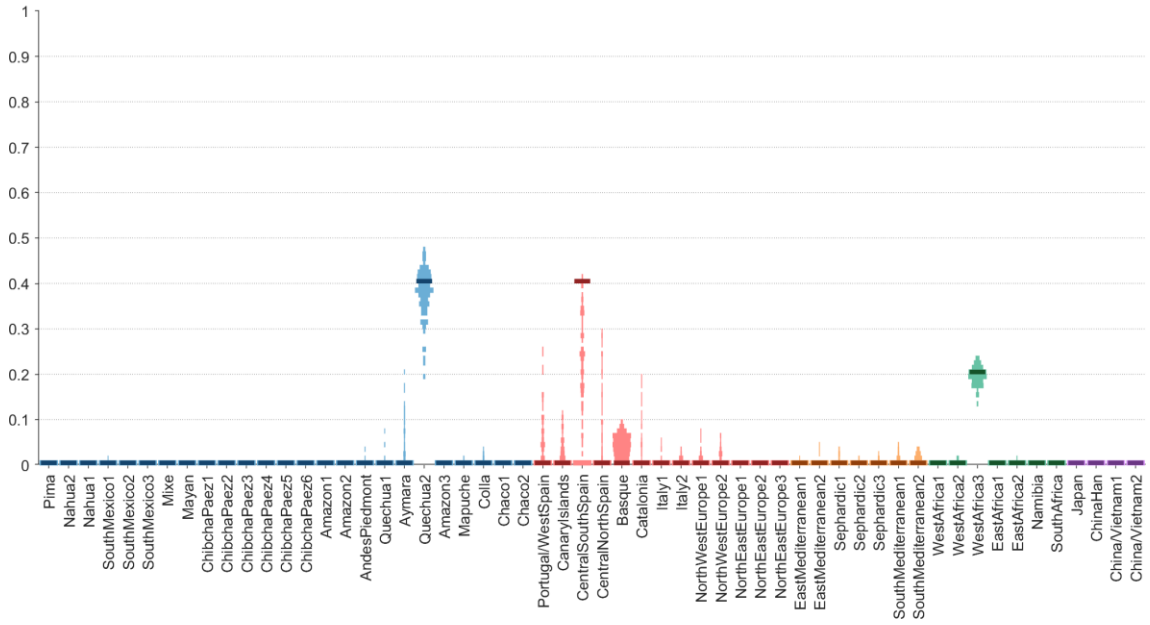

Figure 4.8. Pyramid chart showing the distribution of ancestry proportions assigned to each surrogate cluster across the 100 simulated individuals, as inferred by NNLS. Other details in Figure 4.1 legend. Taken from Chacón-Duque et al. (2018). Generated by JC Chacón-Duque and K Adhikari.

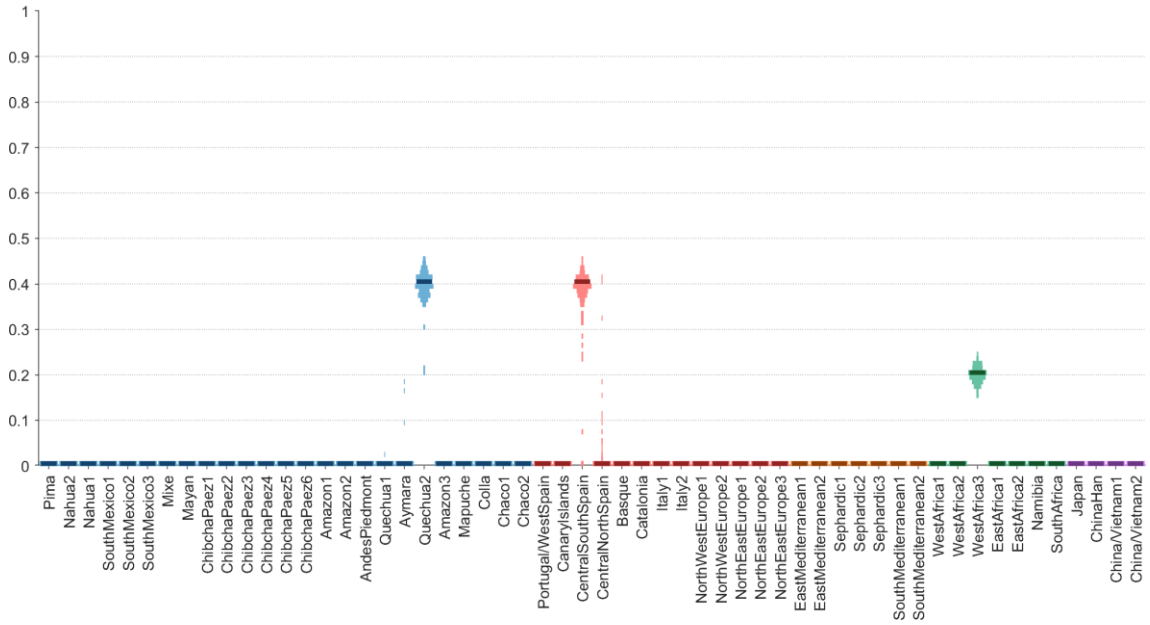

Figure 4.9. Pyramid chart showing the distribution of ancestry proportions assigned to each surrogate cluster across the 100 simulated individuals, as inferred by SOURCEFIND. Other details in Figure 4.1 legend. Taken from Chacón-Duque et al. (2018). Generated by JC Chacón-Duque and K Adhikari.

I also note that NNLS infers a notable spurious contribution from *Basque*, which suggests that inferred Basque-like contributions in the Americas using this approach should be treated with caution (Montinaro et al. 2015).

4.2.4 Closely related Native American ancestries can be quantified and separated accurately

For this simulation I used 6 individuals from *SouthMexico1* (N=16), 3 from *Mayan* (N=7) and 16 from *CentralSouthSpain* (N=48). The percentages of ancestry were set up to be 40%, 40% and 20% respectively (Figure 4.10).

These results suggest that, for NNLS, the presence of different mis-specified signals of ancestry across the Iberian groups is somehow proportional to the amount of true ancestry from these sources (Figure 4.11). This information could allow the establishment of noise thresholds in NNLS inference. For example, if the highest values of Basque ancestry in an individual with 20% *CentralSouthSpain* is around 2%, and around 4% for an individual with 40% *CentralSouthSpain*, we could in theory predict that an individual in the real dataset with 80% *CentralSouthSpain*-like ancestry may have ~8% *Basque* ancestry attributable to noise. SOURCEFIND does not show this problem, instead showing only a slight mis-assignment of this Iberian component to the closest group, *CentralNorthSpain* (Figure 4.12).

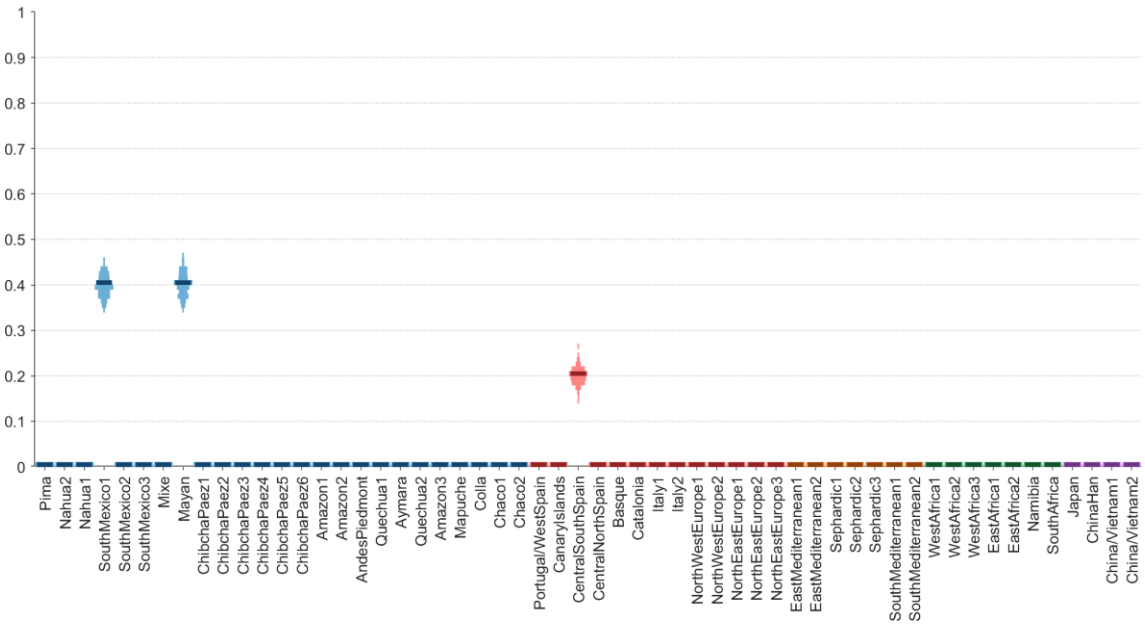

Figure 4.10. Pyramid chart showing the distribution of simulated ancestry proportions from each surrogate cluster across the 100 simulated individuals.
Other details in Figure 4.1 legend. Taken from Chacón-Duque et al. (2018). Generated by JC Chacón-Duque and K Adhikari.

The Native American components, even though from two genetically similar sources, are correctly assigned by both approaches, though with SOURCEFIND

again showing increased precision. As stated above, this could be related to the fact that the relatively higher drift of Native groups allows better differentiation of the haplotype-based copying profiles.

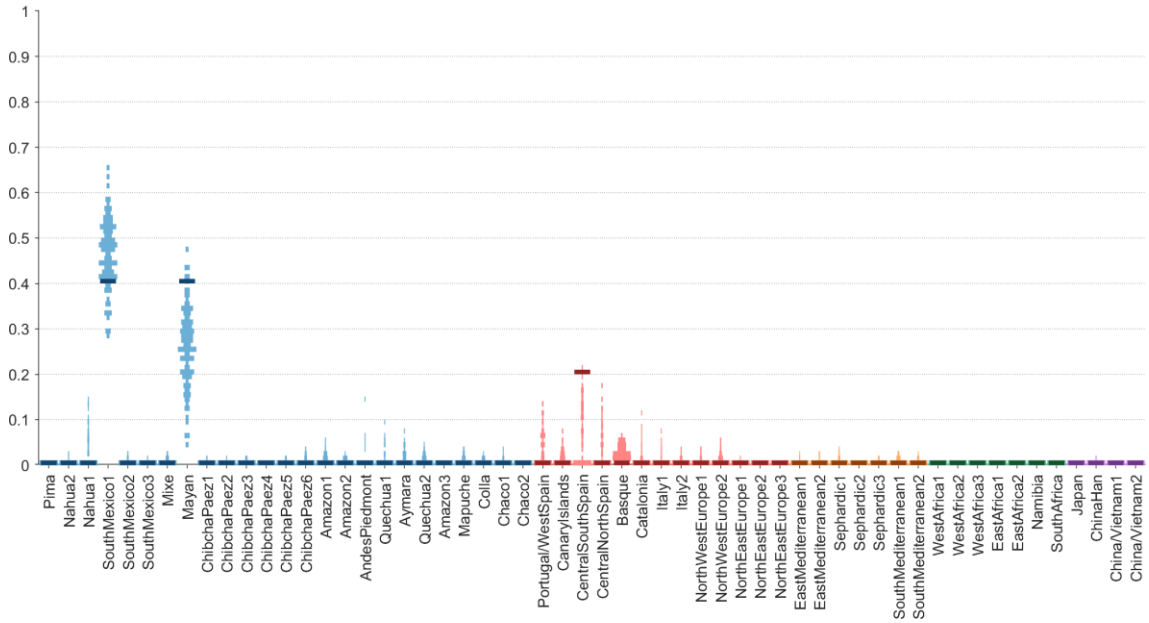

Figure 4.11. Pyramid chart showing the distribution of ancestry proportions assigned to each surrogate cluster across the 100 simulated individuals, as inferred by NNLS. Other details in Fig. 4.1 legend. Taken from Chacón-Duque et al. (2018). Generated by JC Chacón-Duque and K Adhikari.

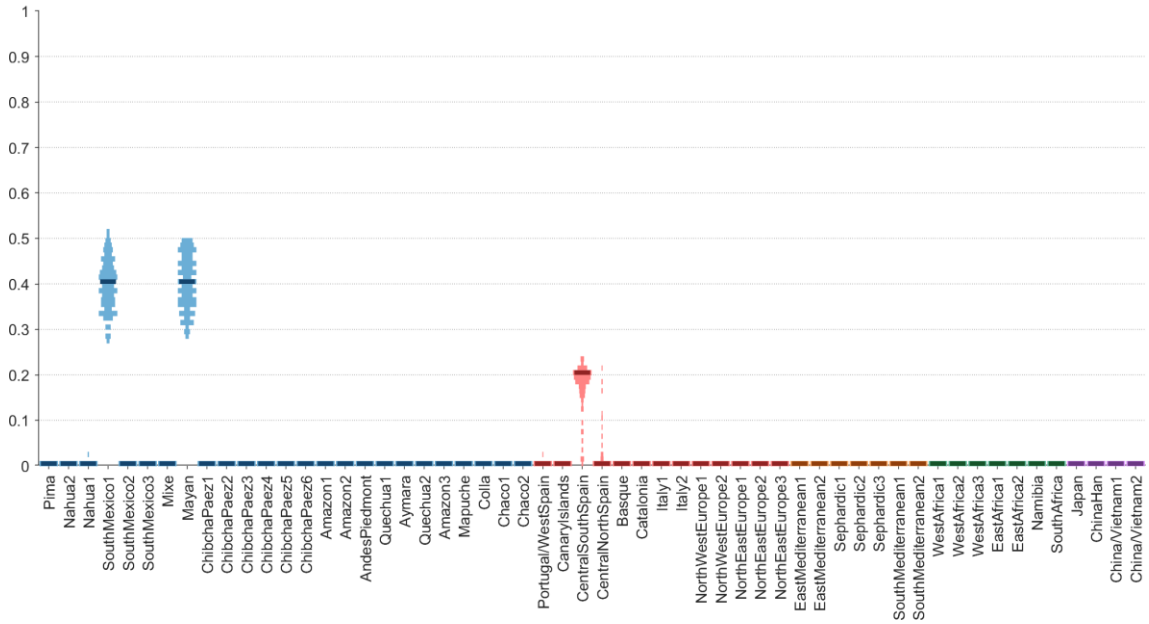

Figure 4.12. Pyramid chart showing the distribution of ancestry proportions assigned to each surrogate cluster across the 100 simulated individuals, as inferred by SOURCEFIND.
Other details in Figure 4.1 legend. Taken from Chacón-Duque et al. (2018). Generated by JC Chacón-Duque and K Adhikari.

4.3 Simulations to assess the accuracy of individual estimations of dates since admixture events

The two sets of simulations below aim to assess the accuracy per individual estimation of time since admixture and the effect of time since admixture on ancestry estimation.

4.3.1 Simulations with a single admixture event

We simulated an additional 1,430 individuals with different proportions of admixture from two sources (*CentralSouthSpain* and *Quechua2*) and different times since admixture. Using the procedure described in Section 4.2, each individual was simulated as descending from an instantaneous admixture event that occurred *g* generations ago, with a proportion *p* of ancestry inherited from *CentralSouthSpain*, and *1-p* ancestry inherited from *Quechua2*. We simulated with 9 different values of *p* = 5, 10, 20, 30, 40, 50, 60, 70, 80, 90 and 95% and 13 different values of *g* = 5, 6,..., 17, with 10 simulated individuals for each combination of *p* and *g*, giving 11x13x10=1,430 simulated individuals in total.

We used 16 *CentralSouthSpain* and 20 *Quechua2* individuals to generate the admixed individuals (same number of individuals from Section 4.2), using our remaining 32 *CentralSouthSpain* and 36 *Quechua2* individuals to define their respective surrogate groups under inference using SOURCEFIND and GLOBETROTTER. Both programs were run separately on each simulated individual, with the slight exception that GLOBETROTTER was allowed to use all surrogates to describe the admixture (i.e. rather than only including surrogates inferred by SOURCEFIND to contribute >1%, which is the procedure I use in Chapter 5 for the CANDELA data).

In contrast to the simulations above, for these simulations I used the alternative, more computationally efficient version of SOURCEFIND (Chapter 2, Section 2.4.2). Here I used *S'*=6 and performed 100,000 total MCMC iterations, sampling posterior values of $\beta_1^r,\ldots,\beta_S^r$ every 5,000 iterations after discarding the initial 50,000 iterations as “burn-in". The results are highly consistent with those produced by the other version of SOURCEFIND (data not shown).

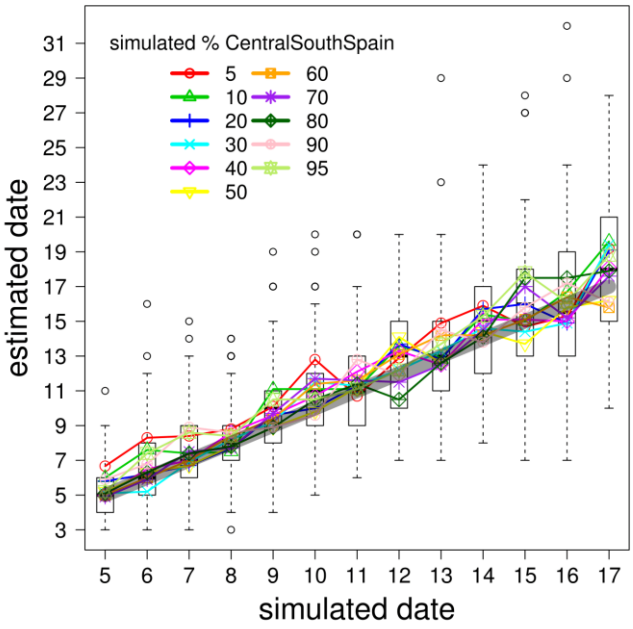

Figure 4.13. GLOBETROTTER's inferred dates (y-axis) across individuals, for simulations mixing *CentralSouthSpain* and *Quechua2* at the given proportions (legend) and times (x-axis).
Coloured dots and lines show mean results across all 10 individuals simulated with the given proportions and dates, with the grey shaded bar highlighting the true simulated dates. Taken from Chacón-Duque et al. (2018). Generated by JC Chacón-Duque and G Hellenthal.

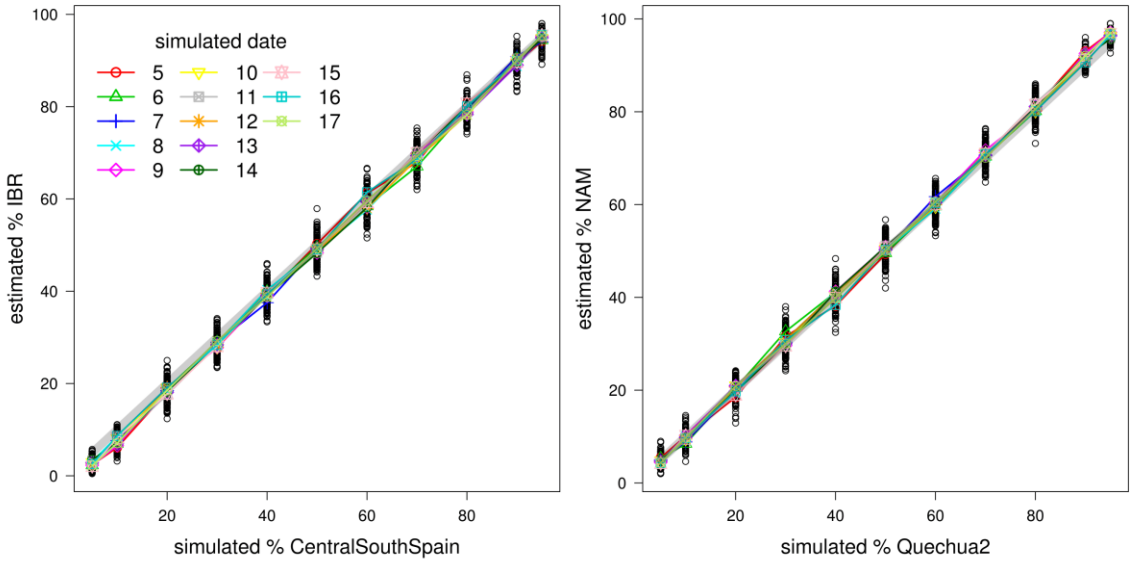

Figure 4.14. SOURCEFIND's inferred proportion of ancestry related to Iberian (IBR) and Native American (NAM) sources (y-axis) across individuals (circles), for simulations mixing *CentralSouthSpain* and *Quechua2* at the given proportions (x-axis) and times (legend).
Coloured lines show mean results across all 10 individuals simulated with the given proportions and dates, with the grey shaded bar highlighting the true simulated proportions. Taken from Chacón-Duque et al. (2018). Generated by JC Chacón-Duque and G Hellenthal.

Figure 4.13 shows that on average, GLOBETROTTER's estimated dates across individuals accurately reflect the simulated dates (grey bar), and that this accuracy is not affected by the true proportion of admixture from each group. Similarly, SOURCEFIND's accuracy in inferring the proportion of DNA contributed from Iberian (IBR) and Native American (NAM) source groups did not depend on the true date of admixture (Figure 4.14). In the case of NAM, the inferred proportion of ancestry almost always matches *Quechua2* (data not shown).

Finally, we tried to replicate the same pattern described in Chapter 5 (Figure 5.20 and Table 5.3), where the amount of Native American ancestry tends to increase as the admixture events become more recent, noting that no such trend should exist in these simulations. To do this, we extracted our 1,297 simulated individuals that were inferred to have a single date of admixture between Native and European source groups. We then binned these simulated individuals based on their inferred date, and calculated the average inferred proportion of DNA matching to European (EUR) versus Native American (NAM) groups across individuals in each date bin. In the figure below, it is clear that the pattern observed in the real data does not exist in this simulated data (Figure 4.15), suggesting that this detected pattern is not an artefact of the GLOBETROTTER estimation. This issue is further explained in Chapter 5.

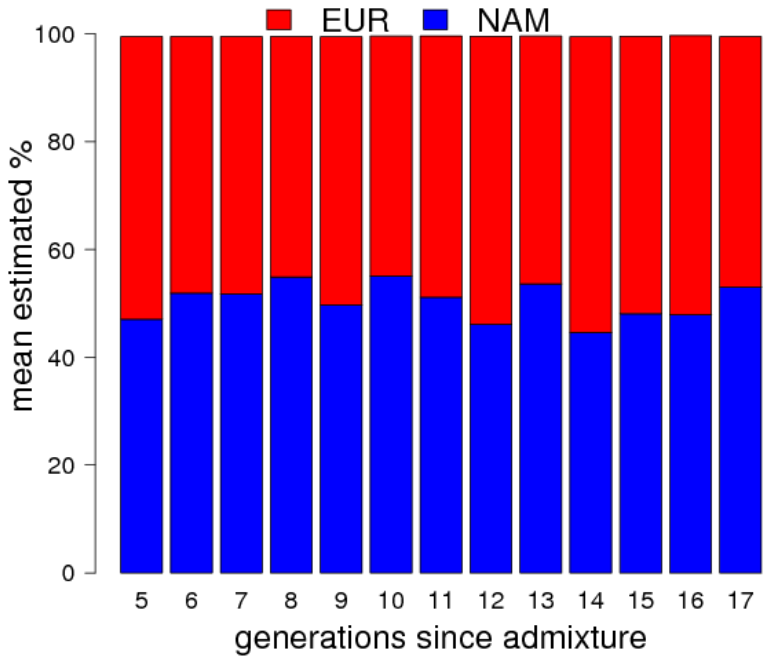

Figure 4.15. Mean ancestry percentages in the simulated individuals estimated by SOURCEFIND grouped by the number of generations since admixture.
Taken from Chacón-Duque et al. (2018). Generated by JC Chacón and G Hellenthal.

4.3.2 Simulations with two sequential admixture events

To further evaluate the trend of increasing Native ancestry at more recent dates of admixture seen in the CANDELA data, 1,050 additional individuals with two sequential admixture events were simulated (this analysis was performed jointly with G. Hellenthal). As before, we simulated different proportions of admixture from two sources (*CentralSouthSpain* and *Quechua2*), and varied the times for the two admixture events.

Using the exponential sampling procedure described in Section 4.2, we first simulated individuals stemming from an instantaneous admixture event occurring 2 generations previously, with 55% *CentralSouthSpain* ancestry and 45% *Quechua2* ancestry. We then simulated a second instantaneous admixture event with *p* ancestry from the population generated in the first admixture event, and 1-*p* ancestry from *Quechua2* occurring *g* generations ago. We simulated *p* = 0.86-0.98 (at 0.02 intervals) and *g* = 5-14 generations, with 15 simulated individuals for each combination of *p* and *g* (1,050 simulated individuals in total). Note that, under this simulation procedure, the first admixture event occurred *g*+2 generations ago, the more recent event occurred *g* generations ago, and the final expected proportion of ancestry from *CentralSouthSpain* is 0.55**p*.

SOURCEFIND and GLOBETROTTER were run separately on each simulated individual as before. As with the previous section, for these simulations we used the more computationally efficient version of SOURCEFIND (Chapter 2, Section 2.4.2) to infer proportions.

In 923 (~88%) of the 1,050 individuals, GLOBETROTTER concluded only a single date of admixture, which is not surprising given the inherent difficulty in distinguishing between two pulses of admixture separated by only 2 generations that involve the same source groups. Figure 4.16 shows results when assuming a single date of admixture, which infers dates that typically are 2 generations above *g* (simulated date given with the grey bar). Therefore, GLOBETROTTER most often concludes a single date of admixture, with the inferred date reflecting mainly the older event.

Figure 4.17 illustrates that SOURCEFIND accurately estimates the admixture proportions in the simulated individuals (grey bar gives simulated proportion).

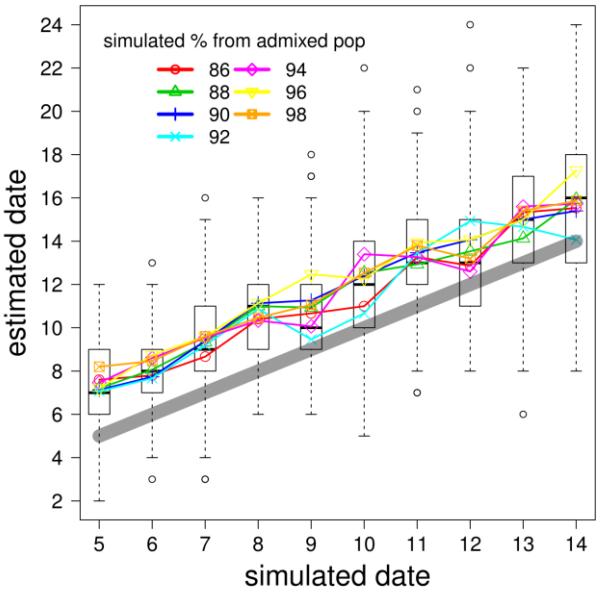

Figure 4.16. GLOBETROTTER's inferred dates (y-axis) across individuals, for simulations with two sequential admixture events, at the given proportions (legend) and times (x-axis).
Coloured dots and lines show mean results across all 15 individuals simulated with the given proportions and dates, with the grey shaded bar highlighting the true simulated dates. Taken from Chacón-Duque et al. (2018). Generated by G Hellenthal.

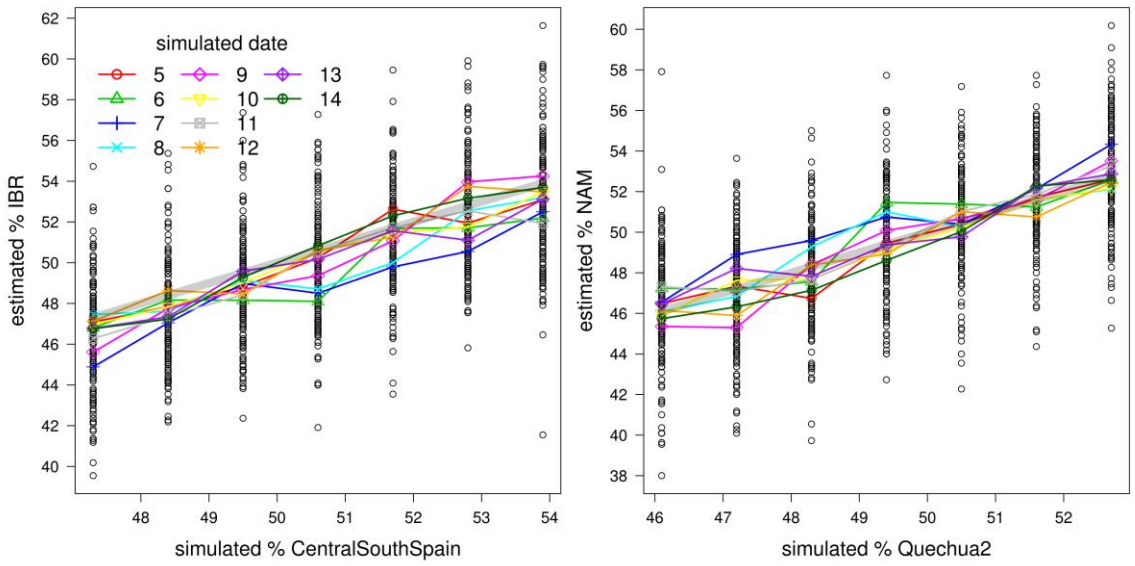

Figure 4.17. SOURCEFIND's inferred proportion of ancestry related to Iberian (IBR) and Native American (NAM) sources (y-axis) across individuals (circles), for simulations with two sequential admixture events, at the given proportions (x-axis) and times (legend).
Coloured lines show mean results across all 15 individuals simulated with the given proportions and dates, with the grey shaded bar highlighting the true simulated proportions. Taken from Chacón-Duque et al. (2018). Generated by G Hellenthal.

In addition, as above, we extracted the 923 simulated individuals that GLOBETROTTER inferred to have a single admixture event between source

groups that best-matched Native and European surrogate groups. We binned these individuals based on their inferred admixture date, and calculated the average ancestry inferred proportions in each bin. While not as striking as that observed in our real data (Chapter 5, Figure 5.20), this set of simulations shows an analogous trend for decreasing Native American ancestry at increasing *g* that is significant ($p<0.001$) under the same simple linear regression model used for analysing this trend in the real data (Chapter 5, Table 5.3). While we did not simulate increasing Native ancestry over time, individuals here are simulated with different proportions of admixture from the earlier admixture event occurring *g*+2 generations ago. Individuals with more simulated ancestry from this earlier admixed group have (*i*) more European ancestry and (*ii*) inferred dates that may be biased to be slightly older by retaining more signal from this older admixture event. Indeed, a simple linear regression of the bias in date estimates for these 923 individuals on their expected proportion of Spanish ancestry shows a significantly positive association ($p<0.007$). In contrast, for the 1,297 simulated individuals described in the previous section with only a single simulated admixture date, there is no such significant trend ($p=0.33$). Overall these simulation results suggest that mixture between unadmixed and admixed Natives over time, such as that we simulated in this section, could lead to the trend we observe in Chapter 5 (Figure 5.20).

4.4 Discussion and limitations

The analyses performed in this chapter confirm the accuracy of haplotype-based sub-continental ancestry, and admixture time and sources estimations, in an appropriate setting for Latin America, corroborating that the approaches I have utilized possess enough resolution to distinguish different sub-continental ancestries and to characterize admixture events in single individuals. This can have useful applications for the analysis of recently admixed populations, as this is the first time that a fully haplotype-based analysis has been performed on each individual separately.

However, it is important to consider all the limitations of these simulated scenarios. Firstly, we are using the surrogate clusters and not the real ancestral populations as templates for the simulations. This makes the ancestry inference more

straightforward in the simulations compared to the real data, as we do not know how close the surrogate clusters are to the original source populations. Even if we have sampled the most genetically similar groups, it is difficult to establish how different they are compared to the groups that actually contributed the DNA in the past (Chapter 3, Section 3.9).

Secondly, these simulations are a simplified representation of the demographic and evolutionary processes behind the current genetic make-up of Latin American populations, also making easier the ancestry inference in the simulations in comparison to the real data. Even though simulations using coalescent approaches could help to reconstruct these complex demographic scenarios in greater detail (Hoban et al. 2012; Hudson 2002), it is not an simple problem to overcome considering that the populations within Latin America can have very different demographic histories. These admixed populations dispersed through the vast region since the beginning of colonial times have experienced varied demographic events at different magnitudes, mainly including deep bottlenecks and differential gene flow from the ancestral populations (Koehl and Long 2018).

4.5 Summary

This chapter assessed the performance of the different methods used in this thesis, in particular suggesting our methods accuracy to (*i*) identify sources and proportions of sub-continental ancestry and (*ii*) infer dates of admixture when analysing single individuals simulated to mimic genetic features of Latin Americans. Overall, these results provide strong support for my conclusions about the genetic history of Latin American populations in the remaining chapters of this thesis.

5 Genetic history of Latin America: increasing resolution with haplotype-based approaches

5.1 Overview

In the previous chapter I have tested through simulations the accuracy of the methods applied in this thesis to identify sources and proportions of sub-continental ancestry and to infer dates of admixture when analysing single individuals using the reference panel established in Chapter 3. In this chapter I explore patterns of sub-continental genetic ancestry in more than 6,500 Latin American individuals across five countries (Mexico, Colombia, Peru, Chile and Brazil). I interpret results according to the history of the region, and estimate the timings and sources involved in the main genetic admixture events that have taken place in the region.

The increase in resolution due to these methods and our large sample size provide a unique opportunity to further reconstruct details of the demographic history of the populations sampled. New findings include fine-grained contributions related to individuals sampled from specific geographic areas within the Iberian Peninsula and local Native American groups, a widespread signal of East / South Mediterranean-like ancestry that is likely to be related to the migration of “Converso” Jews into the American continent during the colonization, and additional small but significant signals from North-western European and East Asian populations. The times and sources of admixture also match historical records of outside migrations related to these regions, from the beginning of the colonial period to the more frequent migrations in the last century.

Overall, these results provide the most comprehensive description to date of the genetic ancestry of Latin America.

5.2 Methods

The CANDELA dataset (Chapter 1, Section 1.6) has been used for all the analyses in this chapter. After quality controls and exploratory analyses, 6,589 individuals and 546,780 autosomal SNPs were retained as described in Chapter 3 (Section 3.3).

Chapter 2 gives theoretical background and technical details on the methods applied here. Comparisons with the analyses performed in reference population individuals (Chapter 3) and simulated individuals (Chapter 4) are described where appropriate.

5.2.1 Estimation of ancestry using allele-frequency-based approaches

As described in Chapter 3 (Section 3.8), after pruning the chip dataset for linkage disequilibrium (LD), 150,858 SNPs were retained for an unsupervised analysis on the same individuals used as surrogates in the haplotype-based analyses (which also covers most of the populations used as references). Additionally, a supervised analysis was implemented, using the same reference population individuals used in the haplotype-based analysis, grouped into the main continental groups in the following scenarios:

(i) Five groups (i.e. using K=5 clusters) – Native American, East Asian, Sub-Saharan African, European and East/South Mediterranean

(ii) Four groups – Native American, East Asian, Sub-Sharan African and European + East/South Mediterranean

(iii) Four groups – Native American, East Asian, Sub-Saharan African and European.

PCA was also performed along with the reference samples as described in Chapter 3.

5.2.2 Inference of haplotype similarity profiles

In order to estimate the haplotype similarity profiles used for sub-continental ancestry estimation, I set up CHROMOPAINTER to provide estimates of the proportion of DNA in every CANDELA individual (recipients) that is most closely related to each reference population individual (donors), allowing us to reconstruct haplotype similarity profiles for all individuals in terms of the surrogate reference clusters' samples. I used exactly the same parameters and donors described in Chapter 3.

5.2.3 Estimation of sub-continental ancestry

The 56 surrogate clusters defined by fineSTRUCTURE from the reference dataset were used as representatives for the ancestral populations contributing to ancestry in Latin America (Chapter 3). I ran SOURCEFIND for 200,000 iterations, sampling every 1,000th iteration. Also, for each recipient individual, I combined results across 50 independent runs, extracting the estimates with the highest posterior probability in each run and then taking a weighted (by posterior probability) average of these 50 estimates. This weighted average (where the more likely values are given higher weight) is equivalent to the posterior mean, which is an estimator of the true value of the mean parameter under Bayesian theory. Informally, this procedure accounts for the uncertainty of the individual ancestry estimations.

To compare continental ancestry assignments with those obtained using ADMIXTURE, SOURCEFIND results were collapsed into continental groups (Section 5.2.1) by adding the inferred ancestry values of the sub-components included within each continental group.

5.2.4 Estimation of number of generations since admixture

The times and sources of major admixture events were estimated with GLOBETROTTER (Chapters 2 and 4). Each CANDELA individual was tested separately for admixture, restricting to the 6,352 individuals inferred by SOURCEFIND to match DNA to more than one surrogate cluster in order to include only admixed individuals.

For each individual, I ran GLOBETROTTER using default settings (see Chapter 2, Section 2.5 for details) and allowing only the subset of ≤ 56 reference groups that contributed >1% to that individual, as inferred by SOURCEFIND, to act as ancestry surrogates when identifying and describing the admixture event. For each CANDELA individual, GLOBETROTTER categorized admixture inference into one of three types: (*i*) one date of admixture involving two sources, (*ii*) one date of involving more than two sources, suggestive of admixture among multiple genetically different groups within a short time span, and (*iii*) multiple dates of admixture between two or more sources (not necessarily the same two), suggesting a more complicated history but which GLOBETROTTER attempts to describe as two major pulses of admixture (Chapter 2, Section 2.5).

For simplicity, the admixture history of the individuals included in type *iii* was described as two distinct events, with each event characterized as having two inferred admixing groups and a single inferred date of mixing. I represent the two admixing sources using GLOBETROTTER's "best-guess" results, which describes each admixing source by the single (included) surrogate group (out of the subset of 56 included in that individual's GLOBETROTTER analysis) that is inferred to be most genetically similar to that (unknown) admixing source group.

To convert time to years, I used the formula proposed in Hellenthal et al. (2014):

$$y = 1990 - 28 \times (g + 1)$$

Where *y* is the year of admixture, *1990* the average birth year across CANDELA individuals, *28 years* the average human generation length according to a cross-cultural estimation using demographic data (Fenner 2005), and *g* is GLOBETROTTER's inferred date (in generations). One generation is added (*g+1*) to account for the fact that recombination inference start from the genetic information on the grandparents.

5.2.5 Testing for patterns in the distributions of inferred admixture dates related to different source groups

In Figure 5.18, I plot histograms for the dates of inferred events involving each of the major geographic labels "Iberia", "NorthWestEurope & Italy", "East Mediterranean & Sephardic", "Sub-Saharan African (SSA)" and "East Asia". These plots

contain the inferred dates for all admixture events that involved a reference group categorized under that major geographic label, with:

"Iberia": *CanaryIslands, Portugal/WestSpain, CentralSouthSpain, Central-NorthSpain, Basque* and *Catalonia*.

"NorthWestEurope & Italy": *Italy1* and *NorthWestEurope1*.

"East Mediterranean & Sephardic": *Sephardic1, EastMediterranean1* and *EastMediterranean2*.

"Sub Saharan Africa": *WestAfrica1, WestAfrica2, WestAfrica3, EastAfrica1, EastAfrica2, Namibia* and *SouthAfrica*.

"East Asia": *Japan, ChinaHan, China/Vietnam1* and *China/Vietnam2*.

The following analysis was conducted jointly with G. Hellenthal. We used "wilcox.test" in R (R-Core-Team 2013) to perform a Wilcoxon rank-sum test (also known as a Mann-Whitney U test) to test the alternative hypothesis that the distribution of admixture dates for each geographic label X= ("East Asia", "NorthWestEurope & Italy", "East Mediterranean & Sephardic", "SSA") is skewed towards more recent dates relative to the "Iberia" geographic label, versus the null hypothesis that distributions are the same. Though they may represent genuine admixture events, for these tests and the histograms I removed events with an inferred date of one generation. This was done both to avoid such dates dominating inference due to their high frequency (8% of all events in Iberia have inferred dates of one generation, with East Asia = 21%, NorthWestEurope & Italy = 6%, East Mediterranean & Sephardic = 10%, SSA = 13%) and because such events have been interpreted as evidence of "no admixture" in past applications of GLOBETROTTER (e.g. Hellenthal et al. 2014).

For the Wilcoxon rank-sum test, we further excluded dates ≥ 30 generations to avoid admixture events that occurred prior to colonial-era migrations. In addition, this analysis assumes each inferred event is an independent observation, even though some individuals have two inferred events. However, we note that conclusions and trends do not change if we restrict to one inferred event per individual (results omitted), e.g. by excluding individuals who infer multiple dates of admixture (e.g. case (iii) described in Section 5.2.4) and only including the more strongly

signalled event in individuals who infer more than two sources of admixture at the same time (e.g. case (ii) described in Section 5.2.4).

To assess the significance of the observed trend of decreasing Native ancestry versus time since admixture (Figure 5.20) and test for the presence of such a trend in the simulations of Chapter 4 (Section 4.3), we performed a linear regression of proportion Native ancestry versus inferred date since admixture, including only individuals inferred to have a single date of admixture between two sources (e.g. case (*i*) described in Section 5.2.4) that are best represented by Native and European reference groups.

5.3 Results

In this section, I first describe ancestry estimations using conventional allele-frequency-based approaches and some of their limitations to estimate ancestry at subtler levels. Then I show how SOURCEFIND estimates presented at a continental scale highly correlate with the ancestry estimates from allele-frequency-based methods. Finally, I present the results for sub-continental ancestry estimations as well as the estimation of the times since admixture and the sources involved in these admixture processes, thoroughly discussing how haplotype-based methods outperform previous approaches and which historical processes are likely to explain the observed results.

5.3.1 Allele-frequency-based approaches cannot infer sub-continental ancestry accurately

In the first CANDELA report (Ruiz-Linares et al. 2014), Sub-Saharan African, European and Native American ancestry proportions were reported using 30 AIMs, carefully chosen to capture genetic differences between these continental groups. Additionally, in Adhikari et al. 2016 we reported the same information for a supervised analysis at K=3 using 93,328 SNPs, showing the consistency of these reports and confirming the prevalent intra and inter-population variation in ancestry in these five Latin American population.

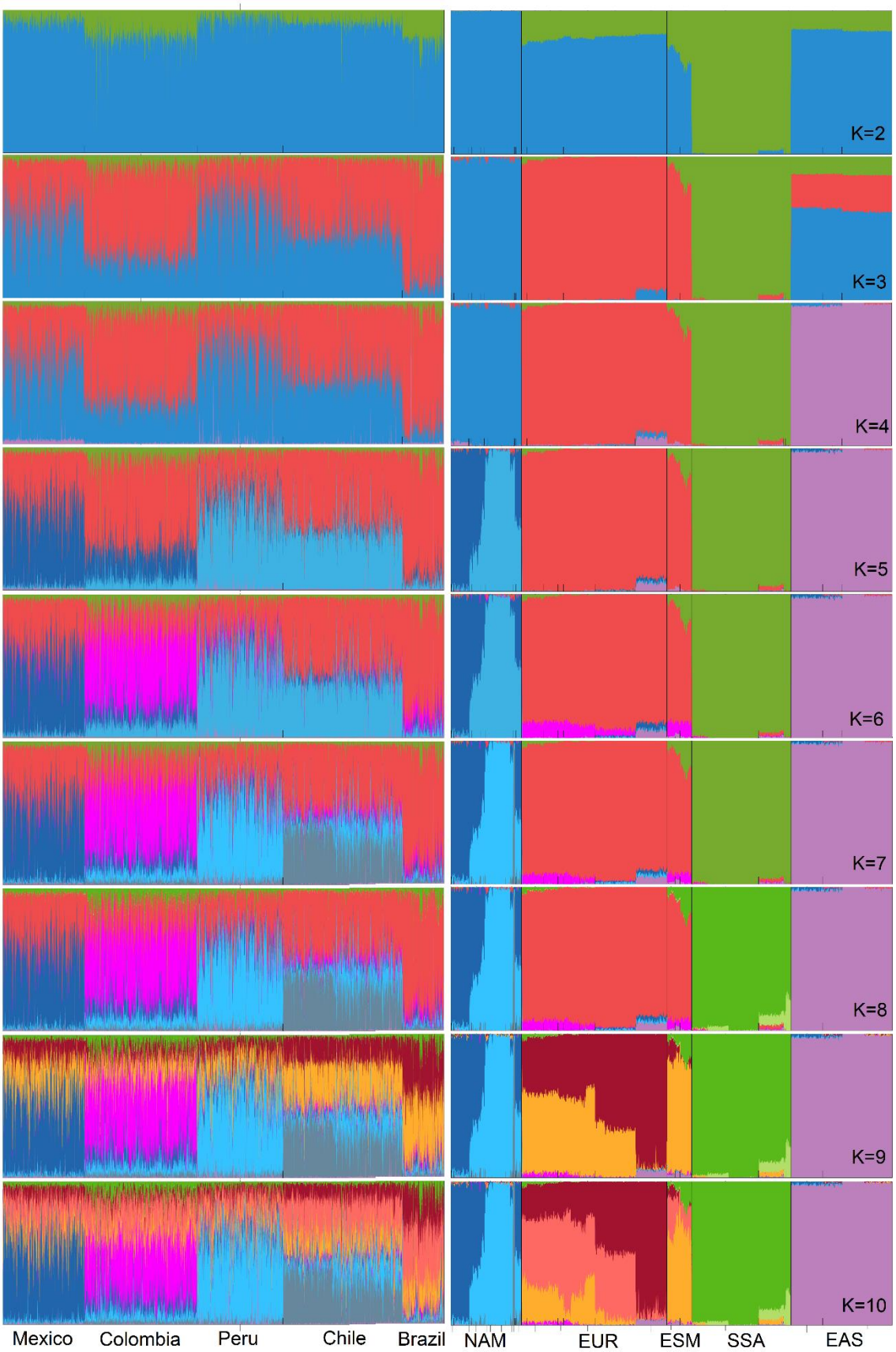

Figure 5.1. Unsupervised ADMIXTURE analysis in the CANDELA dataset. Detailed description of the reference clusters in Chapter 3 (Section 3.8.1 and Figure 3.4). Adapted from Chacón-Duque et al. (2018). Script provided by K. Adhikari.

For the unsupervised admixture analysis (Figure 5.1), main continental groups present in the sample are clearly defined at K=3 (European, Native American and Sub-Saharan African) and K=4 (East Asia). All CANDELA countries have considerable amounts of the first three components. The highest Native American and the lowest European mean ancestry are present in the Peruvian sample (64.2% and 30.6% respectively), while the Brazilian sample shows the lowest Native American and the highest European mean ancestry (8.4% and 83%, respectively). As described in Chapter 1 (Section 1.6), the Brazilian sample was collected in a region with recent European migration, hence the high European-like contribution. African mean ancestry is the lowest out of these three continental components, ranging from 2.7% in Chile to 9.4% in Colombia. PC1 and PC2 show equivalent information with all Latin Americans falling between the axes defining this three major groups (Figure 5.2).

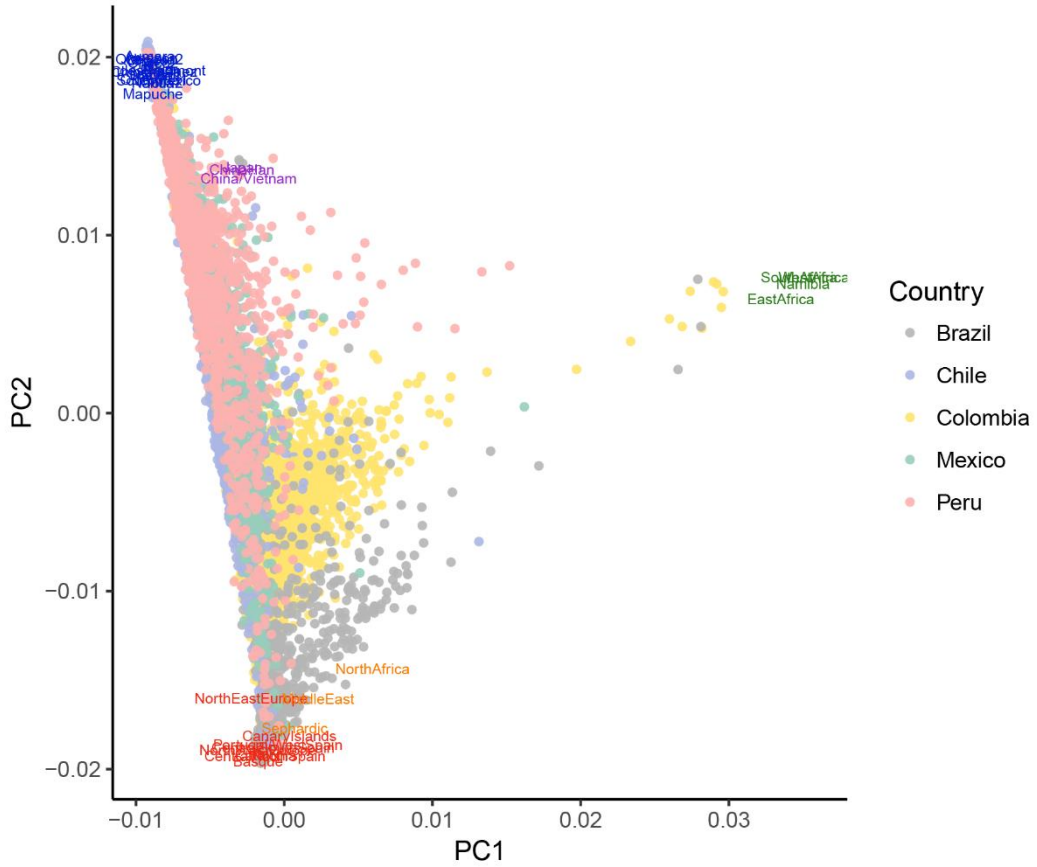

Figure 5.2. Principal component analysis of the merged CANDELA + reference populations' dataset.
(Part 1 of 3). Dots are shown for all CANDELA individual (coloured by country). For the reference population individuals a label has been placed at the median PC scores for that cluster. Detailed description of the reference clusters in Chapter 3 (Section 3.8.2 and Figures 3.5 - 3.10). Adapted from Chacón-Duque et al. (2018). Generated by K. Adhikari.

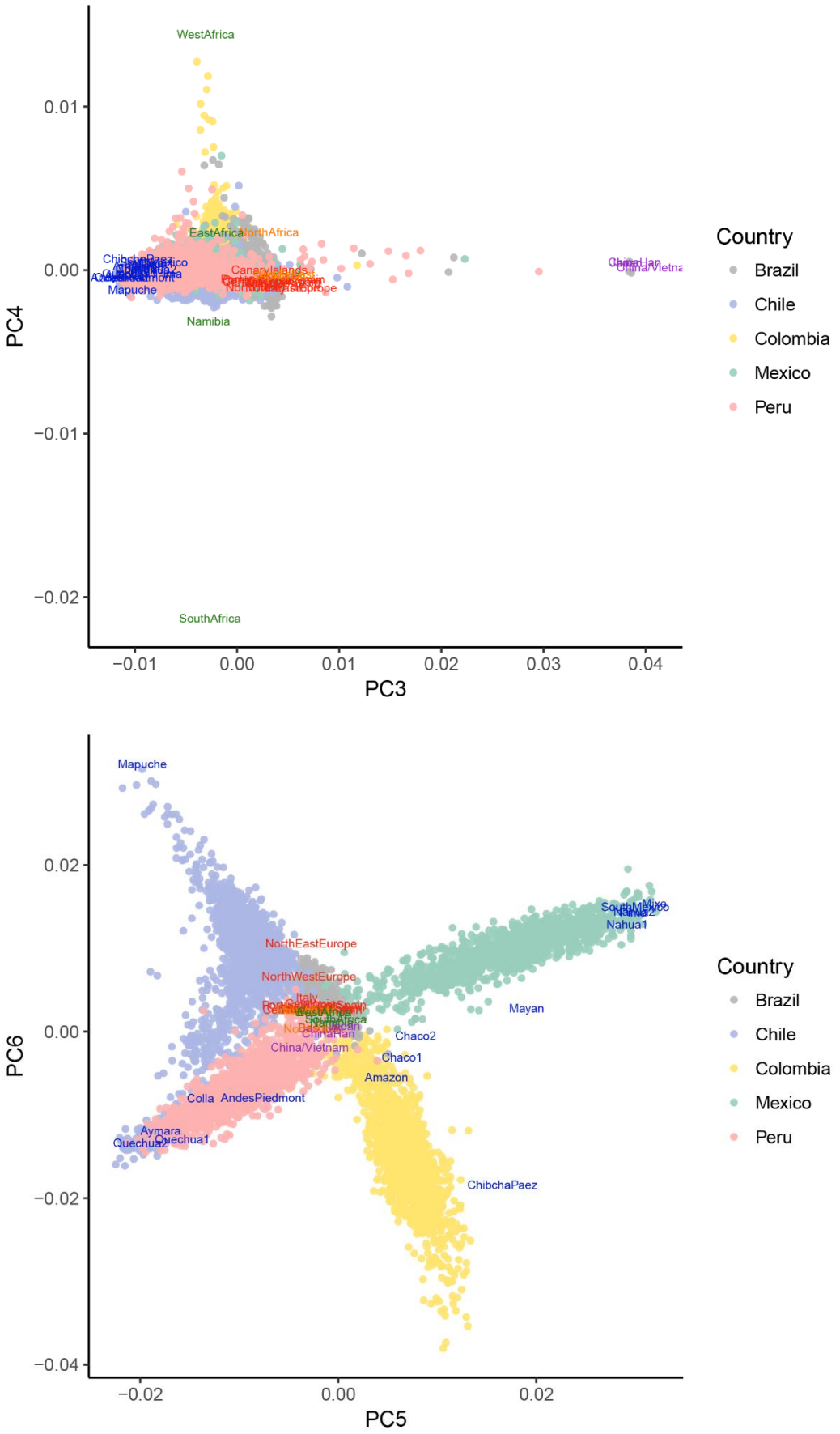

Figure 5.2 (Part 2 of 3). Principal component analysis of the merged CANDELA + reference populations' dataset.

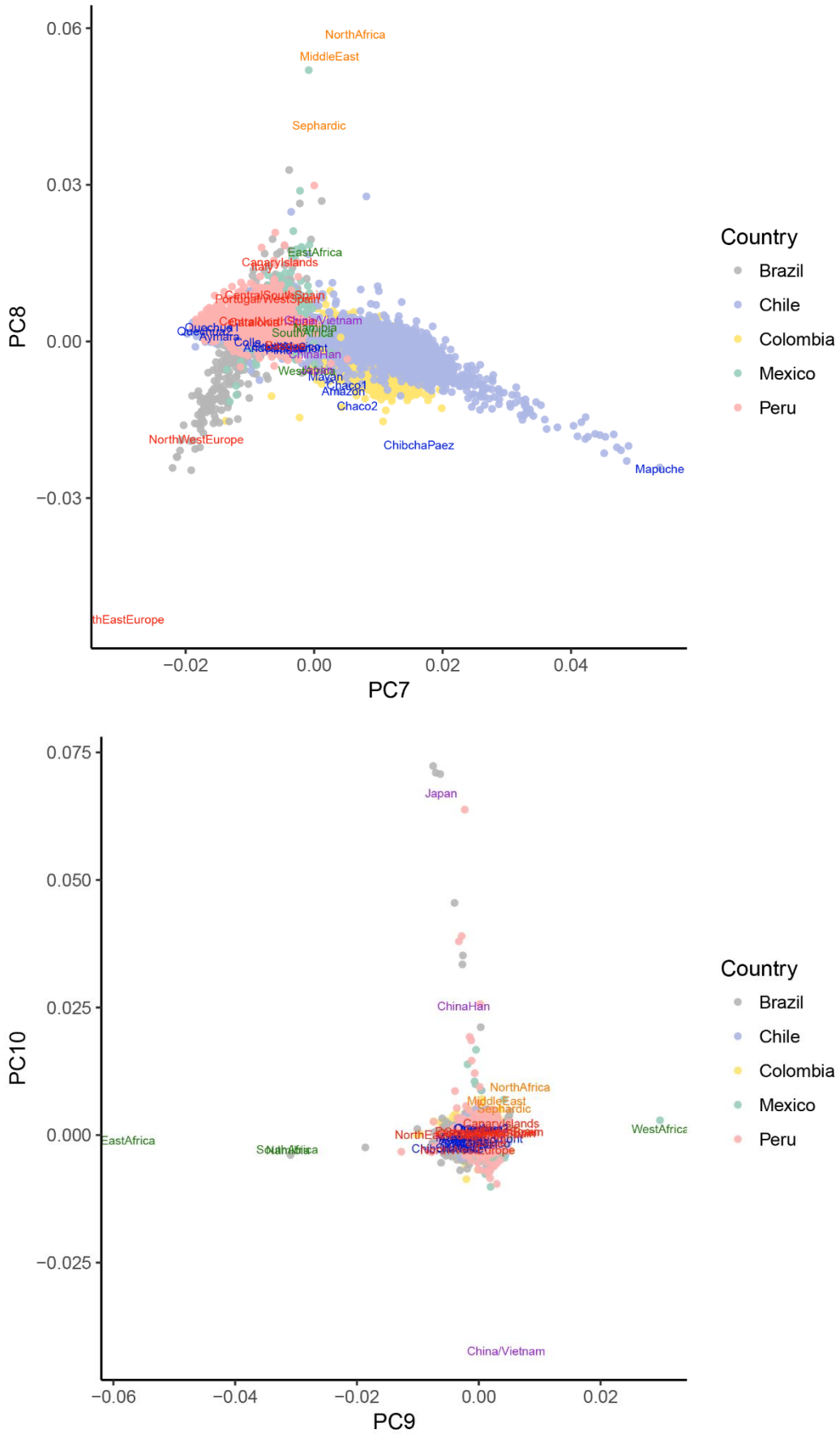

Figure 5.2 (Part 3 of 3). Principal component analysis of the merged CANDELA + reference populations' dataset.

Admixture may be overestimating East Asian Ancestry at lower levels (K=4), as Native Americans and CANDELA samples from Mexico show a consistent but marginal amount of East Asian Ancestry (2.9%, Interquartile Range (IQR) = 2.1 - 3.2 %). However, it is well known that – relative to South American populations – North American Native populations are genetically most similar to Asian populations relative to other world-wide groups (Wang et al. 2007). Thus, as we did not include Siberian samples in this analysis, the East Asian signal in Native Americans may represent an ancestral North-east Asian population, which is suggested by the fact that this marginal amount of ancestry is homogenous across individuals (Chapter 3, Section 3.8.1). However, some individuals show considerably higher values of East Asian ancestry (especially in Peru), more likely indicating a direct and more recent ancestry in these individuals, which is supported by haplotype-based estimates as described in sections 5.3.2.1 and 5.3.2.6. East Asian ancestry is also differentiated in PC3 (Figure 5.2).

The first level of sub-structure in Native American ancestry is inferred at K=5 and PC5, displaying the Mesoamerica-to-Andes cline described in Chapter 3 (Sections 3.8.1 and 3.8.2). Although the proportions from each of the two clusters composing this cline vary across CANDELA samples considerably, clear patterns are present within each country, providing the first line of evidence of a likely Native American genetic sub-structure reflected in the current-day admixed Latin Americans (section 5.3.6.1). This variation across populations has been addressed by previous studies, including one from the Ruiz-Linares Lab (Conley et al. 2017; Homburger et al. 2015; Moreno-Estrada et al. 2013; Wang et al. 2008). The Mexican sample predominantly shows ancestry matching Mesoamerican samples in the reference dataset (mean 55.7%), Peru and Chile display affinity with Andean groups, and Colombia and Brazil have a mixture of both components (with more Mesoamerican in Colombia). The Mapuche component detected at K=7 and PC6 is present primarily in Chile. Interestingly, at PC5 (and to a lesser extent in PC7) the Colombian samples skew away from the axis that points at the sampled Native American groups that are most likely to be related to the indigenous ancestors of current Colombians (*ChibchaPaez*). This skewing could be related to the lack of sampling of more Native American groups related to the original Native admixing source, e.g. due to the extinction of the ancestral population

or a strong effect of genetic drift between that original population and their present-day descendants (as discussed in the next paragraph, a large proportion of the Colombian sample is thought to be part of a genetic isolate).

The component detected at K=6 is especially predominant in Eastern Antioquia in Colombia (ranging between 70 and 100%), and its presence is associated with a reduction of both European and Native American ancestries inferred for smaller Ks. This population has been widely reported as a genetic isolate (Bedoya et al. 2006; Carvajal-Carmona et al. 2000), and the appearance of this ancestral component at a low K is probably due to a strong founder event. Only 111 of 1713 (6.5%) Colombians showed less than 20% from this component, while all individuals outside Colombia had <15%. This component could also be related to the skew seen in Colombians in PC5 and PC7 that shifts these individuals outside the axis of Native American variation. Considering that this component does not represent a single ancestral population, I assume ADMIXTURE results after this point need to be interpreted with caution, as the population structure and the proportions inferred for every K beyond this point cannot be interpreted simply as the product of admixture.

The additional Southeast Sub-Saharan African component arising at K=8 (Chapter 3, Section 3.8.1) is virtually absent in Admixed Latin Americans, as only 11 individuals in the total sample show between 2% and 4% of this component, corroborated by results in PC9 where only a few individuals tend to cluster close to *SouthAfrica* and *Namibia* groups. The Mediterranean-like component at K=9 and the Basque-like component at K=10 (chapter 3, section 3.8.1) are also present in all CANDELA populations, but it is not clear whether these components could represent three different ancestral sources or just different patterns of population structure in Europe and the Mediterranean region. Interestingly, PC8 seems to be more informative about the Mediterranean-to-North East Europe cline as it clearly shows some individuals lying close to Mediterranean populations, while a considerable number of Brazilians cluster with North-western European populations. Overall, unsupervised ADMIXTURE and PCA results are highly correlated and in general seem to show less resolution than haplotype-based methods as I demonstrate later (section 5.3.6, Lawson et al. 2012).

I also performed a series of supervised ADMIXTURE analyses (Figure 5.3), aiming to evaluate the resolution of the software between East / South Mediterranean and European populations, as though the unsupervised analysis is not able to detect a single individual entirely assigned to the homogeneous Mediterranean or sub-continental European component. When inferring ancestry from these two sources separately (Figure 5.3.A), the estimations are totally different compared to unsupervised results and the assignments are clearly inaccurate. For instance, in Colombia almost all of the European component is replaced by the Mediterranean one, which is inconsistent with unsupervised ADMIXTURE, PCA and haplotype-based methods (Section 5.3.6) and known history (Boyd-Bowman 1976; Sánchez-Albornoz 1994). When combining East / South Mediterranean and European sources into a single group (Figure 5.3.B) and when excluding completely the former (Figure 5.3.C) the results seem unaffected; this can be due to the small sample size of our Mediterranean sources compared to the European ones. Given these results, I did not try any sub-continental ancestry estimation based on supervised analyses.

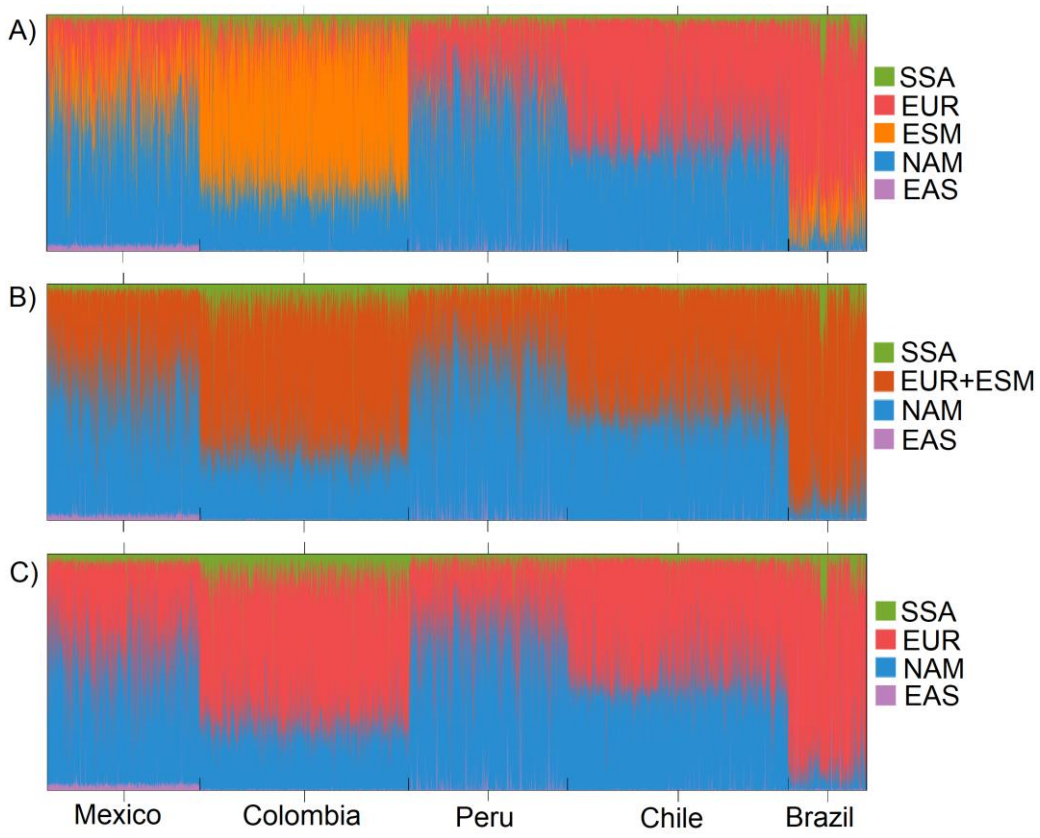

Figure 5.3. Supervised ADMIXTURE analysis in the CANDELA dataset.
(**A**) 5 continental groups, (**B**) 4 groups combining European and East / South Mediterranean sources, (**C**) 4 groups (excluding Mediterranean sources). More details in Section 5.2.3. Script provided by K. Adhikari.

5.3.2 Increasing resolution with haplotype-based approaches

The simulations described in Chapter 4 suggest that, using haplotype-based approaches, I can reliably identify sources and proportions of sub-continental ancestry in single individuals with admixture analogous to the CANDELA individuals. In this section I describe the ancestry estimation results obtained with haplotype-based methods and compare them to those obtained using allele-frequency-based approaches.

5.3.2.1 Continental ancestry estimations with SOURCEFIND and ADMIXTURE are highly correlated

Continental ancestry estimates obtained with ADMIXTURE and SOURCEFIND are highly correlated (Figure 5.4). This correlation is high even in the case of East Asian ancestry ($r > 0.96$), which is clearly overestimated by ADMIXTURE at lower levels (section 5.3.1). The mean East Asian ancestry estimated for Mexico using SOURCEFIND is only 0.2%, suggesting that the ancestry seen with ADMIXTURE could be related to ancestral relationships and lack of resolution between Northern Native Americans and East Asians not related to recent admixture processes.

One further difference between methods is that European and East / South Mediterranean ancestries as estimated by SOURCEFIND are equivalent to the unsupervised European component detected using ADMIXTURE. Given that ADMIXTURE is not able to distinguish ESM as a separate component and describes it as a mixture of European and Sub-Saharan African sources (Chapter 3, Section 3.8.1), it is likely that some of the ancestry related to ESM populations in Latin America could potentially be identified as Sub-Saharan African ancestry by ADMIXTURE. This could explain the decrease of mean SSA ancestry estimated by SOURCEFIND (3.7% (IQR= 0-4.5)) compared to the one estimated by ADMIXTURE at K=4 (5.1% (IQR= 1.7-6)), with a similar trend observed when performing supervised ADMIXTURE analyses defining a Caucasian-like cluster with and without ESM sources (mean SSA ancestry: 3.8% and 4.9% respectively).

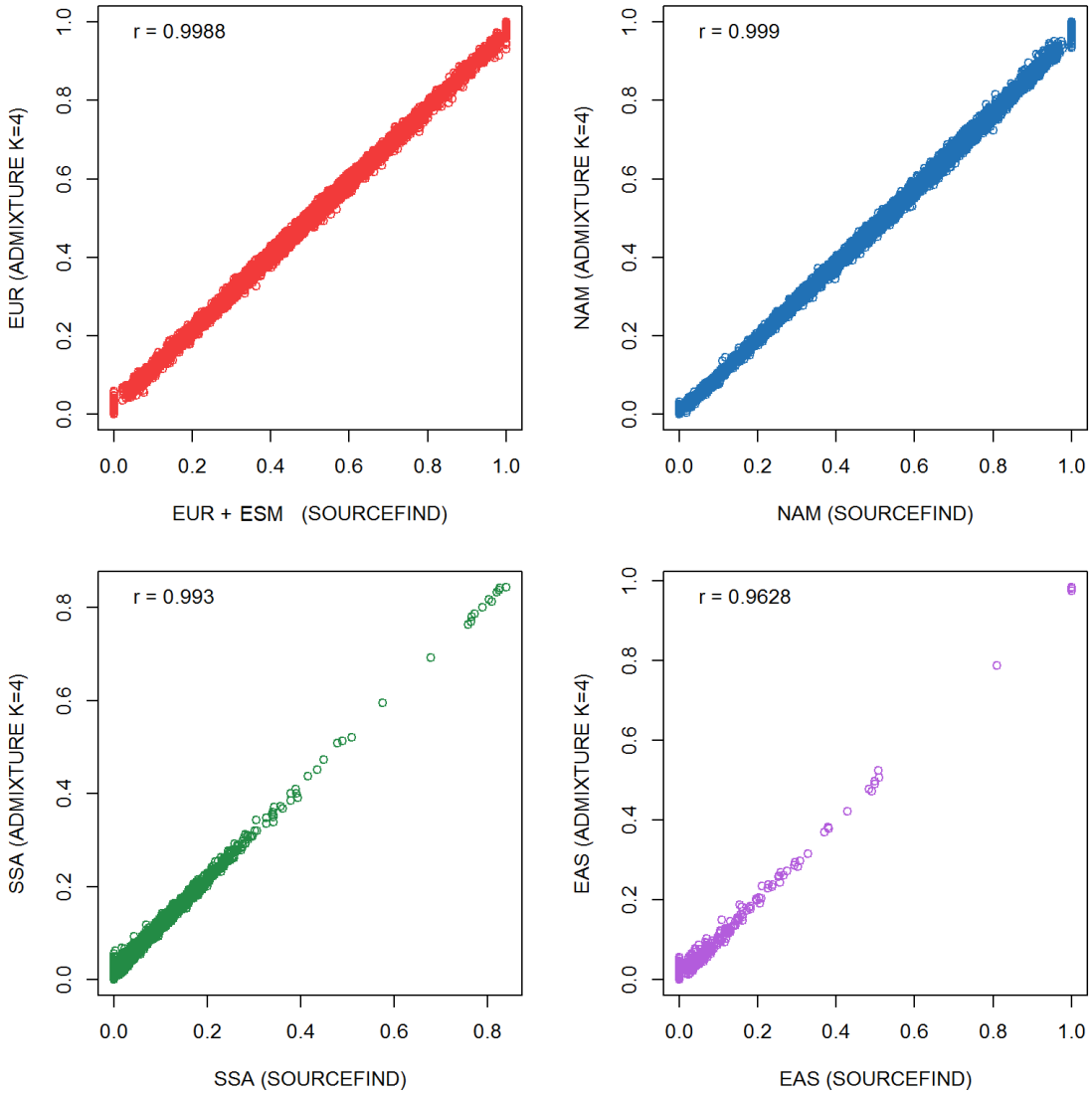

Figure 5.4. Comparison of continental ancestry estimates for the CANDELA sample obtained using SOURCEFIND or ADMIXTURE

5.3.2.2 Pre-Columbian Native American genetic sub-structure is mirrored in Latin Americans

There is a clear correspondence between the location of current Native American populations and the Native ancestry sub-components in the admixed individuals, suggesting a scenario in which local Native populations interbred extensively with immigrants at the onset of the colonization and where the Native American populations have continued inhabiting the same areas. As Native Americans show high levels of genetic structure, relative to other continental populations, previous genetic analyses have demonstrated that pre-Columbian Native American popu-

lation structure is detectable in admixed Latin Americans, even with allele-frequency-based approaches, as discussed at the end of this section. The results presented here add to this by showing a sharp regional differentiation between the five countries examined (Figure 5.5), and are supported by the realization that the individual variation in Native American sub-components' proportions (using geographic location of participants' birthplaces) matches to the genetic profiles of the Native Americans used as surrogates sampled in the surrounding areas (Figure 5.6).

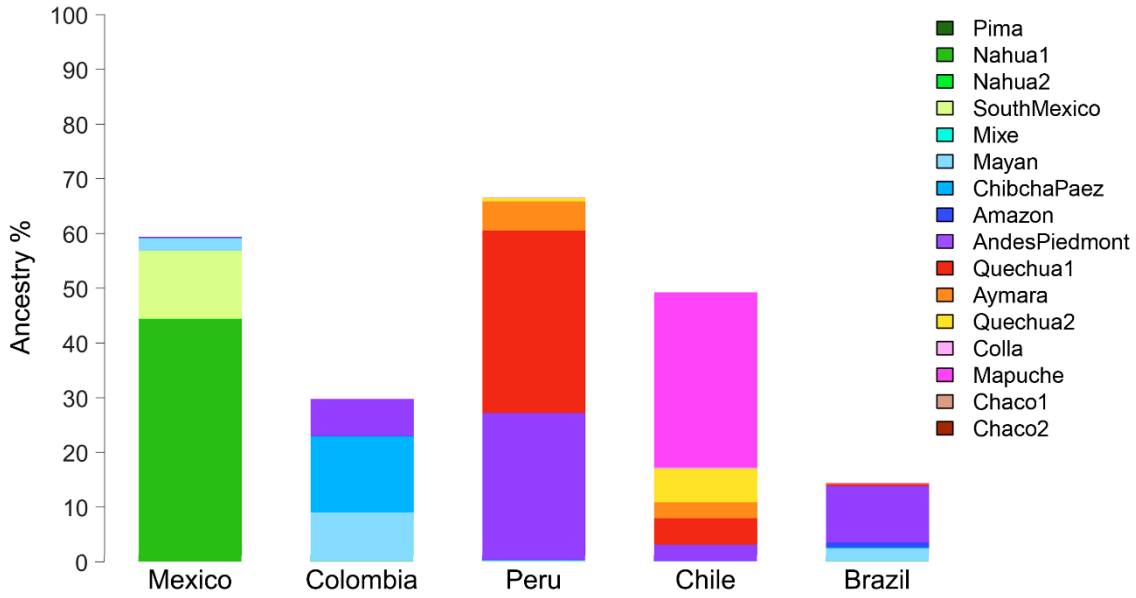

Figure 5.5. Proportion of Native American ancestry sub-components inferred with SOURCEFIND, across all individuals with >5% total Native American ancestry.
In each sampled CANDELA country (using the same colour scheme of the 35 groups shown in Figures 3.3 and 5.6.Total sample sizes for each country are: Mexico (N=1,288), Colombia (N=1,713), Peru (N=1,284), Chile (N=1,891), and Brazil (N=676). Adapted from Chacón-Duque et al. (2018). Generated by J.C. Chacón-Duque and K. Adhikari.

The Mexican sample shows a strong differentiation in Native American ancestry compared to South America. The Native American component is subdivided into three regional sub-components: a predominant Nahua-like sub-component mainly present in northern and central Mexico (*Nahua1*; 44.2%, IQR=29.3-61.8%), one related to Natives of South Mexico widely matching people from the same area (*SouthMexico*; 12.4%, IQR=0-14.1%), and a Maya-like component mostly present in Mexicans from the Yucatan Peninsula (*Mayan*; 2.2%, IQR=0-0%). This result is consistent with previous reports (Moreno-Estrada et al. 2014; Romero-Hidalgo et al. 2017). Moreno-Estrada et al. (2014) characterized the fine-structure of Mexican Native American and admixed populations and reported a

similar trend. Even though they had a broader coverage of the country, the methods applied did not allow the quantification of sub-continental ancestry. From the six clusters they detected, three are related to those here described (Northern, Southern and Mayan) and the other three (Seri, Tojolabal and Lacandon), which are restricted to small geographical areas, resemble likely drifted Native Americans and/or admixed individuals from geographic locations not covered by the CANDELA sampling effort, performed primarily in Mexico City. This finding has been replicated by Romero-Hidalgo et al. (2017).

In Colombians, Native ancestry is subdivided into three sub-components. The principal one is represented by Chibchan-Paezan Natives from Colombia and lower Central America and is more prominent in North-western Colombians (*ChibchaPaez*; 14%, IQR=9.8-16.8%). According to a recent study using the population CLM (Colombians in Medellin) from 1KGP has found that the closest Native American populations to this sample (which is located in the same city where the CANDELA sample was conducted, see Chapter 1) are Embera and Waunana, two of the populations included in our Chibchan-Paezan reference group (Chapter 3, Table 3.1).

The other two components are not related to Native American surrogates from the country, probably reflecting distantly related ancestral groups that are either not represented in the Colombian Native American surrogates included or ancestors from different geographic areas outside of present-day Colombia. The second most prevalent component is represented by the Central American Maya and is widespread through the country (*Mayan*; 8.9%, IQR=0-16.3%). Different anthropological studies have suggested the cultural diversity of Native American populations in Colombia being likely influenced by continued migrations from Central America (Gómez 1970; Reichel-Dolmatoff et al. 1998; Rivet 1943). Finally a Peruvian Andean-Piedmont component (represented by samples from northern Peru) is present and especially predominant in Southern Colombians, coinciding with the northernmost expansion achieved by the Inca Empire in pre-Columbian times (*AndesPiedmont*; 6.7%, IQR=0-11.5%).

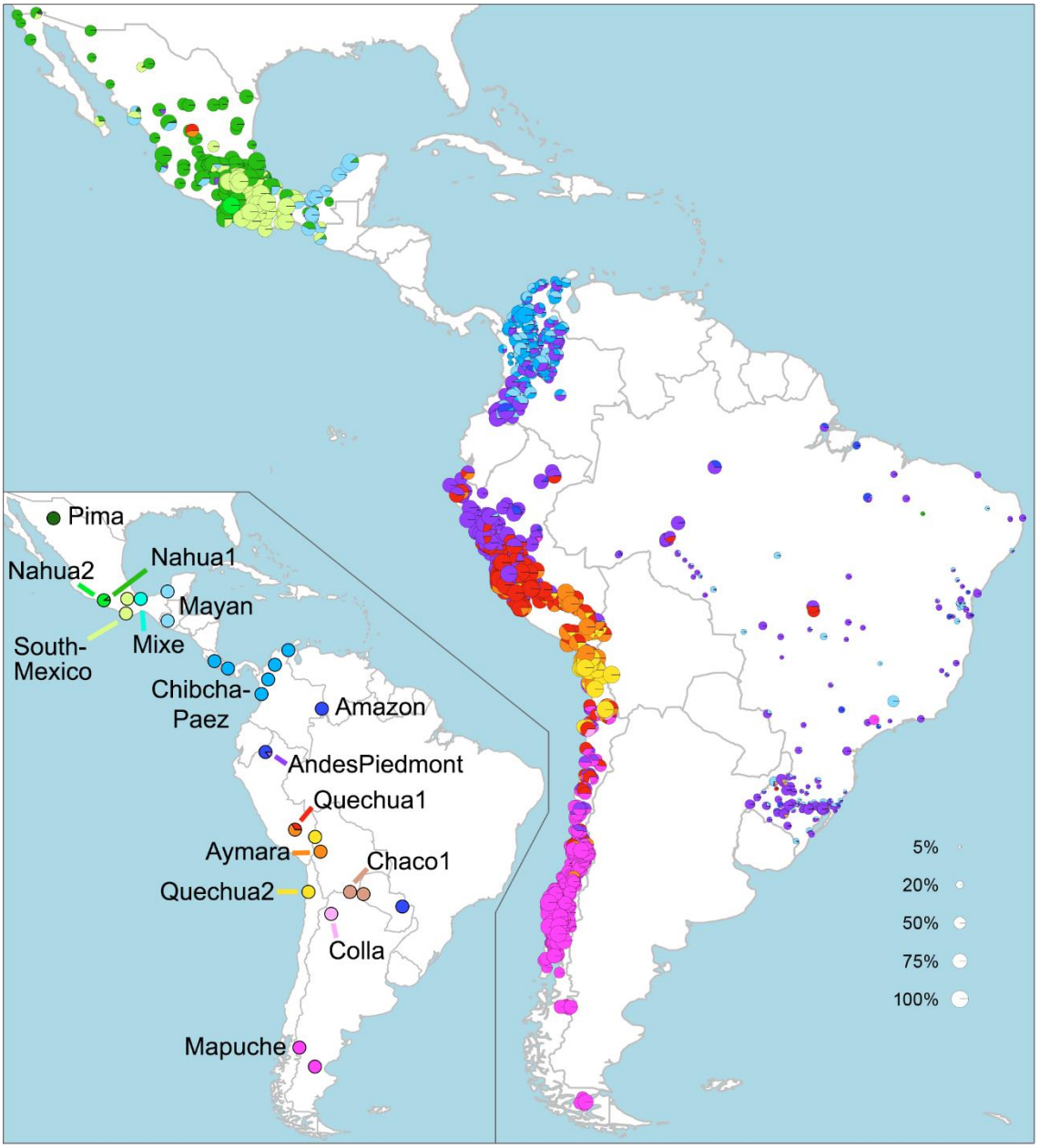

Figure 5.6. Geographic distribution of Native American ancestry sub-components in CANDELA individuals.
Each pie in the main map corresponds to an individual placed according to their birth-place, and shows the proportion of ancestry that individual matches to each regional source group (colours) in the inset map. Pie sizes are proportional to the total ancestry these individuals match to all regional source groups in the legend, with individuals only depicted if their total Native American ancestry is >5%. Since many individuals share the same birthplace, jittering (addition of random noise to the coordinates to avoid overlap of data points) has been performed based on pie size and how crowded the area around a pie is. Pies in the inset map indicate the approximate geographic location of the Native American reference populations (Fig 3.1) that were included in the set of 35 surrogate groups and the colouring represents the proportion of individuals from that population in one of the 35 groups (excluding *Chaco2* as it does not contribute >5% to any individual). More details on the inset map in Figure 3.3. Adapted from Chacón-Duque et al. (2018). Generated by JC Chacón-Duque and K Adhikari.

Native American ancestry in Peru is subdivided into four sub-components, all of them related to Central Andean Natives. The predominant sub-component is Quechua-related showing the highest values in central Peru (*Quechua1*; 33.3%, IQR=0.1-54.7%), followed by a Peruvian Andes-Piedmont sub-component concentrated in Northern Peruvians (*AndesPiedmont*; 26.8%, IQR=0-46.1%), a small Aymara sub-component mostly seen in Southern Peruvians (*Aymara*; 5.4%, IQR=0-0.4%) and a marginal sub-component showing its higher proportion in the border with Chile (*Quechua2*; 0.6%, IQR=0-0%).

These four sub-components are also present in Northern Chile but in different proportions (the Quechua sub-component shared with Southern Peruvians being the most prevalent (*Quechua2*; 6.1%, IQR=0-0%)), and when added together represent the second highest Native American ancestry in Chile (16.6%, IQR=0-24%). The overlap in Andean Native ancestry in Southern Colombia, Peru and Northern Chile, match the areas that according to historical records were under Inca control in the period of greatest expansion of the empire and coincide with the brief political subdivisions created by the Inca administration (Torero 2005).

The Chilean sample is likely to be the most homogenous sample in terms of Native American genetic ancestry. Except for Northern Chileans, all samples show mostly Mapuche-like Native American ancestry (*Mapuche*; 32%, IQR=21.4-43%). The fact that this surrogate accurately represents the Native American ancestors without the need for other contributions from more diverse populations in the reference panel (i.e. Maya or Andes-Piedmont), even though it is a small (N=5) and highly drifted sample is evidence that most of the admixed Chileans can potentially trace their Native American ancestry to direct ancestors of the modern Mapuches. From historical records, it is widely suggested that these populations were mostly absorbed or exterminated by the admixed populations during colonial times (Crow 2013).

The characterization of Native American sub-components in the Brazilian sample is challenging not only because the average Native ancestry is the lowest across the sampled countries, but also because of the lack of better proxies for the Native American ancestors of modern Brazilians. Therefore, these results need to be taken with caution. Some of the individuals with high levels of Native American ancestry are recent immigrants from other Latin American countries, and for the

rest of the Brazilian samples, Andean-Piedmont ancestry from North-eastern Peru is by far the group that best represents their Native ancestors, which could eventually suggest a common ancestral origin in the Amazon basin. Figure 5.7 provides a more detailed depiction of the Native American sub-components in the Brazilians with >5% Native American ancestry.

Compared to other studies, overall these results not only corroborate previous findings but also increase the resolution substantially. A study led by A. Ruiz-Linares demonstrated for the first time - through the use of microsatellites in diverse Native and admixed samples collected through Latin America - that Native American sub-structure can be detected and that such structure is reflected amongst admixed individuals (Wang et al. 2008). However, they were cautious with their limited ability to estimate ancestry proportions at the sub-continental level and cautioned against interpreting their results as ancestry proportions reflecting underlying admixture processes between Native American populations, but rather as a genetic profile heavily influenced by genetic relatedness.

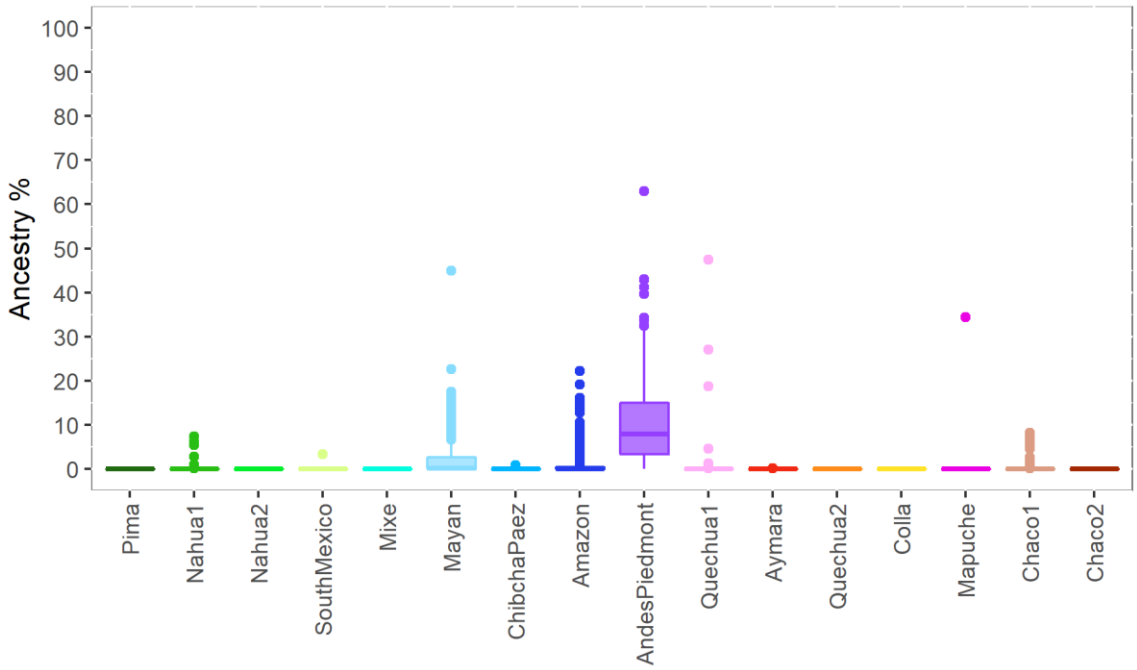

Figure 5.7. Proportion of Native American ancestry sub-components for the 367 Brazilians with >5% Native American ancestry.

The same haplotype similarity estimation approach that I apply here has been previously used in Latin American populations by Montinaro et al. (2015). They calculated sub-continental ancestry with NNLS, which has good resolution for Native American sub-components (Chapter 4, Section 4.2) but used an East Asian

population as a surrogate for Native American ancestry. It has been demonstrated that these populations are not good proxies and can considerably distort the results (de Moura et al. 2016), which is the case with NNLS, which can clearly separate Native American from East Asian populations. Other investigations have made use of Ancestry Specific PCA (AS-PCA, details in Chapter 1, Section 1.3.2) and detect some of the sub-structure here mentioned, but do not allow the estimation of sub-continental ancestry proportions (Browning et al. 2016; Conley et al. 2017; Homburger et al. 2015; Moreno-Estrada et al. 2014; Moreno-Estrada et al. 2013).

Altogether these results provide a high-resolution picture of how Native American population structure is widely reflected in admixed Latin Americans, confirming that pre-colonial genetic sub-structure can be analysed in individuals with high levels of admixture and reporting for the first time estimates of sub-continental ancestry proportions. In Chapter 6, these patterns of population structure are exploited for evaluating association of regional Native American ancestry with variation in physical features.

5.3.2.3 European sub-components trace major migrations back to documented places of origin in the Iberian Peninsula

As described in Chapter 1 (Section 1.2.2), Latin American countries are the consequence of the invasion perpetrated by the two main Iberian Kingdoms, which divided the territory between them early in the colonization process, and the current political borders reflect this history. While the Portuguese territories remained as a single political entity, the former Spanish colonies were fragmented into several countries encompassing southern North America, Central America, and western South America. Various studies have found that the European genetic ancestry of Latin Americans resembles modern Iberian populations (Conley et al. 2017; Homburger et al. 2015; Montinaro et al. 2015; Moreno-Estrada et al. 2013). My analysis provides, for the first time, a breakdown and quantification of specific ancestry sub-components in Latin Americans that are related to the Iberian Peninsula (Figures 5.8 and 5.9), with accuracy demonstrated by the simulations performed in Chapter 4 (Section 4.3.1).

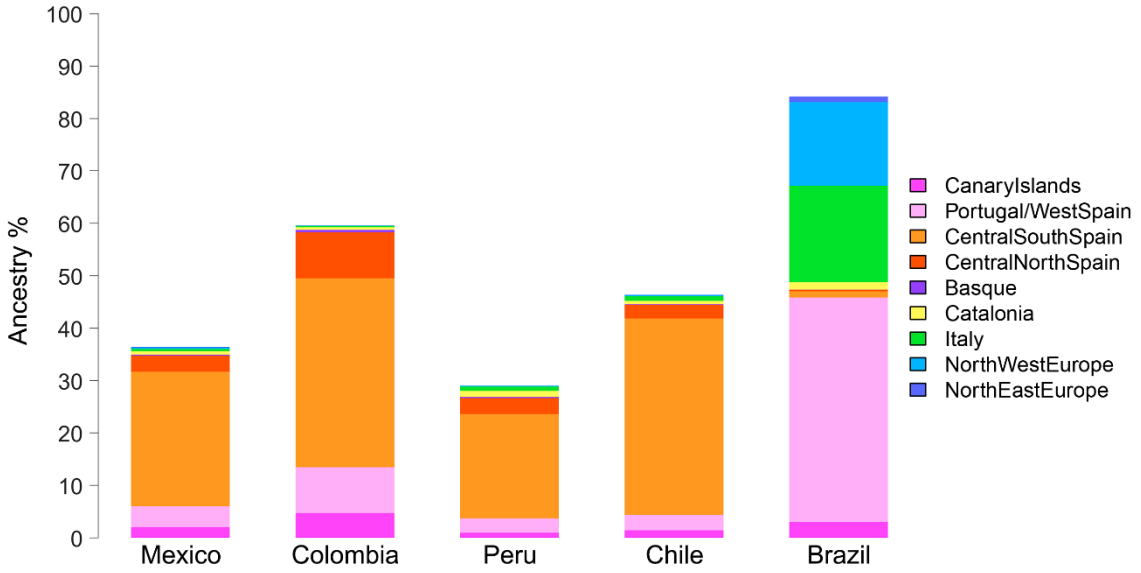

Figure 5.8. Proportion of European ancestry sub-components inferred with SOURCEFIND, across all individuals with >5% total Native American ancestry. Other details in Figure 5.5. Adapted from Chacón-Duque et al. (2018). Generated by JC Chacón-Duque and K Adhikari.

Consistent with historical records (Chapter 1), my analysis of the CANDELA samples shows for the first time a strong regional differentiation in European subcontinental ancestry between Brazil and Spanish America. The most prevalent sub-component in Brazil is represented by the *Portugal/WestSpain* cluster (42.8%, IQR=18.5-65.6%). By contrast in Mexico, Colombia, Peru and Chile, each former Spanish colonies, are predominantly represented by reference populations from South and Central Spain (*CentralSouthSpain*; Mexico: 24.8%, IQR=4-38.5%; Colombia: 36%, IQR=22.8-50.8%; Peru: 19%, IQR=1.5-31%; Chile: 36.4%, IQR=28.6-48.8%).

In the latter, the homogeneity of genetic profiles across countries and the relatively small contribution from other Spanish populations, such as the Basque or the Catalans, could evidence the strong founder effect established at the early stages of the colonial era. AS-PCA analyses also have suggested, based on clustering patterns, that the European component of some Latin American populations could be substantially differentiated from modern Iberian populations, likely attributable to the genetic drift generated by strong founder effects (Moreno-Estrada et al. 2013).

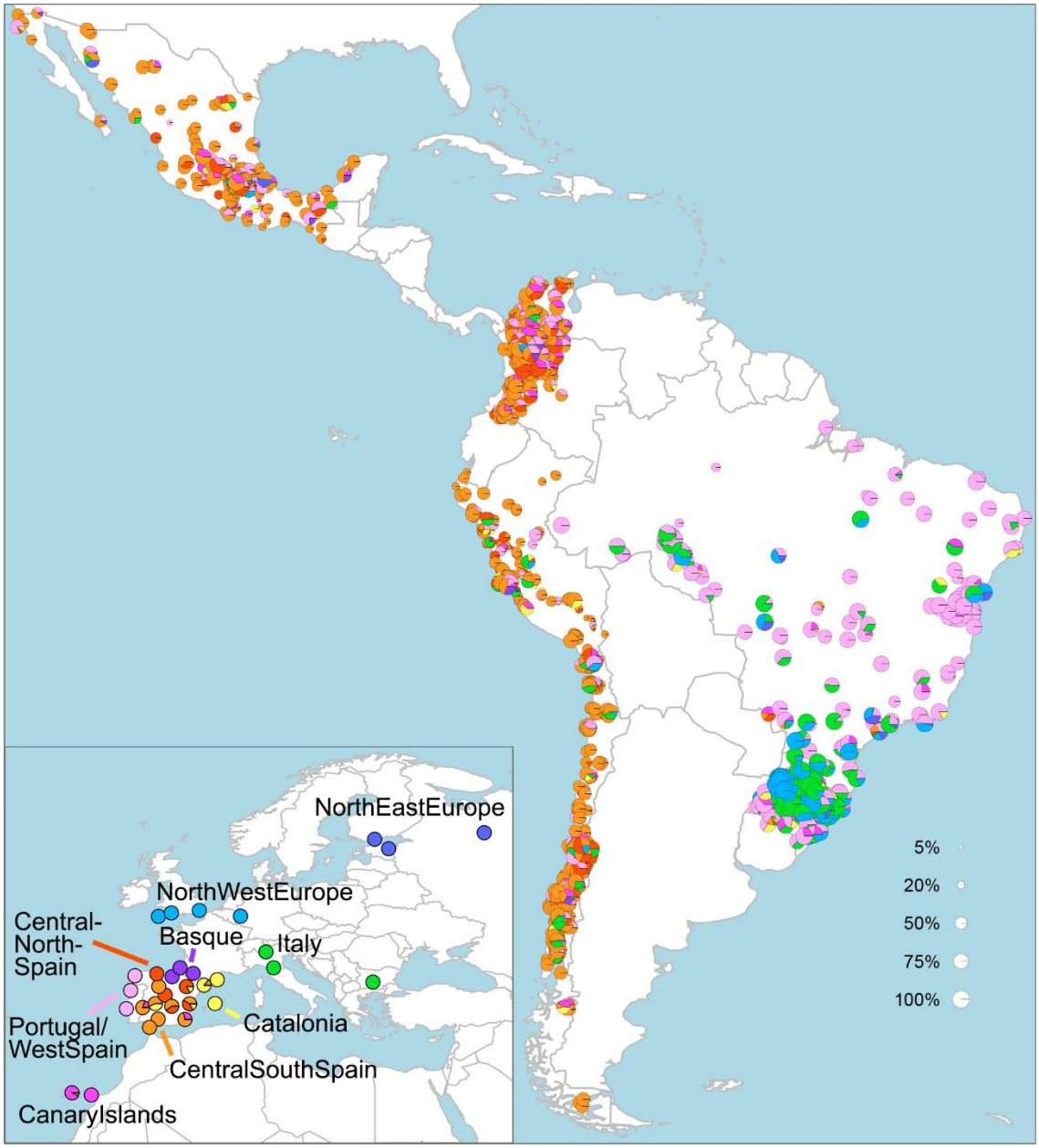

Figure 5.9. Geographic distribution of European ancestry sub-components in CANDELA individuals.
More details in Figure 5.6. Adapted from Chacón-Duque et al. (2018). Generated by JC Chacón-Duque and K Adhikari.

There also is some evidence that partially contradicts these findings. The study that implemented haplotype similarity patterns for inferring sub-continental ancestry in Latin American populations (previous section) claimed significant contributions of Basque and Italian ancestors in a Colombian sample (Montinaro et al. 2015). However, the simulations in Chapter 4 (Section 4.2) show that NNLS is unable to discriminate European sub-components entirely, suggesting that it could be noise related to older ancestral relationships between the populations.

In addition to their Iberian ancestry, the Brazilian sample - particularly the individuals located in the South of the country - (Figure 5.9) have considerable amounts of ancestry matching to the Italian (18.4%, IQR=0-31.7%) and North-western European (16%, IQR=0-23.9%) surrogate clusters. This is consistent with the well documented state-fostered migration of large numbers of immigrants in the 19th century (preferentially to the south of Brazil), with migrants from Germany, Italy, Portugal and Spain, the most common sources of these immigrants (Sánchez-Albornoz 1994).

Overall, these results provide a detailed picture of the European sub-continental ancestry on Latin America, differentiating among contributions from several populations within the Iberian Peninsula accurately. The relative homogeneity observed across Spanish America contrasts with the high genetic structure evidenced in Native American ancestry.

5.3.2.4 Widespread South/East Mediterranean ancestry is detected

The eventual involvement of people of East and South Mediterranean - particularly of Sephardic - origin in the colonization process has been extensively discussed and difficult to probe, given the scarcity of historical records (Sachar 1994). These migrations have been suggested to be somewhat clandestine, as the colonization of the American continent coincided with the upsurge of religion-based discriminatory policies propitiated by the increasing power of the Christian rulers, including prohibiting emigration to the newly established colonies (Chapter 1). Previous evidence based on genetic data, mostly from rare mutations causing Mendelian disorders (Berg et al. 1994; Ellis et al. 1998; Mullineaux et al. 2003) and Y-chromosome haplogroups (Carvajal-Carmona et al. 2000; Velez et al. 2012) detected in Latin American populations' samples, support the possibility of a contribution of these populations to the genetic background of Latin Americans (Chapter 1). However, the analyses presented here constitute the first robust assessment of the overall genetic contribution from Mediterranean populations to Latin Americans.

For this purpose, we genotyped reference population samples from the East and South Mediterranean, including individuals self-identified as Sephardic Jews (Chapter 3; Figures 3.1, 5.10 and 5.11). The analyses described in Chapter 3

demonstrate that these three groups are distinguishable according to fineSTRUCTURE, with the simulations performed in Chapter 4 (Section 4.2) indicating that they can also be distinguished from European sources.

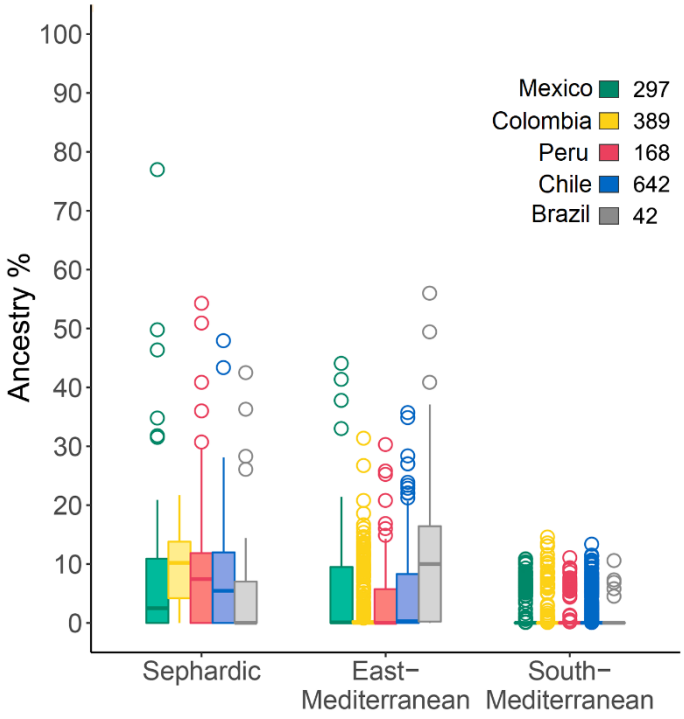

Figure 5.10. Inferred ancestry sub-components in individuals with more >5% Sephardic/East/South Mediterranean ancestry in each of the five CANDELA countries. Adapted from Chacón-Duque et al. (2018). Generated by JC Chacón-Duque and K Adhikari.

According to SOURCEFIND analyses, this Sephardic / East / South Mediterranean contribution is detectable at low-levels in all the countries (Figure 5.10) and is widespread throughout the region (Figure 5.11), with ~23% of CANDELA showing >5%. In these individuals, the most noticeable sub-component is represented by people of Sephardic origin, with an average of 7.3%, while non-Sephardic contributions are significantly lower (East Mediterranean 3.9% and South Mediterranean 1%).

It is likely that some of the individuals with considerable amounts of these sub-components are descendants of recent migrants as confirmed in many cases by the genealogical information available, something common in different Latin American countries but restricted to specific areas (Chapter 1). In fact, for 16 of the 42 individuals with >25% Sephardic or East Mediterranean ancestry, genealogical information confirmed recent ancestry in the Eastern Mediterranean region. In contrast, even though Colombia displays the highest mean Sephardic

component in the sample (Figure 5.10), no recent immigration from this region was documented.

Figure 5.11. Geographic distribution of East/South Mediterranean ancestry sub-components in CANDELA individuals.
More details in Figure 5.6. Adapted from Chacón-Duque et al. (2018). Generated by JC Chacón-Duque and K Adhikari.

These analyses provide evidence that individuals from Sephardic/East/South Mediterranean origin or individuals with high amounts of these ancestries (beyond that found in the sampled present-day Iberian groups) accompanied colonial-era migrants, perhaps at higher levels than suggested by historical records, resulting in a contribution that is widespread across Latin America.

5.3.2.5 Sub-Saharan African ancestry comes mainly from West Africa

As described in Section 5.3.2.1, the average Sub-Saharan ancestry in the full CANDELA sample is relatively low, with only 1,472 (~22%) individuals showing >5% of this ancestry. This is due in part to the biased sampling, as most of the populations of (mainly) African descent are located in the periphery of the countries and also because the CANDELA sampling favoured people of Native American and/or European descent (Chapter 1, Section 1.6).

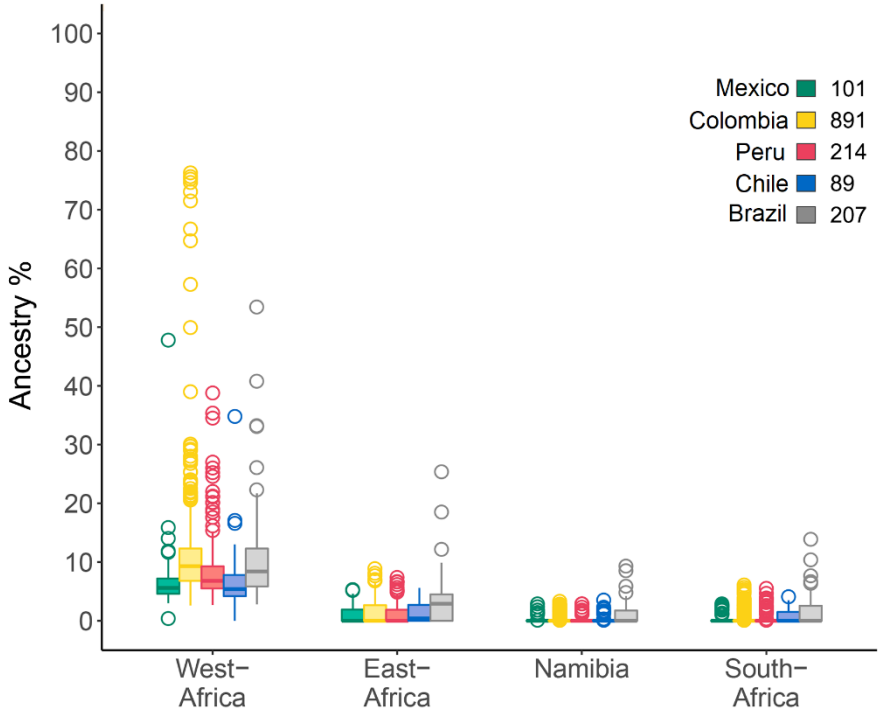

Figure 5.12. Inferred ancestry sub-components in individuals with more >5% Sub-Saharan African ancestry in each of the five CANDELA countries.
Adapted from Chacón-Duque et al. (2018). Generated by JC Chacón-Duque and K Adhikari.

These individuals with >5% Sub-Saharan African ancestry show a higher proportion of the West African sub-component (Figures 5.12 and 5.13), particularly in the Spanish American countries: while the West African sub-component accounts for ~82% of the Sub-Saharan African (SSA) ancestry in these individuals, it only represents ~66% of SSA ancestry in the Brazilians (Figure 5.14). This trend is consistent with historical information indicating that the slave trade to Brazil involved East / South Africa to a greater extent than the Spanish colonies (Kehdy et al. 2015). At the individual level (Figure 5.13), the higher amounts of East /

South African ancestry in Brazil are detected in the south, also consistent with historical evidence (Chapter 1).

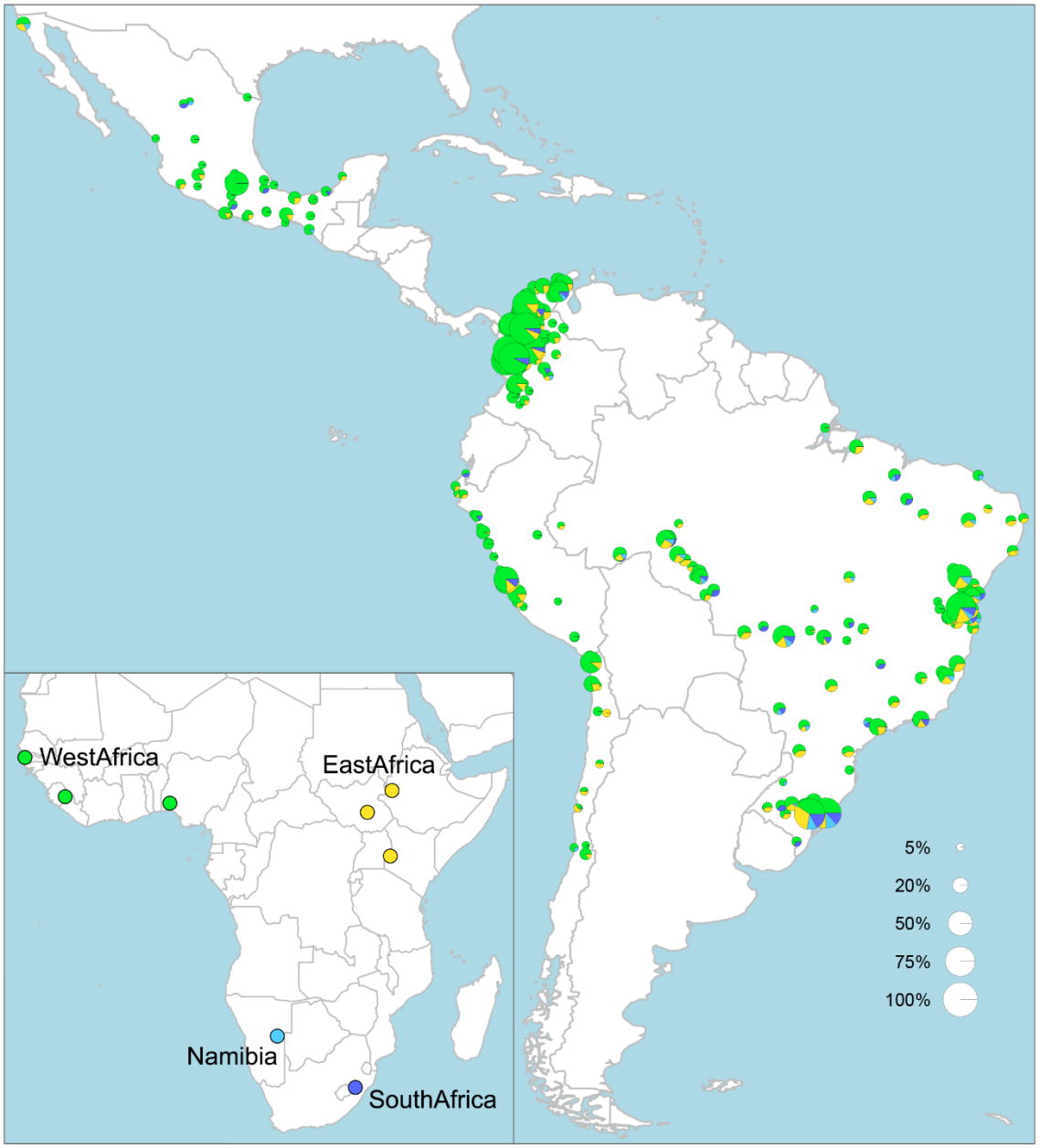

Figure 5.13. Geographic distribution of Sub-Saharan African ancestry sub-components in CANDELA individuals.
More details in Figure 5.6. Adapted from Chacón-Duque et al. (2018). Generated by JC Chacón-Duque and K Adhikari.

Previous analyses of Latin American populations in the Caribbean and Colombia with higher levels of Sub-Saharan African ancestry, have found evidence of two pulses of Sub-Saharan African migration, with the oldest pulse related to coastal West African populations and the newest to Central-West African populations (represented by Nigerian Yoruba populations, widely covered in genetic studies)

(Conley et al. 2017; Moreno-Estrada et al. 2013). I also explored different sub-components within West Africa, as established by the fineSTRUCTURE analyses (Chapter 3, Figure 3.2 and Table 3.4). In the 1,472 individuals with >5% Sub-Saharan African Ancestry (average 12%), the averages of the three West African sub-components are: *WestAfrica1* (Gambia) 2.5%, *WestAfrica2* (Sierra Leone) 0.2% and *WestAfrica3* (Nigeria) 7.1%. It is possible that the higher average of the Nigeria-related sub-component reflects a higher contribution of the most recent (involuntary) migrants from Sub-Saharan Africa in the genetic make-up of the CANDELA sample. However, it could also be that the African ancestors of these Latin Americans can be represented as a mixture of these two putative ancestral components. However, given the low African ancestry in CANDELA, these results need to be taken with caution.

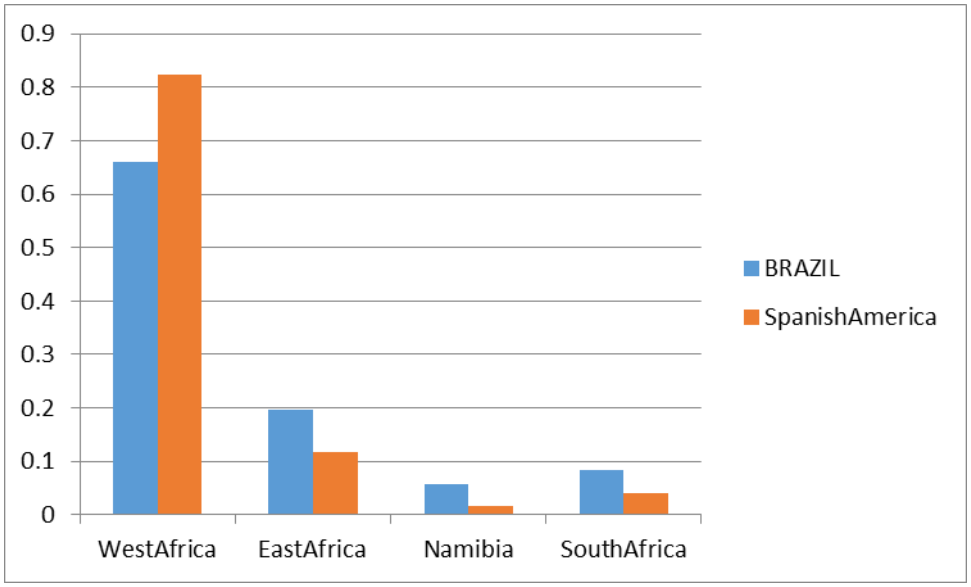

Figure 5.14. Average sub-continental ancestry proportion for the 1,472 individuals with >5% Sub-Saharan ancestry and the Spanish American countries sampled. Generated by JC Chacón-Duque and A Ruiz-Linares.

5.3.2.6 East Asian Ancestry is closely related to Chinese sources

Historical information indicates that some considerable migrations from East Asia took place in Latin America after the abolition of slavery and the independence, with Peru the most popular destination (Crawford and Campbell 2012) and having East Asian genetic ancestry detected previously (Homburger et al. 2015; Sandoval et al. 2013). In this analysis I expand this finding to four other countries

(Mexico, Colombia, Chile and Brazil) and match East Asian ancestry to specific regional sub-components.

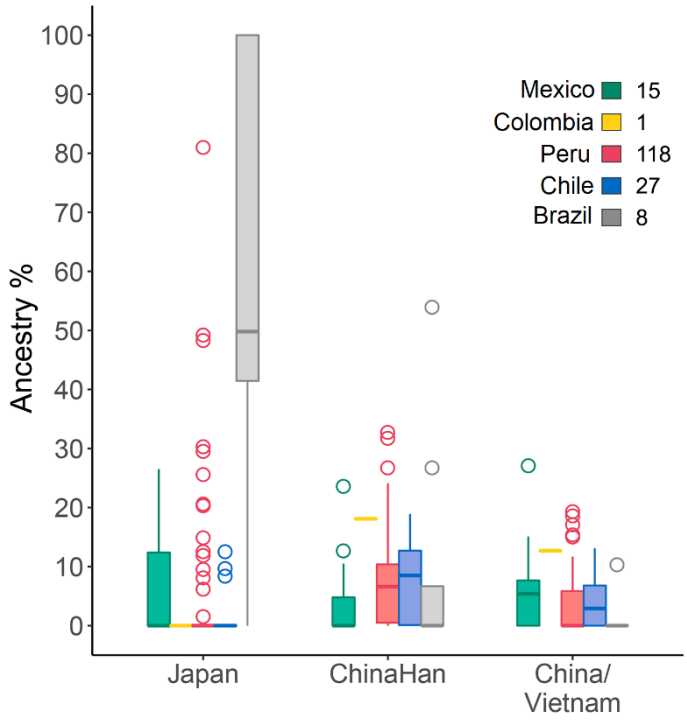

Figure 5.15. Inferred ancestry sub-components in individuals with more >5% East Asian ancestry in each of the five CANDELA countries.
Adapted from Chacón-Duque et al. (2018). Generated by JC Chacón-Duque and K Adhikari.

SOURCEFIND analyses indicate that East Asian ancestry is almost negligible in Mexico (0.24%), Colombia (0.02%), Chile (0.21%) and Brazil (0.82%) and low in Peru (1.4%). However, there are 169 individuals with >5% East Asian ancestry (average of 16.5%): 15 in Mexico, one in Colombia, 27 in Chile, eight in Brazil and 118 in Peru (Figure 5.15). The most common ancestry sub-component relates to the Chinese surrogate clusters (10.7%) and to a lesser extent the Japanese (5.8%). These results match historical records regarding East Asian migrations to Latin America. Although East Asian migrations have been reported since colonial times, associated with the Trans-Pacific routes of the Kingdom of Spain, only in the 19th century was an important influx of immigrants is reported, particularly to Peru. Primarily, these migrations were from Southern China (Guangdong Province), and SOURCEFIND results may be pointing in this direction, as the East Asian ancestors of this group of Latin American are represented, on average, as a mixture of *ChinaHan* (7.2%) and *China/Vietnam* (3.5%) clusters.

Figure 5.16. Geographic distribution of East Asian ancestry sub-components in CANDELA individuals.
More details in Figure 5.6. Adapted from Chacón-Duque et al. (2018). Generated by JC Chacón-Duque and K Adhikari.

The exception to this trend is Brazil, where seven out of eight individuals show exclusively Japanese ancestry (Figure 5.16), including three individuals with 100% Japanese ancestry. The occurrence of a Japanese migration to Brazil in the 20th century is well documented.

5.3.2.7 Sub-continental ancestry estimations are not affected by changes in the reference panel

To evaluate the robustness of ancestry inference when excluding CANDELA reference individuals, G. Hellenthal also ran SOURCEFIND as described in section 5.2.3 on each CANDELA individual after excluding all CANDELA individuals from both the donor and surrogate groups. He furthermore removed any individuals that were excluded from the surrogate groups for other reasons (as described in Chapter 3), so that this analysis contained only 55 donor and surrogate groups. This is one less surrogate group than for the main analysis because the “Germany” surrogate group consisted entirely of CANDELA individuals. As noted in Chapter 2 (Section 2.4.2), for this analysis he used an alternative, more efficient version of SOURCEFIND that used a truncated Poisson prior on the number of contributing surrogates and allowed a maximum of eight surrogates to contribute at each MCMC iteration.

Inferred proportions of ancestry are shown in Figure 5.17. Ancestry matching to European, East/South Mediterranean and Sub-Saharan African groups are largely consistent with results depicted in Figures 5.6, 5.9, 5.11 and 5.13. For Native American ancestry, results are similar across most of the CANDELA sample, but there is a marked decrease in inferred ancestry related to the *AndesPiedmont* and *Quechua1* surrogate groups. This makes sense given that these clusters each contain only one individual after removing CANDELA samples, which is expected to decrease power. The ancestry contributions of these groups are, for the most part, replaced by inferred ancestry matching to other geographically nearby Native surrogate groups.

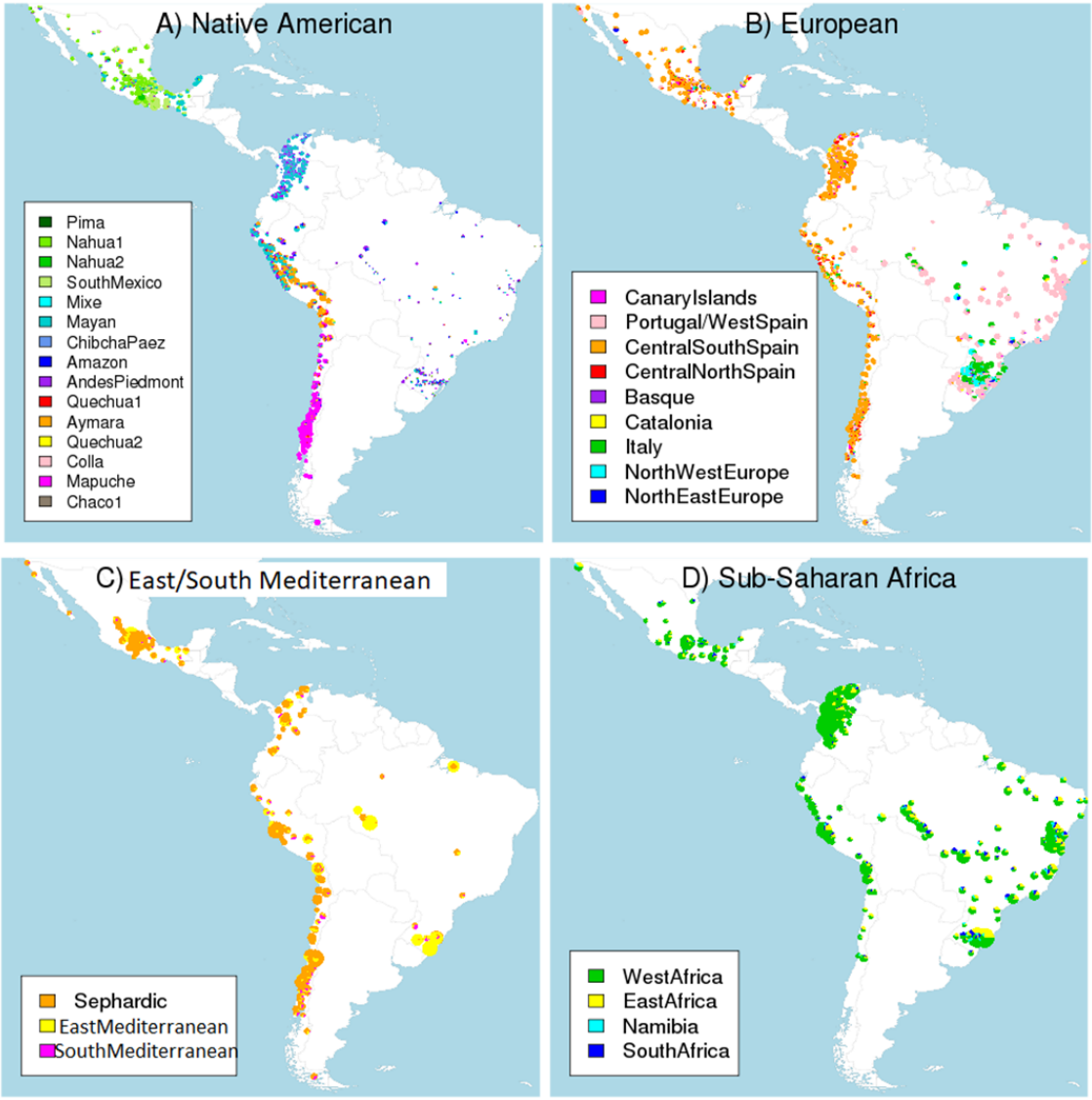

Figure 5.17. Individual pie-maps showing SOURCEFIND analyses when not including any CANDELA reference samples as surrogates or donors.
More details in Figure 5.6. Adapted from Chacón-Duque et al. (2018). Generated by JC Chacón-Duque, K Adhikari and G Hellenthal.

5.3.2.8 Sub-continental ancestry matches genealogical information

As described in Chapter 1 (Section 1.6), CANDELA volunteers provided information about their parents and grandparents origins when possible. Even though these kind of records can be inaccurate, there is a big overlap of self-reported genealogical ancestry and sub-continental ancestry components as can be seen in Table 5.1, supporting the accuracy of these estimations at the individual level in real data.

Table 5.1. Number of individuals reporting a grandparent and/or parent from each region* (columns) and with SOURCEFIND inferred proportion of ancestry (A) 10% and (B) >25% from each reference group** (rows)

(A)

	EEur	**NWEur**	**SEur**	**Iberia**	**SMed**	**EMed**	**WAfr**	**EAsia**	**Other**
total	37	62	43	136	5	21	6	28	60
NE.Eur>10%	**25**	2	2	3	0	0	0	0	0
NW.Eur>10%	15	**24**	3	5	1	0	0	0	16
Italy >10%	10	19	**23**	5	1	9	0	0	9
Iberia >10%	32	61	41	**136**	5	12	6	18	57
S.Med >10%	0	0	0	2	**0**	0	0	0	0
E.Med/Seph >10%	2	13	18	20	1	**16**	6	6	19
SSA > 10%	0	6	4	20	0	2	**6**	0	6
EAS > 10%	0	0	2	0	0	0	0	**28**	2

(B)

	EEur	**NWEur**	**SEuro**	**Iberia**	**SMed**	**EMed**	**WAfr**	**EAsia**	**Other**
NE.Eur>25%	**16**	0	0	0	0	0	0	0	0
NW.Eur>25%	8	**21**	1	3	1	0	0	0	9
Italy >25%	3	11	**20**	4	1	9	0	0	6
Iberia >25%	23	53	33	**131**	5	10	6	12	48
S.Med >25%	0	0	0	0	**0**	0	0	0	0
E.Med/Seph >25%	1	0	5	2	0	**16**	0	2	11
SSA >25%	0	0	0	6	0	0	**6**	0	0
EAS >25%	0	0	2	0	0	0	0	**20**	2

***"EEur": Ukraine, Czech Republic, Czechoslovakia, Finland, Latvia, Poland, Romania, Russia, Yugoslavia, Croatia. "NWEur": Germany, Austria, Belgium, UK, France, Ireland, Sweden, Switzerland, the Netherlands. "SEur": Italy, Greece. "Iberia": Spain, Portugal. "SMed": Algeria, Morocco. "EMed": Lebanon, Turkey, Libya. "WAfr": Senegal. "EAsia": Japan, South Korea, China. "Other": Argentina, Bolivia, Cuba, Guatemala, Paraguay, Uruguay, Venezuela, Canada, USA, India. ******NE.Eur (*NorthEastEurope*), NW.Eur (*NorthWestEurope*), *Italy, Iberia*, S.Med (*SouthMediterranean*). E.Med/Seph (*EastMediterranean + Sephardic*).

5.3.3 Timings and sources of admixture with non-Native ancestors match documented migratory flows

Considering the large variation in individual continental ancestry proportions, times and sources of admixture were inferred for each individual separately. The

simulations described in Chapter 4 (Section 4.3.1, Figures 4.13 and 4.14) corroborate the accuracy of GLOBETROTTER for this inference. Such single individual estimates provide a huge amount of data, allowing us to describe trends in the dates and the inferred sources compared to relying on single estimates per population (Homburger et al. 2015; Moreno-Estrada et al. 2013). Figure 5.18 shows the differences in the distribution of admixture dates for events according to the sources involved, which usually correspond to various historical records documenting migration processes in Latin America.

A total of 8,167 inferred admixture events involved an Iberian source and had a median of 10 (IQR=7-13) generations, corresponding to ~1680CE (1600-1760CE), or about the middle of the colonial period. It is interesting to note that historical information indicates that European migration to Latin America appears to have declined from about the middle of the 17th century onwards (Sánchez-Albornoz 1994). These estimates are also consistent with previous estimations based on genetic data (Homburger et al. 2015). As expected from the historical records documenting recent arrival of other European populations to Brazil (Chapter 1, Section 1.2.4), admixture events involving German or Italian-like sources had a significant skew towards more recent dates (Figure 5.18A; Wilcoxon rank-sum test one-sided p-value=3.3x10^{-8}).

Dates for events involving an East Asian source were also significantly more recent (median = 3; IQR 2-5 generations ago) than those involving European sources (Figure 5.18B; Wilcoxon rank-sum test one-sided p-value<10^{-15}), and consistent with the documented migration of labourers from China during the 19th century and from Japan in the 20th century, primarily to Peru and Brazil, respectively (Chapter 1, Section 1.2.4).

Admixture events involving a Sub-Saharan African source occurred mostly (80%) in individuals with an inferred complex admixture, involving multiple dates and/or more than two groups admixing at approximately the same time (Table 5.2). This suggests that Sub-Saharan Africans started admixing simultaneously with Native Americans and with Europeans in narrow time spans, and contributed to the admixture process later on. It can be seen as evidence for a less extensive admixture than that taking place between Native Americans and Europeans, although it is worth considering that the low representation of African ancestry due to the

sampling protocol may eventually bias the results in this way. These individuals show a 7-fold increase in Sub-Saharan African ancestry, compared to those in which a single admixture event was inferred (Figure 5.19). The distribution of dates involving Sub-Saharan African admixture mostly overlaps with that for Iberian admixture, although a high proportion of recent dates were also inferred (Figure 5.18C), likely reflecting continued episodes of intermixing between Africans and Americans in the regions sampled here.

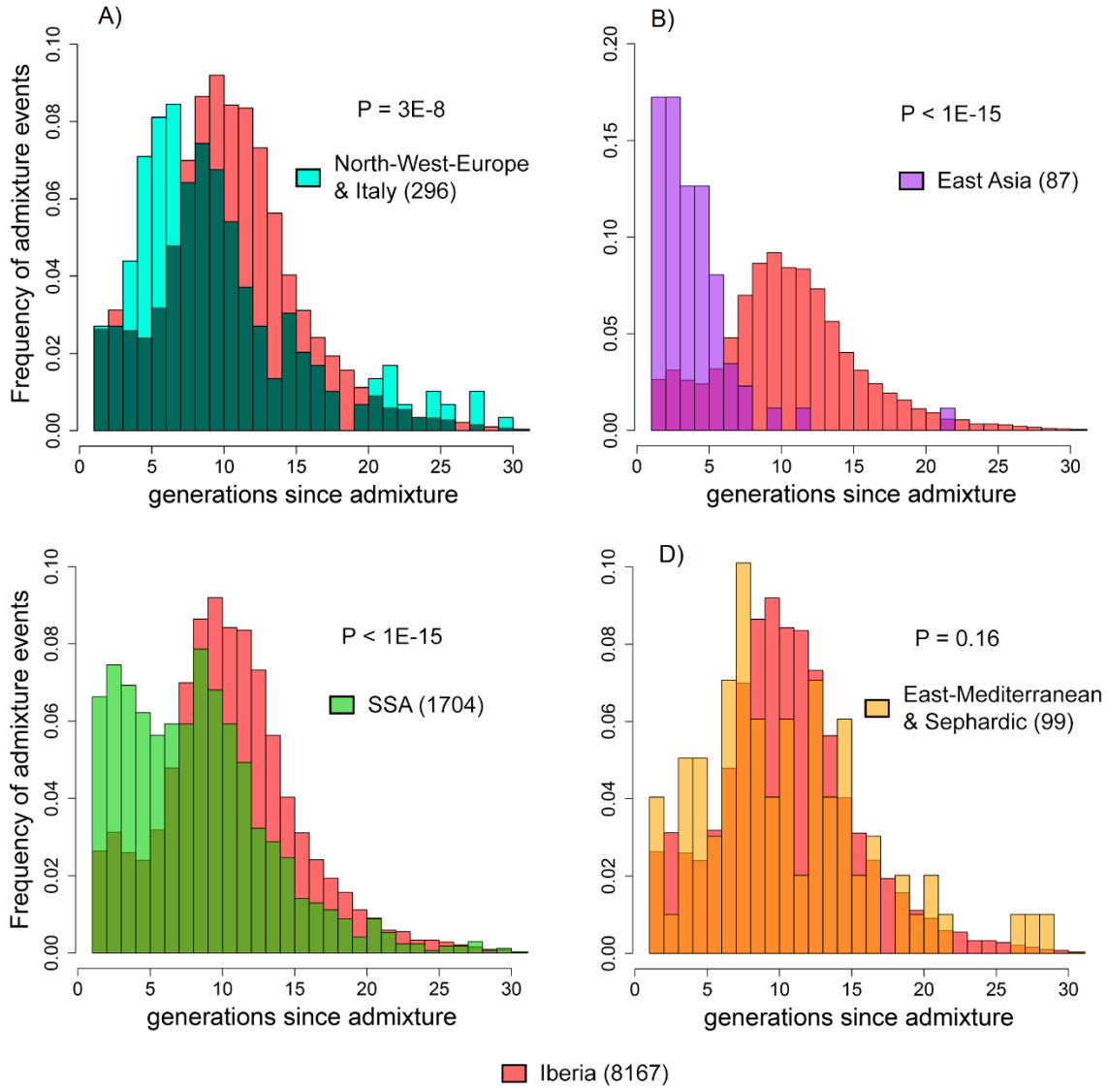

Figure 5.18. Frequency distributions of admixture events in the total CANDELA sample involving an Iberian-like source (red), contrasted with events involving sources related to (A) NorthWestEurope & Italy (B) East Asia, (C) Sub-Saharan African and (D) East Mediterranean & Sephardic.
The p-values were obtained using a Mann-Whitney U test (Section 5.2.5).

Table 5.2. Proportion of inferred admixture events with given GLOBETROTTER conclusion, for all events inferred to have at least one admixing source group best-matched by the given reference group

Source*	n	One-date	One-date, multiway	Multiple-dates, recent	Multiple-dates, older
Iberia	8167	0.4	0.09	0.24	0.26
NorthWest Europe & Italy	296	0.57	0.15	0.15	0.12
E.Mediterranean & Sephardic	99	0.41	0.04	0.25	0.29
Sub Saharan Africa	1704	0.02	0.28	0.52	0.18
East Asia	87	0.07	0.02	0.89	0.02
ALL SOURCES		3519	455+455**	2378	2378

*The sources have been defined as explained in section 5.2.5 to represent different historical/demographic processes. **The two events inferred in this scenario are simultaneous.

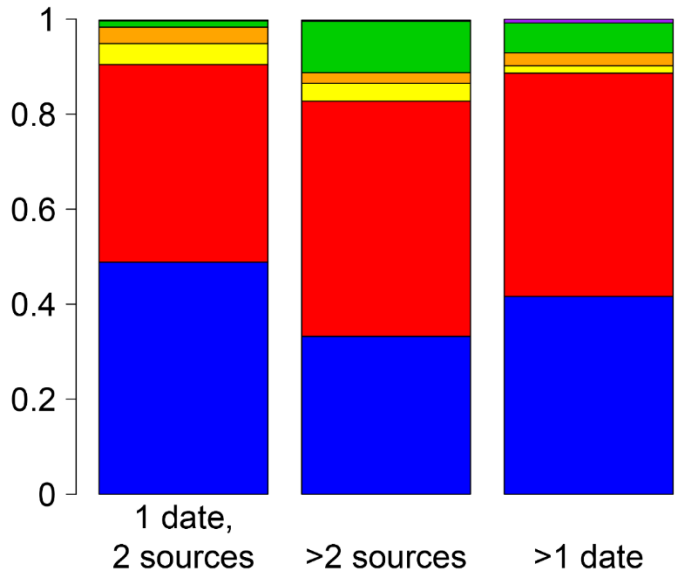

Figure 5.19. Percentage of SOURCEFIND inferred continental ancestry, per type of admixture event as inferred by GLOBETROTTER.
Colours: Yellow: *NorthWestEurope & Italy*, orange: *East Mediterranean & Sephardic*, green: *Sub Saharan Africa*, purple: East Asian, red: Iberian, and blue: Native American.

Interestingly, dates for admixture involving a Sephardic/East Mediterranean source were not significantly different from those involving Iberian sources (Figure 5.18D; Wilcoxon rank-sum test one-sided p-value>0.1), consistent with the scenario that a substantial fraction of the Sephardic/East Mediterranean ancestry detected in Latin Americans was introduced during the colonial period. In this respect, it is noteworthy that admixture dates estimated for seven individuals with only Native American and Sephardic/East Mediterranean ancestry, had a median of 9 generations ago (range 4 to 13), consistent with the view that a proportion of

the Iberian colonial immigrants were of mostly non-European ancestry, possibly recent Christian converts. These results must be interpreted cautiously, as the number of people with a considerably high Sephardic / East Mediterranean-like contributions is very low.

Table 5.3. Results for linear regression of total % Native American ancestry on inferred admixture date, for individuals inferred to have a single date of admixture between two sources best represented by a European and Native American surrogates.
To test robustness, we restricted the regression to individuals whose inferred proportions *p* of Native and European ancestry *each* met the given criterion. **(A)** All individuals. **(B)** Individuals inferred to have a single date of admixture between 5-17 generations ago. Analyses performed jointly with G. Hellenthal.

(A)

Analysis	**Nind**	**Beta**	**se(Beta)**	**t-stat**	**p-value**
all	3,340	-1.41	0.14	-10.4	< 1e-15
0.05 < *p* < 0.95	3,244	-1.56	0.13	-11.7	< 1e-15
0.1 < *p* < 0.9	3,049	-1.52	0.13	-12.1	< 1e-15
0.2 < *p* < 0.8	2,534	-1.21	0.11	-11.2	< 1e-15
Simulations (all)	1,297	-0.11	0.17	-1.04	0.30
Simulations, multiple events (all)	923	-0.11	0.03	-3.73	0.0002

(B)

Analysis	**Nind**	**Beta**	**se(Beta)**	**t-stat**	**p-value**
all	3,274	-1.45	0.15	-9.7	< 1e-15
0.05 < *p* < 0.95	3,189	-1.62	0.14	-11.2	< 1e-15
0.1 < *p* < 0.9	3,000	-1.60	0.14	-11.8	< 1e-15
0.2 < *p* < 0.8	2,495	-1.29	0.12	-11.2	< 1e-15
Simulations (all)	1,083	-0.27	0.25	-1.1	0.28
Simulations, multiple events (all)	832	-0.10	0.04	-2.63	0.009

Among instances in which GLOBETROTTER inferred a single date of admixture involving Native Americans and Europeans, we observe a significant increase (p-value<10^{-15}) in average Native American ancestry as time since admixture decreases (with an average increase of ~1.2-1.6% per generation. Figure 5.20, Table 5.3). The process underlying this pattern is unclear, but simulated scenarios suggest that this trend is real (Chapter 4, Sections 4.3.1 and 4.3.2). It is consistent

with continuing admixture between admixed Latin Americans and unadmixed Natives until as recently as ~200 years ago, possibly as a result of the decline in Iberian immigration after the mid-17th century (Sánchez-Albornoz 1994), concomitant with the demographic recovery of neighbouring Native American populations.

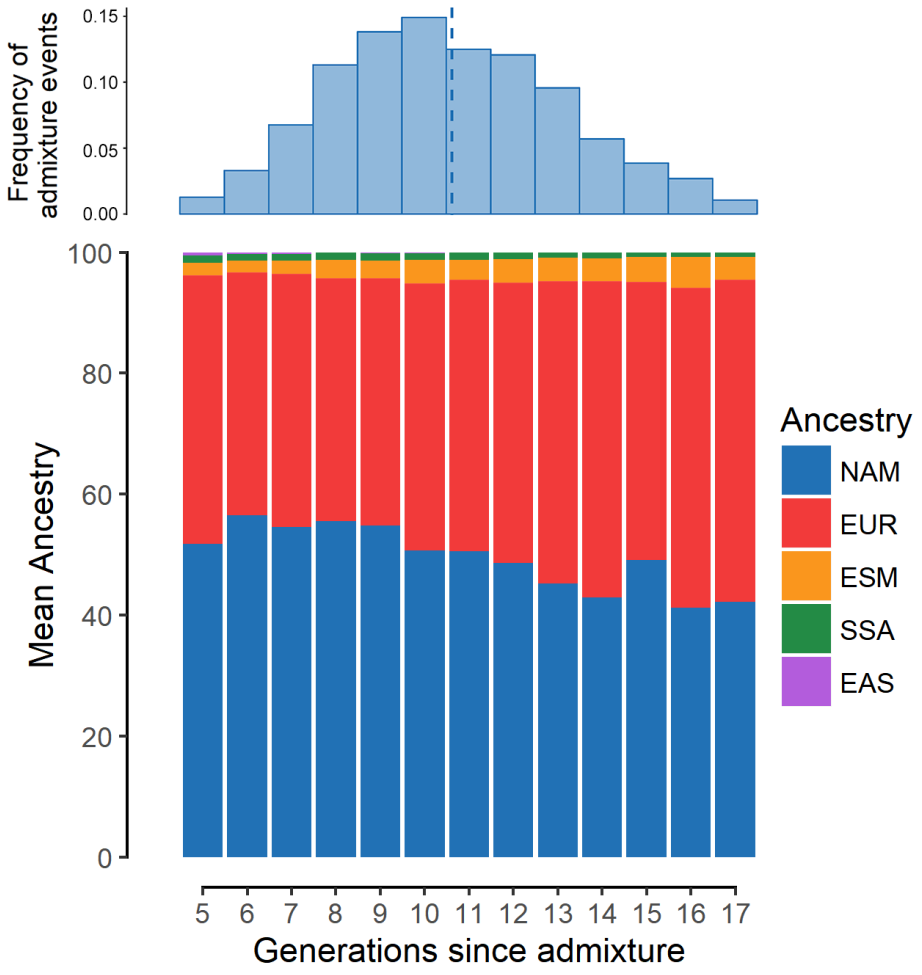

Figure 5.20. Times since admixture estimated with GLOBETROTTER for individuals in which a single time of Native American – European admixture was inferred.
Top: Frequency distribution of admixture times. The dashed line indicates the mean. Bottom: mean continental ancestry (%) as a function of time since admixture. Only time bins including >20 individuals are shown (NAM= Native American, EUR = European, ESM = East/South Mediterranean, SSA= Sub-Saharan African, EAS = East Asian).

5.4 Discussion and limitations

This chapter describes the fine-scale genetic structure of several Latin American populations by means of the inference of individual sub-continental ancestry proportions, achieving a high level of resolution, compared both to allele-frequency-

based approaches and to previous reports exploring fine-grained genetic structure in the region (Browning et al. 2016; Conley et al. 2017; Homburger et al. 2015; Montinaro et al. 2015; Moreno-Estrada et al. 2013; Wang et al. 2008).

I did a detailed description of the limitations of allele-frequency approaches in CANDELA (Section 5.2.1). Even though PCA and ADMIXTURE differentiate among major continental groups, they have difficulty distinguish more subtle population structure in the surrogate clusters (Chapter 3), limiting the inference and interpretation of population relationships in admixed individuals at lower PCs and the estimation of ancestry at the sub-continental level. The most common factors causing this limitation are genetic drift and low sample sizes in the groups included in a given analysis (Lawson et al. 2017; McVean 2009). The case of genetic drift is clearly exemplified in this dataset in the component arising at K=6 with unsupervised ADMIXTURE (Figure 5.1; Chapter 3, Section 3.8), where a considerable number of admixed samples from a well-known genetic isolate in North-west Colombia form into their own cluster, despite the fact they are recently admixed. Additionally, although continental ancestry estimates are consistent between allele-frequency and haplotype-based approaches, the ancestry estimations at low amounts need to be analysed and described carefully as described in Section 5.3.2.1.

While interpreting the haplotype-based inference, it is important to consider that contemporary samples may not be the best surrogates for ancestral source populations, as discussed in Chapters 3 and 4, possibly leading to uncertainty in the ancestry assignments and complicating both the inference and the interpretation of results. For instance, for Mexico, Peru and Chile there are likely good proxies for Native American ancestors in the dataset, and their results match demographic/historical scenarios (like the Inca expansion). In contrast, other populations with small Native American contributions (i.e. the Brazilian sample), and/or with limitations to find good surrogates due to their demographic history involving isolation of both admixed and indigenous populations (i.e. Colombia) are harder to characterize.

Native American ancestry represents perhaps the best scenario to apply our approach for several reasons. First of all, the high levels of genetic drift between Native American populations make them easier to distinguish (although this could

eventually interfere with the inference of haplotype similarity patterns due to high amount of haplotype-sharing within the same population). Secondly, because the admixture is recent, there is perhaps a better chance that good surrogates exist today, which can be potentially sampled. The results presented in this chapter highlight the high structure of Native American ancestry in Latin Americans both between and within countries.

European and East / South Mediterranean contributions are harder to disentangle due to the complex demographic dynamics of the populations located in these regions. From allele-based approaches it is well known that European populations show low differentiation (Novembre et al. 2008). Supported by the analysis on simulated data, I demonstrated that in several scenarios I have enough power to distinguish less differentiated groups, such as different Iberian populations. Independent of the accuracy of the surrogates or the post-Columbian divergence of the populations involved in the analyses (as suggested by AS-PCA in Moreno-Estrada et al. (2013)), the results match the main contributions suggested by historical records. Furthermore, new contributions from Sephardic populations are an invaluable asset on the resolution of contradictory historical records. The timings and sources since admixture clearly show that all these populations have been involved in the admixture processes since colonial times.

In this chapter, I also describe the first estimation of the timings and sources of admixture using single individuals. I consider it an important improvement, given the fact that genetic ancestry proportions in Latin America are highly variable among individuals. The estimations described in this chapter generally match with historical accounts and are of great value to corroborate histories of migration and admixture. For instance, in the “one-date admixture event” scenario, I detect a pattern where Native American ancestry tends to increase as the events are more recent, probably reflecting several demographic events like urbanization and changes on the amount of European migrants to these countries (Sánchez-Albornoz 1994). Furthermore, an increase of Native American ancestry across generations could indicate a steady recovery of population sizes in Native American groups, as well as the increased migration of “mestizos” (admixed individuals) carrying higher amounts of these ancestry into the more European-like urban centres.

Other ancestries not related to the colonization process also contribute to the new novel insights of this thesis. For instance, East Asian flow has been reported since colonial times as a fundamental part of the transpacific Spanish routes (Sánchez-Albornoz 1994). However, from historical records it seems more likely that these signals are due to the massive migration of East Asian workers during the 19th century (Chapter 1) and our results support this conclusion. The time of admixture involving East Asian sources on average is less than 200 years, consistent with the beginning of the Republican Period and the abolition of the slave trade.

It is also essential to consider that ours is a convenient sample with recruitment limited to specific areas (Chapter 1, Section 1.6), and as such the results presented here cannot be taken as an exact and comprehensive representation of an entire country, especially in the case of recently admixed populations, where the individual variation in ancestry proportions is highly variable and where every region within the country may have had a totally different colonial history.

5.5 Summary

In this chapter I provide a detailed picture of ancestry in over 6,500 individuals from five Latin American countries. I observe that Native American population structure is extensively reflected in the ancestry of Latin Americans at a within-country level, with times and sources of Non-Native ancestry matching documented regional migratory flows to the New World. I also detect significant and widespread East/South Mediterranean (particularly Sephardic) ancestry across the region, possibly in connection with the persecution of non-Christians in Spain during the colonial period.

Overall, this chapter enriches historical analyses of the Americas and contributes to a deeper understanding of the heterogeneity of the sources involved in the complex and continuous admixture processes in Latin America. In the next chapter I make use of this information to assess its impact on physical appearance, with a range of phenotypes measured in the CANDELA sample.

6 Impact of sub-continental ancestry on physical appearance in Latin Americans

6.1 Overview

In the previous chapter I provided a comprehensive description of sub-continental ancestry in Latin American populations spanning five countries. For that purpose, I characterized and quantified the diversity of the sources involved in the make-up of current-day Latin American populations, showing how these patterns match the demographic history of the region.

The genetic heterogeneity generated by these extensive admixture processes also impacts the phenotypic diversity, evidenced by the great variation in physical appearance traits observed in Latin American populations and their association with continental ancestry variation (Adhikari et al. 2016c; Ruiz-Linares et al. 2014). Understanding the extent of this impact at finer levels of regional genetic variation could be of great usefulness for exploring the genetic architecture of complex traits and for improving current ways of accounting for genetic variation in GWASs.

In this chapter I evaluate the impact of sub-continental ancestry, as an approximation to fine-scale regional genetic variation, on a range of physical appearance features measured on the CANDELA sample, including aspects of anthropometry, face and ear morphology, facial and scalp hair and pigmentation. I find significant correlations between variation in European ancestry sub-components and pigmentation traits, as well as between variation in Native American sub-components and facial features. It evidences the impact of regional genetic variation on human phenotypic diversity and highlights the importance of taking fine-grained genetic variation into account in human genetic studies.

6.2 Methods

In this section I describe the physical appearance traits included in the analyses and all the considerations for evaluating the association between these traits and sub-continental ancestry estimates using linear regression.

6.2.1 Phenotypes description

I used data for 28 physical appearance traits that were collected, gathered and processed by several researchers/students involved in the CANDELA consortium, which have been previously published in different studies (Adhikari et al. 2016a; Adhikari et al. 2016b; Adhikari et al. 2015; Cerqueira et al. 2014; Quinto-Sanchez et al. 2015; Ruiz-Linares et al. 2014), some of which I actively collaborated on my PhD. These traits were recorded by physical examination of the volunteers and/or by examining facial photographs. In the case of most of the traits recorded from facial photographs, >10% of the images were scored twice by two observers, independently, two weeks apart, in order to evaluate the observer reliability by calculating intra-class correlation coefficients (ICC) (Adhikari et al. 2016b). These traits have been described in detail previously in the above papers. Here I briefly describe how the phenotypes were measured/scored:

- (1) Height. Quantitative measurement (in cm).

Head and hair:

- (2) Monobrow. 1: low, 2: medium or 3: high (thinner to thicker).
- (3) Eyebrow density. 1: low, 2: medium or 3: high (thinner to thicker).
- (4) Beard density. Divided in shaven and unshaven men. 1: low, 2: medium or 3: high.
- (5) Scalp hair shape. 1: straight, 2: wavy, 3: curly or 4: frizzy.
- (6) Scalp hair greying. 1: no greying, 2: predominant no greying, 3: 50% greying, 4: predominant greying or 5: totally white hair.
- (7) Balding. 1: low, 2: medium or 3: high. Measured in men and women.

Pigmentation traits:

- (8) Natural hair colour. 1: blond, 2: dark blond/light brown or 3: brown/black.

- (9) Skin colour (Melanin index). Quantitative measurement using DermaSpectrometer DSMEII reflectometer (Cortex Technology, Hadsund, Denmark). The value used for each individual corresponds to the mean index for both inner arms.
- (10) Eye colour. 1: blue/grey, 2: Honey, 3: Green, 4: light brown, 5: dark brown/black.

Categorical face traits:

- (11) Brow ridge protrusion. The presence and degree of a ridge in lateral view. 0: none, 1: slightly pronounced or 2: strongly pronounced.
- (12) Eye fold. Skin fold of the upper eyelid, covering the inner corner (medial canthus) of the eye. 0: no fold, 1: partial, 2: completely.
- (13) Chin shape. Chin contour in frontal view. 0: pointed, 1: rounded or 2: square.

Quantitative face traits:

These were defined based on landmarks placed on facial photographs (taken at three different angles) as detailed in figure 6.1:

- (14) Forehead profile. Slope of line joining 35-1.
- (15) Nasion position. Distance from landmark 18 to the mid-point of a line joining landmarks 8 and 16.
- (16) Nose bridge breadth. Distance between landmarks 37 and 38.
- (17) Nose wing breadth. Distance between landmarks 20 and 22.
- (18) Columella Inclination. Angle between landmarks 19-21-23.
- (19) Nose protrusion. Distance of landmark 19 to a line joining landmarks 18 and 21.
- (20) Nose tip angle. Angle between landmarks 18-19-21.
- (21) Chin protrusion. Distance of point 30 from line joining 35-36.
- (22) Facial flatness. Distance 30-32/ distance 32-18.

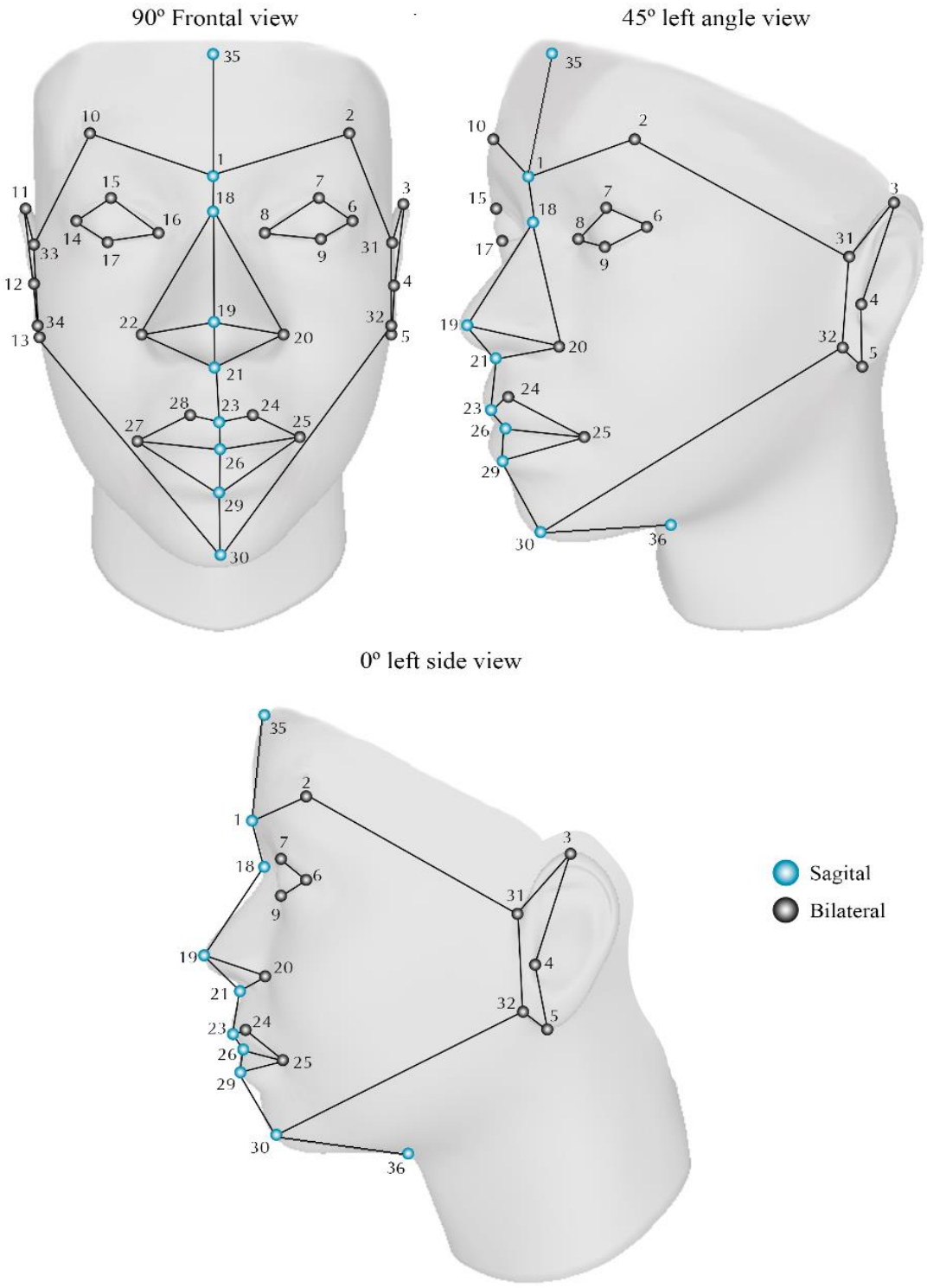

Figure 6.1. Landmarks placed on facial photographs obtained for CANDELA. Adapted from Quinto-Sanchez et al. (2015). Generated by M. Fuentes-Guajardo.

Ear traits:

Location of these features is provided in figure 6.2. All traits were scored on a 3-point scale (low, medium, high).

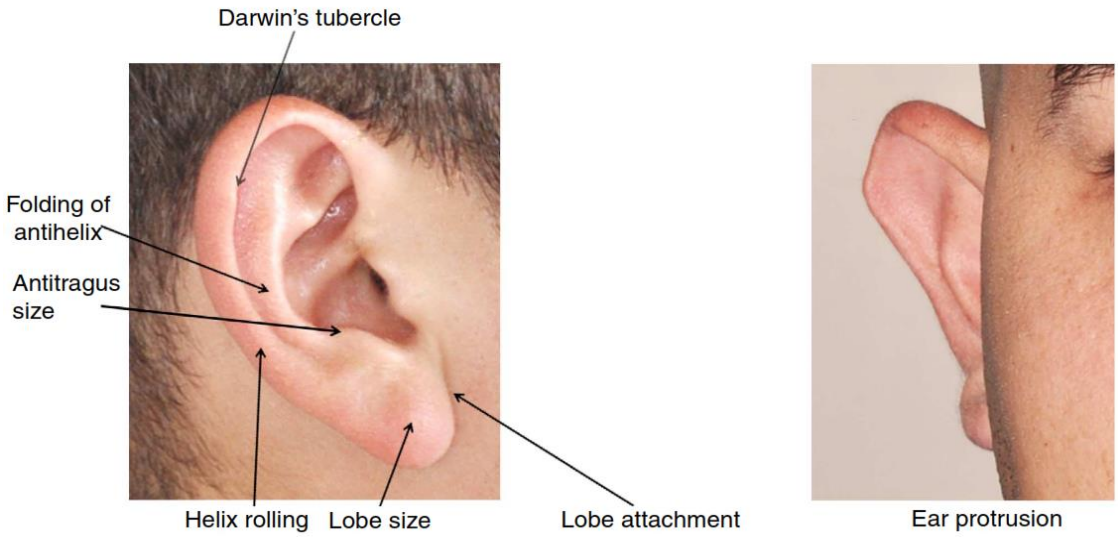

Figure 6.2. Location of ear traits characterized in the CANDELA dataset. Modified from Adhikari et al. (2015). Generated by M. Fuentes-Guajardo.

- (23) Ear protrusion. Degree of protrusion of the ear in relation to the frontal face view (less to more protruded).
- (24) Lobe attachment. Degree of attachment of the inferior part of the pinna to the anteroinferior part of the face (no attachment to complete attachment).
- (25) Lobe size. Small to bigger size.
- (26) Helix rolling. The outer rim of the ear that extends from the superior insertion of the ear on the scalp (root) to the termination of the cartilage at the earlobe (less to more pronounced helix rolling).
- (27) Fold of antihelix. Less to more pronounced fold of antihelix.
- (28) Antitragus size. Small to bigger size. The anterosuperior cartilaginous protrusion lying between the incisura and the origin of the antihelix. The anterosuperior margin of the antitragus forms the posterior wall of the incisura.

6.2.2 Analyses

I assessed the impact of sub-continental ancestry on the traits described in the previous section using linear regression. I paid special attention to the sub-continental ancestry estimations obtained with SOURCEFIND (Chapter 5) as I demonstrated that they are robust estimators, in terms of how interpretations vary when

considering results at increasingly finer scales, compared to allele-frequency-based approaches (Chapter 4). However, I also used the results from the latter approaches and performed additional regressions to compare the results when the sub-continental estimations between methods were somehow equivalent. In the regression models I propose, a series of covariates were included in order to account for any confounding factors. All the analyses were made with the supervision and advice of K. Adhikari.

6.2.2.1 Contrasts of sub-continental ancestry estimates

Considering that ancestry sub-components are (negatively) correlated with other major continental ancestries (e.g. the proportion of the most prevalent Native American sub-component in Peru inversely correlates with broad European ancestry), they should not be used jointly in the linear model in order to avoid confounding. For this reason, I constructed contrasts of sub-continental ancestry components by taking the difference of a given pair of sub-continental components and only retained for the analysis the contrasts that matched the following criteria:

(i) Each sub-continental ancestry component tested should have >10% frequency in at least one country.
(ii) Each pair of contrasted sub-continental ancestry components should add up to at least half of the total continental ancestry for the respective component in a country.
(iii) The contrasted sub-continental ancestry components should show a relatively high genetic differentiation according to the clustering analyses performed in Chapter 3.

In order to reduce colinearity effects and to maximize statistical power, the analyses are focused on contrasts between the most common, highly differentiated sub-continental ancestry sub-components. In the case of European ancestry, only Brazil shows high frequency of more than one European sub-component (Chapter 5, Section 5.3.2.3), which is the contrast of *NorthWestEurope1* against *Portugal/WestSpain.*

For Native American ancestry, to enable a contrast, some closely related sub-components estimated by SOURCEFIND (*Quechua1, Quechua2, Colla* and *Aymara*) were merged into a group called *CentralAndes*. This group was used to create the contrast against *Mapuche*, which is relatively differentiated both genetically and geographically from the former. The merge was necessary as the non-Mapuche ancestry in Chile (the country with the highest amount of *Mapuche* and a decent amount of *CentralAndes*) only add to >10% if all Quechua and Aymara related contributions were merged. Similar components were defined by Principal Component (PC) 7 and by ADMIXTURE at K=7, allowing us to test for consistency between estimates from different methods.

6.2.2.2 Regression models and additional covariates

The basic regression models tested were:

Phenotype ~ Age + Sex (+BMI) + SES (+Country)* + Total Sub-Saharan African ancestry + Total European ancestry + Native component contrast,*

Or,

Phenotype ~ Age + Sex (+BMI) + SES + Total Sub-Saharan African ancestry + Total Native American ancestry + European component contrast.*

*These variables are not always used in the regression, as I explain below.

Age, sex and socioeconomic status (SES) were included as covariates for all analyses.

SES is represented as a wealth index that measures living standards (Filmer and Pritchett 2001; Gwatkin et al. 2007), estimated from a list of material items provided by the participants (including ownership of property and household items as well as the availability of domestic service) using the first PC of a polychoric PCA (Kolenikov and Angeles 2009) as described in Ruiz-Linares et al. (2014). Briefly, a standard PCA is performed using polychoric correlation (Olsson 1979), which is designed for comparing ordinal variables (in this case the list of material items). This wealth index was converted to deciles within each country in order to allow comparisons between them.

SES was included for two main reasons. First, it is known that SES is correlated with continental ancestry (low SES correlates with Native American ancestry) and it also affects physical traits such as height, providing clear evidences on the impact of wealth in physical development (Ruiz-Linares et al. 2014). This covariate will test whether ancestry has an effect on phenotype beyond what can be explained by SES. Second, because it adjusts for the convenience sampling of CANDELA within every country (see Chapter 1 for details).

When undertaking a multi-country analysis, to adjust for the combined convenience sampling (which was different in each country), SES was used as a dummy variable, and additionally, country was set as a dummy variable too. Additional to the biases on the sampling within every country, adding country as a covariate allows us to adjust for the fact that not only the sampling, but also the collection of the information were done independently in every country, and some of the variables, like SES, may not be entirely equivalent between countries.

Body Mass Index (BMI) was included for all the face morphology traits.

Continental ancestries not related to the contrast being analysed were included as covariates to account for any effect related to the total amounts of ancestry. To reduce additional variability from other continental ancestries, we exclude people with more than 10% Sub-Saharan African or East / South Mediterranean ancestry, and/or with >1% East Asian ancestry. This is because, the two main ancestries in our sample are European and Native American, and the contrasts proposed were only encompassing these two.

Bonferroni correction of the significance threshold, for all traits (28) an all contrasts (3), gives a final significance cut-off of $-\log P = 3.22$ equivalent to and alpha of 0.05 and 84 observations (0.05/84). Finally, In order to make the Betas comparable for display items, the Beta value for each trait was standardized to report results as a factor of the standard deviation (SD), i.e. by dividing them by the SD of the respective trait.

6.2.2.3 Differences in allele frequencies of GWAS hits between *Mapuche* and *CentralAndes*

In the GWAS we reported in Adhikari et al. (2016b), several loci were identified as being associated with facial features (Chapter 1, Section 1.4.1). We wanted to test whether allele frequencies in these loci differ between individuals with Mapuche and Central Andean ancestries, but given the fact that there are only five individuals in the *Mapuche* surrogate cluster, it was not possible to obtain reliable allele frequencies for this group. To overcome this limitation, we performed a local continental ancestry analysis in two subsets of phased CANDELA individuals with considerable amounts of *CentralAndes* or *Mapuche* ancestry, in order to extract the information for the Native American segments containing these loci in every individual and combine this information with the respective surrogate samples from each cluster.

This analysis was performed by K. Adhikari and J. Mendoza using RFMix (Maples et al. 2013), with three continental reference groups (107 IBS, 101 YRI and 125 Native American samples). RFMix assigns local continental ancestry to each allele of each CANDELA haplotype, providing both the continental ancestry and the inferred allele at that site. The software accounts for errors in genotyping, marginal amounts of admixture in the reference groups and phasing switch errors (Maples et al. 2013).

Using SOURCEFIND results, we selected the two subsets according to their subcontinental ancestry proportions. For each set, all individuals had >10% inferred ancestry from the Native group of interest, with <1% combined inferred ancestry from all other Native groups and <1% inferred East Asian ancestry. For all individuals in a group, for the index SNPs of all the six genomic regions identified in Adhikari et al. (2016c), all alleles that had local Native American ancestry were used to estimate the allele frequency.

As a sanity check, the allele frequencies obtained for *CentralAndes* with this analysis were compared to the frequencies obtained directly from 49 surrogate individuals in the *CentralAndes* cluster that have more than >99% Native American ancestry ($r^2 > 0.99$). Finally, a t-test was used to assess whether the allele frequencies were significantly different in *CentralAndes* vs. *Mapuche* individuals.

The FDR (false discovery rate) procedure was used to control the Type-I error rate at 0.05 level.

6.2.2.4 Comparisons

Several of the phenotypes recorded in the CANDELA sample and used in these analyses were taken from the Anthropological Atlas of male facial features (Ohlrogge 2008). A subsequent paper published by the same authors (Ritz-Timme et al. 2011) evaluated 300 people from each Germany and Italy to assess the frequencies for some of the categorical traits initially published in the Atlas. For the traits that overlap with our traits, we calculated chi-square p-values to assess the differentiation of the trait between *NorthWestEurope1* (Germany) and *Portugal/WestSpain*.

In our Brazilian dataset, samples with *Portugal/WestSpain* ancestry have predominantly Portuguese ancestry. Though Ritz-Timme et al. (2011) did not study Iberian samples, we take the Italian samples as a proxy for Iberian ancestry.

6.3 Results

Figure 6.3 summarizes the results of the linear regressions of sub-continental ancestry contrasts (obtained from SOURCEFIND results) against the 28 phenotypes described in Section 6.2.1. As explained in Section 6.2.2, only two strong contrasts based on SOURCEFIND results accomplished all the criteria we established.

6.3.1 A contrast of *CentralAndes* versus *Mapuche* ancestry is associated with facial morphology traits

The contrast *CentralAndes* versus *Mapuche* in the full CANDELA sample is significantly associated with variation in several facial features (Figure 6.3). To take one example, Figure 6.4 shows a scatterplot with the regression line for one of the associated traits, Nose Bridge Breadth.

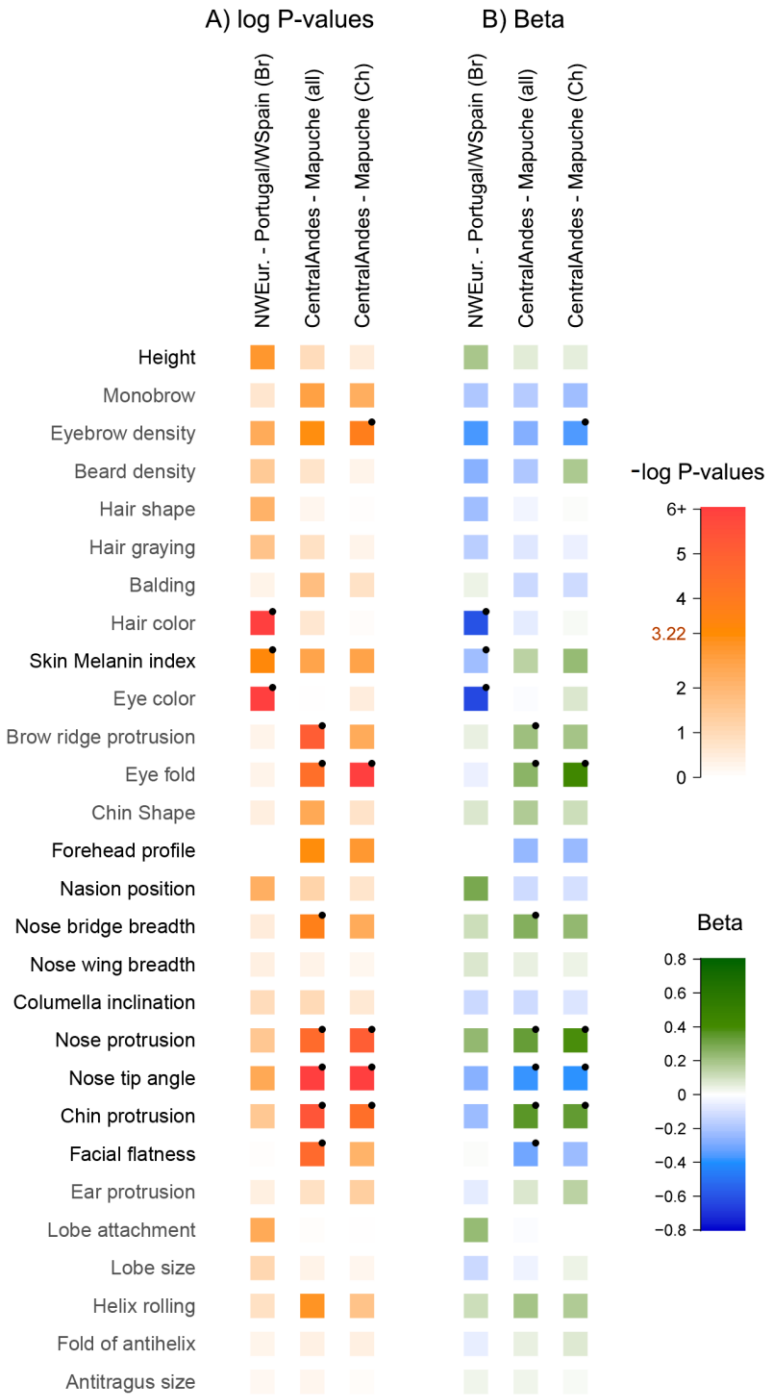

Figure 6.3. Sub-continental ancestry and physical appearance.
(**A**) -log P-values for a linear regression of variation in the 28 traits described in section 6.2.1 (with categorical traits listed in grey and quantitative traits in black) against the contrasts between two sub-continental ancestry components estimated by SOURCEFIND. The left column details results for the *Portugal/WestSpain* and *North-WestEurope1* contrast in the Brazilian sample (Br). The two right columns present the contrast between *CentralAndes* and *Mapuche* in the full CANDELA sample (all) or restricted to Chile (Ch). Bonferroni-corrected P-value significance threshold (alpha=0.05) is shown on the –log P-value scale as 3.22. (**B**) Regression coefficients (Betas), divided by the standard deviation (SD) for that trait, for the contrasts in (**A**) (hence in units of SD). In panels (**A**) and (**B**) colour intensity reflects variation in –log P-values or beta coefficients, as indicated on the scale. Bonferroni-corrected significant values are highlighted with a dot. Adapted from Chacón-Duque et al. (2018). Generated by J.C. Chacón-Duque and K. Adhikari.

Validation analyses using the same contrast but limited to Peru and Chile or only to Chile, an equivalent contrast generated from unsupervised ADMIXTURE results (K=7) and PC7 produced similar results, showing consistency for several phenotypes, with four traits significantly associated in all analyses: Eye Fold, Chin Protrusion, Nose Protrusion and Nose tip angle (Figure 6.5 and Table 6.1).

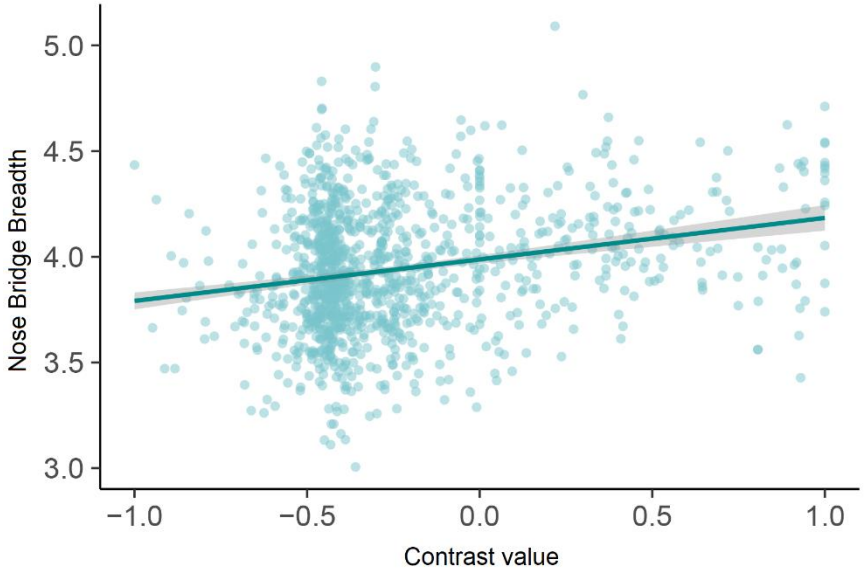

Figure 6.4. Scatterplot and regression line (with 95% confidence interval) for nose bridge breadth and the SOURCEFIND contrast between *CentralAndes* and *Mapuche* in Peru and Chile.
Adapted from Chacón-Duque et al. (2018). Generated by J.C. Chacón-Duque and K. Adhikari.

The associations with nose-related traits are perhaps the most interesting given previous evidence. In the analysis I present here, the Mapuche sub-component is associated with a less protruded nose (-log P-value=4.61) and equivalently with a broader nose tip angle (-log P-value=6.83), consistent with physical anthropology studies indicating that the Mapuche have a flatter, wider nose compared to Aymaras and Quechuas, the two main Central Andean groups (Bustamante et al. 2011b; Comas 1960; Davies 1932). Furthermore, this variation has also been documented in world-wide human populations (including some Andean and Southern Chilean populations) and an association between nose protrusion and dry and cold conditions has been found, interpreted as an evidence for climatic adaptation (Davies 1932; Hubbe et al. 2009; Leong and Eccles 2009).

A research that compared the divergence of quantitative nose shape traits and neutral molecular markers (using $Q_{ST} - F_{ST}$ comparisons, (Leinonen et al. 2013;

Wright 1951)) has also suggested that features related to nose width have been influenced by adaptation to cold/dry versus hot/humid environments, as they seem to have differentiated more than it would be expected under genetic drift compared to other nose shape traits (Zaidi et al. 2017).

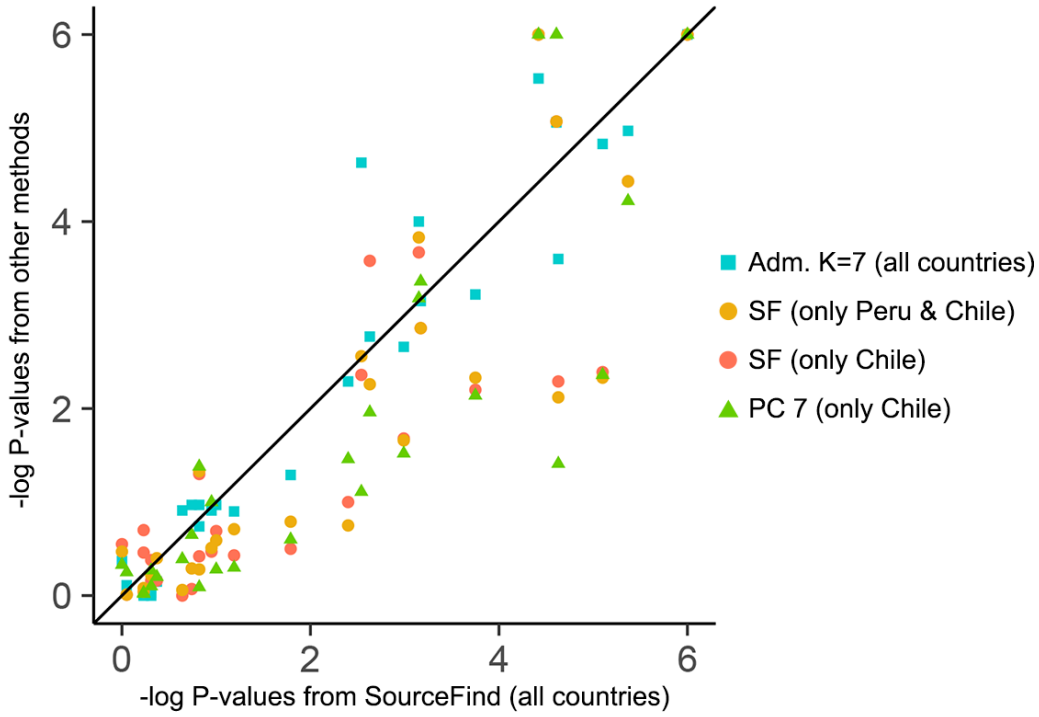

Figure 6.5. Scatterplot of –log P-values from additional phenotypic regression analyses involving *CentralAndes* versus *Mapuche* contrast.
The values used are presented in Table 6.1. The X-axis refers to P-values from the primary analysis using SOURCEFIND estimates (SF) and all the CANDELA data, as shown in the second column of Figure 6.3. The Y-axis refers to –log P-values from four other regression analyses using related ancestry components defined by ADMIXTURE (Adm) at K=7 in all the CANDELA data or using SOURCEFIND or PCA (PC7) ancestry components limited to the Peruvian and/or Chilean data (chapter 3, section 3.8). Sample sizes: all data N=5.794, Peruvian and Chileans N=2.594, Chileans N=1.542. Adapted from Chacón-Duque et al. (2018). Generated by J.C. Chacón-Duque and K. Adhikari.

The nasal cavity is an important regulator of inhaled air, temperature and humidity (Naftali et al. 2005). The nasal airways warm inspired air and saturate it with water vapour, in order to reach the right optimal temperature and moisture required in the respiratory tract (Negus 1954). Regarding the possible effect of environmental adaptation, it has been proposed that narrow respiratory cavities maximize the mucosal contact area in relation to the inhaled air volume, enhancing the airflow and facilitating the exchange of heat and moisture in cold or dry climates (Churchill et al. 2004). For instance, according to simulated data, narrower nasal

airways could be helpful in colder climates as the airflow increases and helps to warm the air quickly (Zhu et al. 2011), suggesting that populations with these characteristics, like the Central Andeans, may have been adapting to altitude.

Table 6.1. –log P-values from additional phenotypic regression analyses involving *CentralAndes* versus *Mapuche* contrast

Trait	ADM.K=7 (all)	SF (all)	SF (Pe-Ch)	SF (Ch)	PC7 (Ch)
Height	0.91	0.95	0.47	0.51	1
Monobrow	2.77	2.63	**3.58**	2.26	1.96
Eyebrow density	**4**	3.15	**3.67**	**3.83**	3.18
Beard density	0.97	0.74	0.07	0.29	0.65
Hair shape	0	0.23	0.7	0.04	0.02
Hair graying	0.97	0.82	0.42	0.28	0.09
Balding	1.29	1.79	0.5	0.79	0.6
Hair color	0.91	0.64	0	0.06	0.39
Skin Melanin index	**4.63**	2.54	2.36	2.56	1.11
Brow ridge protrusion	**4.83**	**5.1**	2.39	2.33	2.36
Eye fold	**5.53**	**4.42**	**6.97**	**10.21**	**11.6**
Chin Shape	2.29	2.4	1	0.75	1.46
Forehead profile	3.15	3.17	2.86	2.86	**3.36**
Nasion position	0.9	1.19	0.43	0.71	0.3
Nose bridge breadth	**3.22**	**3.75**	2.2	2.33	2.14
Nose wing breadth	0.03	0.31	0.38	0.22	0.1
Columella inclination	0.97	1	0.69	0.59	0.28
Nose protrusion	**5.06**	**4.61**	**5.07**	**5.07**	**6.33**
Nose tip angle	**7.84**	**6.83**	**6.01**	**6.06**	**6.96**
Chin protrusion	**4.97**	**5.37**	**4.43**	**4.43**	**4.22**
Facial flatness	**3.6**	**4.63**	2.29	2.12	1.41
Ear protrusion	0.74	0.82	1.3	1.33	1.38
Lobe attachment	0.11	0.05	0.01	0.01	0.25
Lobe size	0	0.31	0.16	0.23	0.27
Helix rolling	2.66	2.99	1.68	1.66	1.52
Fold of antihelix	0.15	0.37	0.16	0.4	0.2
Antitragus size	0.02	0.23	0.46	0.08	0.03

*ADM.: ADMIXTURE, SF: SOURCEFIND, Pe: Peru, Ch: Chile.

It is important to consider that the correlation between nose shape and climate is not always present (Leong and Eccles 2009) and that other causes for this variation must be taken into consideration. Another plausible explanation could be the effect of sexual selection (Darwin 1871), as proposed for other physical appearance traits like skin pigmentation (Aoki 2002) and height (Stulp et al. 2015). However the evidences for sexual selection of facial traits in humans are scarce, so far being mainly supported by studies on the effect of facial attractiveness in mate choice (Little et al. 2011).

6.3.2 Allele frequencies in loci associated with variation in facial traits are significantly differentiated between *CentralAndes* and *Mapuche*

In a recent GWAS we published using CANDELA data (Adhikari et al. 2016b), five of the six genes with alleles significantly associated impacted on nose shape (Chapter 1, Section 1.4.1). Here we compared the allele frequencies of the index SNPs at these loci for the two groups included in the contrast, finding that all the SNPs show significantly differentiated allele frequencies between *CentralAndes* and *Mapuche* (Table 6.2), consistent with the phenotypic effect of the sub-continental ancestry contrasts analysis.

Table 6.2. Allele frequencies in the Central Andes and the Mapuche at index SNPs associated with facial features in the CANDELA sample.
Adapted from Chacón-Duque et al. (2018), elaborated by K. Adhikari.

				Allele frequency (Allele count)		
Chr. Region	**SNP**	**Gene**	**Derived allele**	**Central-Andes**	**Mapuche**	**P-value**
2q12	rs3827760	EDAR	G	0.961 (879)	0.995 (595)	2.18×10^{-4}
2q35	rs2395845	PAX3	A	0.388 (896)	0.683 (635)	6.09×10^{-29}
4q31	rs12644248	DCHS2	G	0.512 (903)	0.725 (699)	3.59×10^{-17}
6p21	rs1285029	SUPT3H/ RUNX2	C	0.585 (880)	0.638 (566)	4.51×10^{-2}
7p13	rs17640804	GLI3	T	0.417 (892)	0.498 (614)	6.19×10^{-3}
20p11	rs927833	PAX1	C	0.700 (888)	0.503 (616)	7.41×10^{-14}

Furthermore, for each SNP, the allele with a higher frequency in *CentralAndes* compared to *Mapuche* had the same direction of effect (same signs of regression coefficient beta) for that allele in the GWAS as compared to the regression coefficient (beta, Figure 6.3B) between the *CentralAndes-Mapuche* contrast and the trait, for all traits that are associated at a genome-wide significant or suggestive significant level with the SNP.

6.3.3 A contrast of *NorthWestEurope1* versus *Portugual/WestSpain* components is associated with pigmentation in Brazil

Regression analysis evidenced a highly significant effect of this contrast on pigmentation traits (skin pigmentation (-log P-value=3.39); hair colour (-log P-value=7.48); and eye colour (-log P-Value=7.5); Figure 6.3). As an example, Figure 6.6 shows a scatterplot with the regression line for one of the associated traits, skin pigmentation.

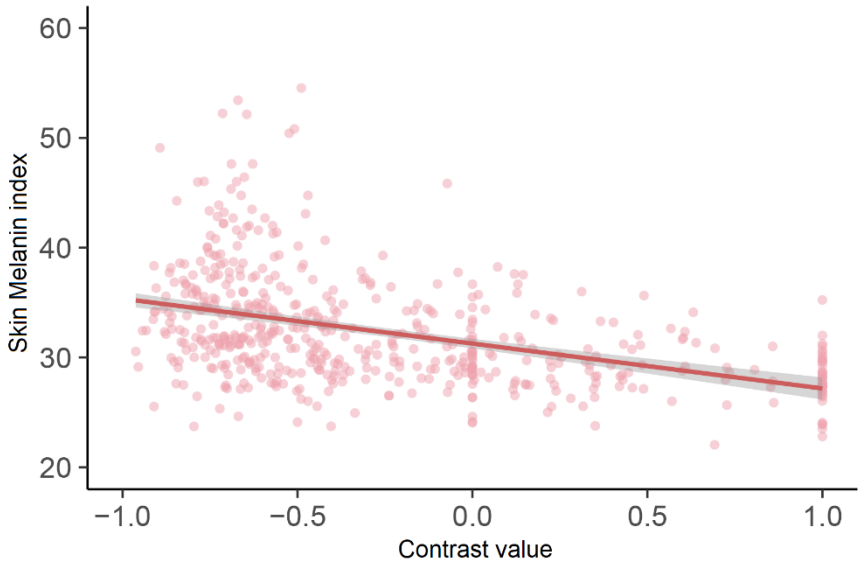

Figure 6.6. Scatterplot and regression line (with 95% confidence interval) for Skin Melanin index and the contrast between *NorthWestEurope* and *Portugal/WestSpain* in the Brazilian sample.
Adapted from Chacón-Duque et al. (2018). Generated by J.C. Chacón-Duque and K. Adhikari.

These results are consistent with the latitudinal gradient that has been reported in Europe, where Northern Europeans show lower pigmentation levels than Southern Europeans, with a striking difference in hair and eye colour (Frost 2014). A subset of anthropological facial features were compared between Northern and

Southern Europeans (Ritz-Timme et al. 2011) showed similar trends to this analysis (Figure 6.7, Table 6.3).

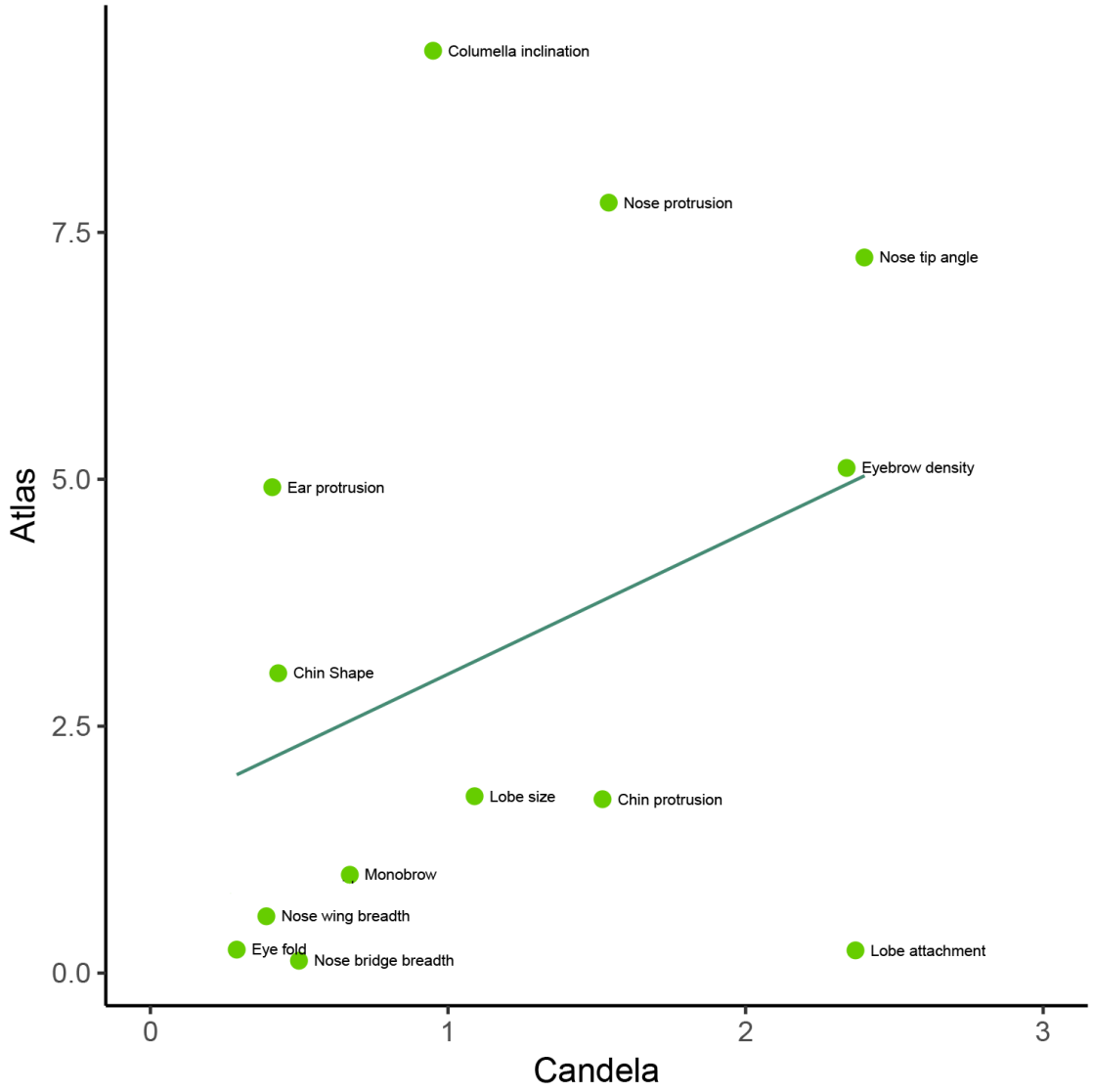

Figure 6.7. –log P-value comparison for North versus South Europe facial phenotypic differences.
Elaborated by K. Adhikari.

-log P-values from the comparison of trait frequencies from the anthropological atlas dataset were compared to the P-values for our *NorthWestEurope1 – Portugal/WestSpain* ancestry contrast in Brazil. While these are not equivalent, the relative ordering between the –log P-values can be compared, and we see considerable agreement between the two, with Spearman's rank correlation = 0.41. This illustrates how phenotype associations with European ancestry can be inferred in admixed individuals that carry additional sources of non-European (e.g. Native American, African) ancestry.

Table 6.3. –log P-values from the two studies, the Anthropological Atlas of male facial features and CANDELA

Trait	Atlas	CANDELA
Monobrow	1.00	0.67
Eyebrow density	5.12	2.34
Eye fold	0.24	0.29
Chin Shape	3.04	0.43
Forehead profile	1.87	NA
Nose bridge breadth	0.13	0.50
Nose wing breadth	0.58	0.39
Columella inclination	9.34	0.95
Nose protrusion	7.80	1.54
Nose tip angle	7.25	2.40
Chin protrusion	1.76	1.52
Ear protrusion	4.92	0.41
Lobe attachment	0.23	2.37
Lobe size	1.79	1.09

6.4 Discussion and limitations

The novelty of the analyses presented in this chapter lies on demonstrating that fine-grained regional genetic variation impacts traits of adaptive significance. Local genetic adaptation plays a major role in evolution and physical appearance has been the model of choice of anthropologists to characterize human origins, migration processes and evolution.

This crucial breakthrough comes, to a certain extent, because of the robustness achieved by SOURCEFIND to quantify sub-continental ancestry accurately, providing an opportunity to assess and interpret the effect of this subtle variation on different phenotypes. I hope this opens new opportunities to try and understand the impact of regional genetic variation on other complex phenotypes, including disease, and encourages the inclusion of fine-scale genetic structure into other research questions, such as studies of genetic association or natural selection.

The biggest limitations of this investigation were the limited availability of surrogates representing the original sources of admixture and the lack of a more geographically widespread sampling. The first limitation has been a recurrent obstacle during this project, as finding the best surrogate for a population is difficult, not only due to the temporal factor but also to the huge effort required for a random and even sampling. With more and better reference populations, it will be possible to obtain more precise estimations of the subcontinental ancestry components. Practical solutions to this limitation could be provided by approaches that do not require reference populations and also increase the resolution of population structure, such as AS-PCA or DAPC. The former has been successfully applied for associations with a phenotype, pulmonary capacity, in Mexican populations (see Chapter 1 - Section 1.4.1 for details).

The other limitation can only be surpassed by continuing the intensive sampling efforts of initiatives like CANDELA, which has taken years to collect a sample that - though one of the biggest representing Latin America - still lacks comprehensive geographical coverage. For instance, CANDELA's Chilean sample, which is probably the best currently available in terms of coverage, as it covers vast territories from the south cone to the central Andes, showed the strongest associations between Native American sub-continental ancestry variation and physical appearance diversity, probably indicating the power conferred by a geographically extensive sampling.

6.5 Summary

In this chapter I explore, for the first time, the relationship between sub-continental ancestry and physical appearance in Latin America. I show that variation in Native sub-continental ancestry in the Andean region significantly impacts on facial features, particularly nose morphology, setting the stage for further analyses on how variation in facial features could reflect environmental adaptation. Secondly, I also show that variation in Northern versus Southern European ancestry significantly impacts on pigmentation phenotypes in Brazilians, demonstrating how sub-continental European genetic information can be extracted in admixed individuals and tested for phenotype associations. Overall, these results highlight the impact of regional genetic variation on human phenotypic diversity.

7 Conclusions and perspectives

7.1 Conclusions

In this thesis I have provided a comprehensive reconstruction of the demographic history of Latin American populations using state-of-the-art approaches in population genetics. I paid a lot of attention to the process of constructing the reference panels and to the assessment of the performance of the methods implemented, aiming to set a strong base for the inference and the accurate interpretation of fine-scale genetic structure and sub-continental ancestry.

In Chapter 3 I demonstrated how haplotype-based methods provide a higher level of resolution for detecting fine-scale population structure compared to frequency-allele-based ones and I established a robust set of surrogates to infer the contribution of specific population groups to the genetic make-up of Latin Americans. The results of these analyses are supported by historical, geographic, linguistic, and genetic evidences.

The analyses presented in Chapter 4 confirm the accuracy of the new haplotype-based approaches we are using to estimate the sub-continental ancestry, and admixture time and sources, in a setting appropriate for Latin American analyses. By providing the first formal assessments of accuracy of these methods, I show that our methods accurately (*i*) identify sources and proportions of sub-continental ancestry and (*ii*) infer dates of admixture when analysing single individuals simulated to mimic genetic features of Latin Americans.

After the work from Chapters 3 and 4, I was able to reconstruct the demographic history of Latin American populations (Chapter 5) with a higher level of resolution than previous studies, highlighting new findings that corroborate historical accounts. This results enrich historical analyses of the Americas and contributes to

a deeper understanding of the heterogeneity of the sources involved in the complex and continuous admixture processes in Latin America.

Finally, I confirm the importance of understanding the fine-grained regional genetic variation in Latin America by establishing clear associations between sub-continental variation in ancestry and phenotypic diversity. More broadly, I demonstrate that physical appearance serves as a model system in which to examine the effect of local genetic variation on complex traits.

All these results demonstrate the importance of studying deeply the central role that sub-continental genetic variation has on the genetic architecture of human phenotypes. This is essential to consider, given the fact that up to this date, most of the surveys of genetic variation have been strongly biased towards European-derived populations.

7.2 Future directions

In the future, I would like to further understand how the complex demographic histories of Latin Americans have shaped their genomes and how the genetic architecture - shaped by endless migrations and deep bottlenecks - has influenced complex traits. With all this knowledge I would like to develop analytical strategies to better understand the evolution of these populations, seeking for biomedical and forensic applications.

It is clear that we need to do more to study populations under-represented in the current surveys of genetic diversity. I think we need to make bigger efforts to collect and analyse more samples. We also need to collect more ancient DNA samples, which could potentially represent better the ancestors of Latin Americans.

One field to explore is the possibility to corroborate poorly documented events further. That is the case of the Converso migration to Latin America. A better and bigger sampling of European and East/South Mediterranean populations will be necessary to confirm whether high numbers of Conversos migrated to America, or if the Spanish that migrated were highly admixed with Semitic peoples.

Given the magnitude of the population collapse suffered by Native populations in the last 500 years, the exploration of the Native American sub-continental ancestry in Latin America will allow the reconstruction of genetic profiles of the Native American peoples that contributed to the make-up of current mestizo populations.

Finally, I consider that considerable effort must go into public engagement and, more importantly, this interaction needs to be contextualized given the amount of cultural diversity of Latin America. This will require collaborative work with social scientists, policy makers and other people participating and impacting on decision making.

Bibliography

1000 Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA et al. . 2015. A global reference for human genetic variation. Nature 526(7571):68-74.

Abe-Sandes K, Silva WA, Jr., and Zago MA. 2004. Heterogeneity of the Y chromosome in Afro-Brazilian populations. Human biology 76(1):77-86.

Adhikari K, Chacon-Duque JC, Mendoza-Revilla J, Fuentes-Guajardo M, and Ruiz-Linares A. 2017. The Genetic Diversity of the Americas. Annual review of genomics and human genetics 18:277-296.

Adhikari K, Fontanil T, Cal S, Mendoza-Revilla J, Fuentes-Guajardo M, Chacon-Duque JC, Al-Saadi F, Johansson JA, Quinto-Sanchez M, Acuna-Alonzo V et al. . 2016a. A genome-wide association scan in admixed Latin Americans identifies loci influencing facial and scalp hair features. Nature communications 7:10815.

Adhikari K, Fuentes-Guajardo M, Quinto-Sanchez M, Mendoza-Revilla J, Chacon-Duque JC, Acuna-Alonzo V, Jaramillo C, Arias W, Lozano RB, Perez GM et al. . 2016b. A genome-wide association scan implicates DCHS2, RUNX2, GLI3, PAX1 and EDAR in human facial variation. Nature communications 7:11616.

Adhikari K, Mendoza-Revilla J, Chacon-Duque JC, Fuentes-Guajardo M, and Ruiz-Linares A. 2016c. Admixture in Latin America. Curr Opin Genet Dev 41:106-114.

Adhikari K, Mendoza-Revilla J, Sohail A, Fuentes-Guajardo M, Lampert J, Chacon-Duque JC, Hurtado M, villegas V, Granja V, Acuna-Alonzo V et al. . Submitted. A genome-wide association scan in latin americans underlines the convergent evolution of lighter skin pigmentation in eurasia. Submitted for publication.

Adhikari K, Reales G, Smith AJ, Konka E, Palmen J, Quinto-Sanchez M, Acuna-Alonzo V, Jaramillo C, Arias W, Fuentes M et al. . 2015. A genome-wide association study identifies multiple loci for variation in human ear morphology. Nature communications 6:7500.

Alarcon-Riquelme ME, Ziegler JT, Molineros J, Howard TD, Moreno-Estrada A, Sanchez-Rodriguez E, Ainsworth HC, Ortiz-Tello P, Comeau ME, Rasmussen A et al. . 2016. Genome-Wide Association Study in an Amerindian Ancestry Population Reveals Novel Systemic Lupus Erythematosus Risk Loci and the Role of European Admixture. Arthritis & rheumatology (Hoboken, NJ) 68(4):932-943.

Alexander DH, and Lange K. 2011. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC bioinformatics 12(1):246.

Alexander DH, Novembre J, and Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome research 19(9):1655-1664.

Alves-Silva J, da Silva Santos M, Guimaraes PE, Ferreira AC, Bandelt HJ, Pena SD, and Prado VF. 2000. The ancestry of Brazilian mtDNA lineages. American journal of human genetics 67(2):444-461.

Amirikia KC, Mills P, Bush J, and Newman LA. 2011. Higher population-based incidence rates of triple-negative breast cancer among young African-American women : Implications for breast cancer screening recommendations. Cancer 117(12):2747-2753.

Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, and Zondervan KT. 2010. Data quality control in genetic case-control association studies. Nature Protocols 5:1564.

Aoki K. 2002. Sexual selection as a cause of human skin colour variation: Darwin's hypothesis revisited. Annals of human biology 29(6):589-608.

Barbujani G, Magagni A, Minch E, and Cavalli-Sforza LL. 1997. An apportionment of human DNA diversity. Proceedings of the National Academy of Sciences of the United States of America 94(9):4516-4519.

Bedoya G, Montoya P, Garcia J, Soto I, Bourgeois S, Carvajal L, Labuda D, Alvarez V, Ospina J, Hedrick PW et al. . 2006. Admixture dynamics in Hispanics: a shift in the nuclear genetic ancestry of a South American population isolate. Proceedings of the National Academy of Sciences of the United States of America 103(19):7234-7239.

Beleza S, Johnson NA, Candille SI, Absher DM, Coram MA, Lopes J, Campos J, Araújo II, Anderson TM, Vilhjálmsson BJ et al. . 2013. Genetic Architecture of Skin and Eye Color in an African-European Admixed Population. PLoS genetics 9(3):e1003372.

Bellwood P. 2004. First Farmers: The Origins of Agricultural Societies: Wiley.

Berg MA, Peoples R, Perez-Jurado L, Guevara-Aguirre J, Rosenbloom AL, Laron Z, Milner RD, and Francke U. 1994. Receptor mutations and haplotypes in growth hormone receptor deficiency: a global survey and identification of the Ecuadorean E180splice mutation in an oriental Jewish patient. Acta paediatrica (Oslo, Norway : 1992) Supplement 399:112-114.

Bethell L. 1984. The Cambridge history of Latin America. Cambridge: Cambridge University Press. 11 v. in 12 : ill. ; 24 cm. p.

Bhatia G, Patterson N, Sankararaman S, and Price AL. 2013. Estimating and interpreting FST: the impact of rare variants. Genome research 23(9):1514-1521.

Boca SM, and Rosenberg NA. 2011. Mathematical properties of Fst between admixed populations and their parental source populations. Theoretical population biology 80(3):208-216.

Bolnick DA, Raff JA, Springs LC, Reynolds AW, and Miró-Herrans AT. 2016. Native American Genomics and Population Histories. Annual Review of Anthropology 45(1):319-340.

Bomba L, Walter K, and Soranzo N. 2017. The impact of rare and low-frequency genetic variants in common disease. Genome biology 18(1):77.

Botigue LR, Henn BM, Gravel S, Maples BK, Gignoux CR, Corona E, Atzmon G, Burns E, Ostrer H, Flores C et al. . 2013. Gene flow from North Africa contributes to differential human genetic diversity in southern Europe. Proceedings of the National Academy of Sciences of the United States of America 110(29):11791-11796.

Boyd-Bowman P. 1964. Indice geobiográfico de cuarenta mil pobladores españoles de América en el siglo XVI. Bogotá: Instituto Caro y Cuervo. v p.

Boyd-Bowman P. 1976. Patterns of Spanish Emigration to the Indies until 1600. The Hispanic American Historical Review 56(4):580-604.

Boyd-Bowman P. 1985. Indice geobiográfico de más de 56 mil pobladores de la América Hispánica. México: Instituto de Investigaciones Históricas Fondo de Cultura Económica. lxvi,275p p.

Brandini S, Bergamaschi P, Cerna MF, Gandini F, Bastaroli F, Bertolini E, Cereda C, Ferretti L, Gómez-Carballa A, Battaglia V et al. . 2017. The Paleo-Indian Entry

into South America According to Mitogenomes. Molecular Biology and Evolution:msx267-msx267.

Broushaki F, Thomas MG, Link V, Lopez S, van Dorp L, Kirsanow K, Hofmanova Z, Diekmann Y, Cassidy LM, Diez-Del-Molino D et al. . 2016. Early Neolithic genomes from the eastern Fertile Crescent. Science.

Browning SR, Grinde K, Plantinga A, Gogarten SM, Stilp AM, Kaplan RC, Aviles-Santa ML, Browning BL, and Laurie CC. 2016. Local Ancestry Inference in a Large US-Based Hispanic/Latino Study: Hispanic Community Health Study/Study of Latinos (HCHS/SOL). G3 6(6):1525-1534.

Bryc K, Auton A, Nelson MR, Oksenberg JR, Hauser SL, Williams S, Froment A, Bodo JM, Wambebe C, Tishkoff SA et al. . 2010. Genome-wide patterns of population structure and admixture in West Africans and African Americans. Proceedings of the National Academy of Sciences of the United States of America 107(2):786-791.

Bryc K, Durand EY, Macpherson JM, Reich D, and Mountain JL. 2015. The genetic ancestry of African Americans, Latinos, and European Americans across the United States. American journal of human genetics 96(1):37-53.

Burchard EG, Ziv E, Coyle N, Gomez SL, Tang H, Karter AJ, Mountain JL, Pérez-Stable EJ, Sheppard D, and Risch N. 2003. The Importance of Race and Ethnic Background in Biomedical Research and Clinical Practice. New England Journal of Medicine 348(12):1170-1175.

Burkholder MA, and Johnson LL. 2003. Colonial Latin America: Oxford University Press.

Busby GB, Band G, Si Le Q, Jallow M, Bougama E, Mangano VD, Amenga-Etego LN, Enimil A, Apinjoh T, Ndila CM et al. . 2016. Admixture into and within sub-Saharan Africa. eLife 5.

Bustamante CD, De La Vega FM, and Burchard EG. 2011a. Genomics for the world. Nature 475:163.

Bustamante F, Olave E, and Binvignat O. 2011b. Estudio de Índices Faciales en Alumnos de la Universidad de La Frontera, Chile. International Journal of Morphology 29:1335-1340.

Campbell DD, Parra MV, Duque C, Gallego N, Franco L, Tandon A, Hunemeier T, Bortolini C, Villegas A, Bedoya G et al. . 2012. Amerind ancestry, socioeconomic status and the genetics of type 2 diabetes in a Colombian population. PloS one 7(4):e33570.

Campos-Sanchez R, Raventos H, and Barrantes R. 2013. Ancestry informative markers clarify the regional admixture variation in the Costa Rican population. Human biology 85(5):721-740.

Carvajal-Carmona LG, Ophoff R, Service S, Hartiala J, Molina J, Leon P, Ospina J, Bedoya G, Freimer N, and Ruiz-Linares A. 2003. Genetic demography of Antioquia (Colombia) and the Central Valley of Costa Rica. Human genetics 112(5-6):534-541.

Carvajal-Carmona LG, Soto ID, Pineda N, Ortiz-Barrientos D, Duque C, Ospina-Duque J, McCarthy M, Montoya P, Alvarez VM, Bedoya G et al. . 2000. Strong Amerind/white sex bias and a possible Sephardic contribution among the founders of a population in northwest Colombia. American journal of human genetics 67(5):1287-1295.

Cavalli-Sforza LL, and Bodmer WF. 1971. The genetics of human populations. San Francisco,: W.H. Freeman. xvi, 965 p. p.

Cavalli-Sforza LL, Menozzi P, and Piazza A. 1996. The history and geography of human genes. Princeton, N.J. ; Chichester: Princeton University Press. xiii,413p p.

Cerqueira CC, Hunemeier T, Gomez-Valdes J, Ramallo V, Volasko-Krause CD, Barbosa AA, Vargas-Pinilla P, Dornelles RC, Longo D, Rothhammer F et al. . 2014. Implications of the admixture process in skin color molecular assessment. PloS one 9(5):e96886.

Chacon-Duque JC, Adhikari K, Avendano E, Campo O, Ramirez R, Rojas W, Ruiz-Linares A, Restrepo BN, and Bedoya G. 2014. African genetic ancestry is associated with a protective effect on Dengue severity in colombian populations. Infect Genet Evol 27:89-95.

Chacon-Duque JC, Adhikari K, Fuentes-Guajardo M, Mendoza-Revilla J, Acuna-Alonzo V, Barquera Lozano R, Quinto-Sanchez M, Gomez-Valdes J, Everardo Martinez P, Villamil-Ramirez H et al. . 2018. Latin Americans show wide-spread Converso ancestry and the imprint of local Native ancestry on physical appearance. bioRxiv doi 10.1101/252155.

Chakraborti R. 1986. Gene Admixture in Human Populations: Models and Predictions. Yearbook of Physical Anthropology 29:1-43.

Chiang CWK, Marcus JH, Sidore C, Al-Asadi H, Zoledziewska M, Pitzalis M, Busonero F, Maschio A, Pistis G, Steri M et al. . 2016. Population history of the Sardinian people inferred from whole-genome sequencing. bioRxiv.

Churchhouse C, and Marchini J. 2013. Multiway admixture deconvolution using phased or unphased ancestral panels. Genetic epidemiology 37(1):1-12.

Churchill SE, Shackelford LL, Georgi JN, and Black MT. 2004. Morphological variation and airflow dynamics in the human nose. American Journal of Human Biology 16(6):625-638.

Comas J. 1960. Manual of physical anthropology. Springfield: Charles C. THomas.

Conley AB, Rishishwar L, Norris ET, Valderrama-Aguirre A, Marino-Ramirez L, Medina-Rivas MA, and Jordan IK. 2017. A Comparative Analysis of Genetic Ancestry and Admixture in the Colombian Populations of Choco and Medellin. G3 7(10):3435-3447.

Conomos MP, Laurie CA, Stilp AM, Gogarten SM, McHugh CP, Nelson SC, Sofer T, Fernandez-Rhodes L, Justice AE, Graff M et al. . 2016. Genetic Diversity and Association Studies in US Hispanic/Latino Populations: Applications in the Hispanic Community Health Study/Study of Latinos. American journal of human genetics 98(1):165-184.

Cook ND. 1998. Born to Die: Disease and New World Conquest, 1492-1650: Cambridge University Press.

Coussens AK, and Daal Av. 2005. Linkage disequilibrium analysis identifies an FGFR1 haplotype-tag SNP associated with normal variation in craniofacial shape. Genomics 85(5):563-573.

Crawford JE, Amaru R, Song J, Julian CG, Racimo F, Cheng JY, Guo X, Yao J, Ambale-Venkatesh B, Lima JA et al. . 2017a. Natural Selection on Genes Related to Cardiovascular Health in High-Altitude Adapted Andeans. The American Journal of Human Genetics 101(5):752-767.

Crawford MH, and Campbell BC. 2012. Causes and consequences of human migration : an evolutionary perspective. Cambridge: Cambridge University Press. xv, 550 p. p.

Crawford NG, Kelly DE, Hansen MEB, Beltrame MH, Fan S, Bowman SL, Jewett E, Ranciaro A, Thompson S, Lo Y et al. . 2017b. Loci associated with skin pigmentation identified in African populations. Science.

Creanza N, and Feldman MW. 2016. Worldwide genetic and cultural change in human evolution. Current Opinion in Genetics & Development 41:85-92.

Crow J. 2013. The Mapuche in Modern Chile: A Cultural History: University Press of Florida.

Curtin PD. 1969. The Atlantic slave trade : a census. Madison: University of Wisconsin Press. xix, 338 : maps ; 322 cm. p.

Darwin C. 1871. The descent of man, and selection in relation to sex. London: J. Murray.

Davies A. 1932. A Re-Survey of the Morphology of the Nose in Relation to Climate. The Journal of the Royal Anthropological Institute of Great Britain and Ireland 62:337-359.

De Mello Auricchio MT, Vicente JP, Meyer D, and Mingroni-Netto RC. 2007. Frequency and origins of hemoglobin S mutation in African-derived Brazilian populations. Human biology 79(6):667-677.

de Moura RR, de Queiroz Balbino V, Crovella S, and Brandao LA. 2016. On the use of Chinese population as a proxy of Amerindian ancestors in genetic admixture studies with Latin American populations. Eur J Hum Genet 24(3):326-327.

Delaneau O, Zagury JF, and Marchini J. 2013. Improved whole-chromosome phasing for disease and population genetic studies. Nature methods 10(1):5-6.

Denevan WM. 1992. The native population of the Americas in 1492. Madison, Wis. ; London: University of Wisconsin Press. xliv, 353 : ill., maps ; 324 cm. p.

DIAGRAM-Consortium, Asian Genetic Epidemiology Network Type 2 Diabetes C, South Asian Type 2 Diabetes C, Mexican American Type 2 Diabetes C, Type 2 Diabetes Genetic Exploration by Next-generation sequencing in multi-Ethnic Samples C, Mahajan A, Go MJ, Zhang W, Below JE, Gaulton KJ et al. . 2014. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. Nature genetics 46:234.

Dillehay TD. 2009. Probing deeper into first American studies. Proceedings of the National Academy of Sciences of the United States of America 106(4):971-978.

Edge MD, and Rosenberg NA. 2015. Implications of the apportionment of human genetic diversity for the apportionment of human phenotypic diversity. Stud Hist Philos Biol Biomed Sci 52:32-45.

Eichstaedt CA, Antao T, Pagani L, Cardona A, Kivisild T, and Mormina M. 2014. The Andean adaptive toolkit to counteract high altitude maladaptation: genome-wide and phenotypic analysis of the Collas. PloS one 9(3):e93314.

Elliott JH. 1984. The Spanish Conquest and settlement of America. In: Bethell L, editor. The Cambridge History of Latin America: Volume 1: Colonial Latin America. Cambridge: Cambridge University Press. p 147-206.

Ellis NA, Ciocci S, Proytcheva M, Lennon D, Groden J, and German J. 1998. The Ashkenazic Jewish Bloom syndrome mutation blmAsh is present in non-Jewish Americans of Spanish ancestry. American journal of human genetics 63(6):1685-1693.

Erlandson JM, and Braje TJ. 2011. From Asia to the Americas by boat? Paleogeography, paleoecology, and stemmed points of the northwest Pacific. Quaternary International 239(1):28-37.

Ettinger NA, Duggal P, Braz RF, Nascimento ET, Beaty TH, Jeronimo SM, Pearson RD, Blackwell JM, Moreno L, and Wilson ME. 2009. Genetic admixture in Brazilians exposed to infection with Leishmania chagasi. Ann Hum Genet 73(Pt 3):304-313.

Excoffier L. 2008. Analysis of Population Subdivision. Handbook of Statistical Genetics: John Wiley & Sons, Ltd. p 980-1020.

Eyheramendy S, Martinez FI, Manevy F, Vial C, and Repetto GM. 2015. Genetic structure characterization of Chileans reflects historical immigration patterns. Nature communications 6:6472.

Fagundes NJ, Kanitz R, Eckert R, Valls AC, Bogo MR, Salzano FM, Smith DG, Silva WA, Jr., Zago MA, Ribeiro-dos-Santos AK et al. . 2008. Mitochondrial population genomics supports a single pre-Clovis origin with a coastal route for the peopling of the Americas. American journal of human genetics 82(3):583-592.

Falush D, Stephens M, and Pritchard JK. 2003. Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies. Genetics 164(4):1567-1587.

Fawcett L, and Posada-Carbo E. 1997. Arabs and Jews in the development of the Colombian Caribbean 1850–1950. Immigrants & Minorities 16(1-2):57-79.

Fejerman L, Ahmadiyeh N, Hu D, Huntsman S, Beckman KB, Caswell JL, Tsung K, John EM, Torres-Mejia G, Carvajal-Carmona L et al. . 2014. Genome-wide association study of breast cancer in Latinas identifies novel protective variants on 6q25. Nature communications 5:5260.

Fejerman L, Chen GK, Eng C, Huntsman S, Hu D, Williams A, Pasaniuc B, John EM, Via M, Gignoux C et al. . 2012. Admixture mapping identifies a locus on 6q25 associated with breast cancer risk in US Latinas. Human molecular genetics 21(8):1907-1917.

Fenner JN. 2005. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. American journal of physical anthropology 128(2):415-423.

Filmer D, and Pritchett LH. 2001. Estimating wealth effects without expenditure data--or tears: an application to educational enrollments in states of India. Demography 38(1):115-132.

Florez JC, Price AL, Campbell D, Riba L, Parra MV, Yu F, Duque C, Saxena R, Gallego N, Tello-Ruiz M et al. . 2009. Strong association of socioeconomic status with genetic ancestry in Latinos: implications for admixture studies of type 2 diabetes. Diabetologia 52(8):1528-1536.

Fortes-Lima C, Gessain A, Ruiz-Linares A, Bortolini MC, Migot-Nabias F, Bellis G, Moreno-Mayar JV, Restrepo BN, Rojas W, Avendano-Tamayo E et al. . 2017. Genome-wide Ancestry and Demographic History of African-Descendant Maroon Communities from French Guiana and Suriname. American journal of human genetics 101(5):725-736.

Foster MW, and Sharp RR. 2002. Race, ethnicity, and genomics: social classifications as proxies of biological heterogeneity. Genome research 12(6):844-850.

Fregel R, Gomes V, Gusmão L, González AM, Cabrera VM, Amorim A, and Larruga JM. 2009. Demographic history of Canary Islands male gene-pool: replacement of native lineages by European. BMC Evolutionary Biology 9(1):181.

Frost P. 2014. The Puzzle of European Hair, Eye, and Skin Color. Advances in Anthropology 4(2):11.

Galanter JM, Gignoux CR, Torgerson DG, Roth LA, Eng C, Oh SS, Nguyen EA, Drake KA, Huntsman S, Hu D et al. . 2014. Genome-wide association study and

admixture mapping identify different asthma-associated loci in Latinos: the Genes-environments & Admixture in Latino Americans study. The Journal of allergy and clinical immunology 134(2):295-305.

Gamerman D. 1997. Markov chain Monte Carlo : stochastic simulation for Bayesian inference. London: Chapman & Hall. 240 p. p.

Gardner LI, Jr., Stern MP, Haffner SM, Gaskill SP, Hazuda HP, Relethford JH, and Eifler CW. 1984. Prevalence of diabetes in Mexican Americans. Relationship to percent of gene pool derived from native American sources. Diabetes 33(1):86-92.

Goebel T, Waters MR, and O'Rourke DH. 2008. The late Pleistocene dispersal of modern humans in the Americas. Science 319(5869):1497-1502.

Goetz LH, Uribe-Bruce L, Quarless D, Libiger O, and Schork NJ. 2014. Admixture and clinical phenotypic variation. Hum Hered 77(1-4):73-86.

Goldberg A, Mychajliw AM, and Hadly EA. 2016. Post-invasion demography of prehistoric humans in South America. Nature 532(7598):232-235.

Goldberg A, and Rosenberg NA. 2015. Beyond 2/3 and 1/3: The Complex Signatures of Sex-Biased Admixture on the X Chromosome. Genetics 201(1):263-279.

Goldberg A, Verdu P, and Rosenberg NA. 2014. Autosomal admixture levels are informative about sex bias in admixed populations. Genetics 198(3):1209-1229.

Gómez LD. 1970. Los quimbayas: Instituto Colombiano de Antropología.

Gonzalez BE, Borrell LN, Choudhry S, Naqvi M, Tsai HJ, Rodriguez-Santana JR, Chapela R, Rogers SD, Mei R, Rodriguez-Cintron W et al. . 2005. Latino populations: a unique opportunity for the study of race, genetics, and social environment in epidemiological research. Am JPublic Health 95(12):2161-2168.

Gravel S. 2012. Population genetics models of local ancestry. Genetics 191(2):607-619.

Green LD, Derr JN, and Knight A. 2000. mtDNA affinities of the peoples of North-Central Mexico. American journal of human genetics 66(3):989-998.

Griffiths AJF. 2012. Introduction to Genetic Analysis: W. H. Freeman and Company.

Grugni V, Battaglia V, Perego UA, Raveane A, Lancioni H, Olivieri A, Ferretti L, Woodward SR, Pascale JM, Cooke R et al. . 2015. Exploring the Y Chromosomal Ancestry of Modern Panamanians. PloS one 10(12):e0144223.

Guo Y, He J, Zhao S, Wu H, Zhong X, Sheng Q, Samuels DC, Shyr Y, and Long J. 2014. Illumina human exome genotyping array clustering and quality control. Nature Protocols 9:2643.

Gwatkin DR, Rutstein S, Johnson K, Suliman E, Wagstaff A, and Amouzou A. 2007. Socio-economic differences in health, nutrition, and population within developing countries: an overview. Nigerian journal of clinical practice 10(4):272-282.

Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, Brandt G, Nordenfelt S, Harney E, Stewardson K et al. . 2015. Massive migration from the steppe was a source for Indo-European languages in Europe. Nature 522(7555):207-211.

Haldane JBS. 1940. The Blood-Group Frequencies of European Peoples, and Racial Origins. Human biology 12(4):457-480.

Han E, Carbonetto P, Curtis RE, Wang Y, Granka JM, Byrnes J, Noto K, Kermany AR, Myres NM, Barber MJ et al. . 2017. Clustering of 770,000 genomes reveals post-colonial population structure of North America. Nature communications 8:14238.

Hellenthal G, Busby GB, Band G, Wilson JF, Capelli C, Falush D, and Myers S. 2014. A genetic atlas of human admixture history. Science 343(6172):747-751.

Hernandez-Pacheco N, Flores C, Alonso S, Eng C, Mak ACY, Hunstman S, Hu D, White MJ, Oh SS, Meade K et al. . 2017. Identification of a novel locus associated with skin colour in African-admixed populations. Scientific Reports 7:44548.

Hoban S, Bertorelle G, and Gaggiotti OE. 2012. Computer simulations: tools for population and evolutionary genetics. Nature Reviews Genetics 13:110.

Hodoğlugil U, and Mahley RW. 2012. Turkish Population Structure and Genetic Ancestry Reveal Relatedness among Eurasian Populations. Annals of human genetics 76(2):128-141.

Hofmanova Z, Kreutzer S, Hellenthal G, Sell C, Diekmann Y, Diez-Del-Molino D, van Dorp L, Lopez S, Kousathanas A, Link V et al. . 2016. Early farmers from across Europe directly descended from Neolithic Aegeans. Proceedings of the National Academy of Sciences of the United States of America 113(25):6886-6891.

Hoggart CJ, Shriver MD, Kittles RA, Clayton DG, and McKeigue PM. 2004. Design and analysis of admixture mapping studies. American journal of human genetics 74(5):965-978.

Holsinger KE, and Weir BS. 2009. Genetics in geographically structured populations: defining, estimating and interpreting FST. Nature Reviews Genetics 10:639.

Homburger JR, Moreno-Estrada A, Gignoux CR, Nelson D, Sanchez E, Ortiz-Tello P, Pons-Estel BA, Acevedo-Vasquez E, Miranda P, Langefeld CD et al. . 2015. Genomic Insights into the Ancestry and Demographic History of South America. PLoS genetics 11(12):e1005602.

Howie B, Marchini J, and Stephens M. 2011. Genotype imputation with thousands of genomes. G3 1(6):457-470.

Hubbe M, Hanihara T, and Harvati K. 2009. Climate signatures in the morphological differentiation of worldwide modern human populations. Anatomical record (Hoboken, NJ : 2007) 292(11):1720-1733.

Hudson RR. 1990. Gene genealogies and the coalescent process. Oxford Surveys in Evolutionary Biology 7:1-44.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18(2):337-338.

Hunter P. 2014. The genetics of human migrations. EMBO reports 15(10):1019-1022.

International HapMap C, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P et al. . 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449(7164):851-861.

Jackson JE. 2003. A user's guide to principal components. Hoboken, N.J: Wiley-Interscience. xvii, 569 : ill. ; 524 cm. p.

Jakobsson M, Edge MD, and Rosenberg NA. 2013. The relationship between F(ST) and the frequency of the most frequent allele. Genetics 193(2):515-528.

Jobling M, Hollox E, Hurles M, Kivisild T, and Tyler-Smith C. 2014. Human evolutionary genetics. New York: Garland Science. xviii, 670 pages : colour illustrations, maps, charts ; 628 cm p.

Johnson NA, Coram MA, Shriver MD, Romieu I, Barsh GS, London SJ, and Tang H. 2011. Ancestral components of admixed genomes in a Mexican cohort. PLoS genetics 7(12):e1002410.

Jombart T, Devillard S, and Balloux F. 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genetics 11(1):94.

Kamen H. 2002. Spain's road to empire: the making of a world power, 1492-1763: Allen Lane.

Kayser M. 2015. Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes. Forensic science international Genetics 18:33-48.

Kehdy FS, Gouveia MH, Machado M, Magalhaes WC, Horimoto AR, Horta BL, Moreira RG, Leal TP, Scliar MO, Soares-Souza GB et al. . 2015. Origin and dynamics of admixture in Brazilians and its effect on the pattern of deleterious mutations. Proceedings of the National Academy of Sciences of the United States of America 112(28):8696-8701.

Kent RB. 2016. Latin America : regions and people (second edition). New York: Guilford Press.

Kerminen S, Havulinna AS, Hellenthal G, Martin AR, Sarin AP, Perola M, Palotie A, Salomaa V, Daly MJ, Ripatti S et al. . 2017. Fine-Scale Genetic Structure in Finland. G3 7(10):3459-3468.

Kingman JFC. 1982. The Coalescent. Stochastic Processes and their Applications 13:235 - 248.

Koehl AJ, and Long JC. 2018. The contributions of admixture and genetic drift to diversity among post-contact populations in the Americas. American journal of physical anthropology 165(2):256-268.

Kolenikov S, and Angeles G. 2009. Socioeconomic Status measurement with discrete proxy variables: is Principal Component Analysis a reliable answer? Review of Income and Wealth 55(1):128-165.

Lamason RL, Mohideen MA, Mest JR, Wong AC, Norton HL, Aros MC, Jurynec MJ, Mao X, Humphreville VR, Humbert JE et al. . 2005. SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. Science 310(5755):1782-1786.

Lavrin A. 1992. Sexuality and Marriage in Colonial Latin America: University of Nebraska Press.

Lawson D, van Dorp L, and Falush D. 2017. A tutorial on how (not) to over-interpret STRUCTURE/ADMIXTURE bar plots. bioRxiv.

Lawson DJ, and Falush D. 2012. Population identification using genetic data. Annual review of genomics and human genetics 13:337-361.

Lawson DJ, Hellenthal G, Myers S, and Falush D. 2012. Inference of population structure using dense haplotype data. PLoS genetics 8(1):e1002453.

Lee C, Abdool A, and Huang CH. 2009. PCA-based population structure inference with generic clustering algorithms. BMC bioinformatics 10 Suppl 1:S73.

Leinonen T, McCairns RJS, O'Hara RB, and Merilä J. 2013. QST–FST comparisons: evolutionary and ecological insights from genomic heterogeneity. Nature Reviews Genetics 14:179.

Leong SC, and Eccles R. 2009. A systematic review of the nasal index and the significance of the shape and size of the nose in rhinology. Clinical otolaryngology : official journal of ENT-UK ; official journal of Netherlands Society for Oto-Rhino-Laryngology & Cervico-Facial Surgery 34(3):191-198.

Leslie S, Winney B, Hellenthal G, Davison D, Boumertit A, Day T, Hutnik K, Royrvik EC, Cunliffe B, Wellcome Trust Case Control C et al. . 2015. The fine-scale genetic structure of the British population. Nature 519(7543):309-314.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL et al. . 2008. Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation. Science 319(5866):1100-1104.

Li N, and Stephens M. 2003. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics 165(4):2213-2233.

Lindo J, Huerta-Sanchez E, Nakagome S, Rasmussen M, Petzelt B, Mitchell J, Cybulski JS, Willerslev E, DeGiorgio M, and Malhi RS. 2016. A time transect of exomes from a Native American population before and after European contact. Nature communications 7:13175.

Little AC, Jones BC, and DeBruine LM. 2011. Facial attractiveness: evolutionary based research. Philosophical Transactions of the Royal Society B: Biological Sciences 366(1571):1638-1659.

Liu F, van der Lijn F, Schurmann C, Zhu G, Chakravarty MM, Hysi PG, Wollstein A, Lao O, de Bruijne M, Ikram MA et al. . 2012. A genome-wide association study identifies five loci influencing facial morphology in Europeans. PLoS genetics 8(9):e1002932.

Liu F, Visser M, Duffy DL, Hysi PG, Jacobs LC, Lao O, Zhong K, Walsh S, Chaitanya L, Wollstein A et al. . 2015. Genetics of skin color variation in Europeans: genome-wide association studies with functional follow-up. Human genetics 134(8):823-835.

Llamas B, Fehren-Schmitz L, Valverde G, Soubrier J, Mallick S, Rohland N, Nordenfelt S, Valdiosera C, Richards SM, Rohrlach A et al. . 2016. Ancient mitochondrial DNA provides high-resolution time scale of the peopling of the Americas. Science Advances 2(4).

Loh PR, Lipson M, Patterson N, Moorjani P, Pickrell JK, Reich D, and Berger B. 2013. Inferring admixture histories of human populations using linkage disequilibrium. Genetics 193(4):1233-1254.

Loveman M. 2014. National Colors: Racial Classification and the State in Latin America: Oxford University Press.

Maca-Meyer N, Arnay M, Rando JC, Flores C, González AM, Cabrera VM, and Larruga JM. 2003. Ancient mtDNA analysis and the origin of the Guanches. European Journal Of Human Genetics 12:155.

Malécot G. 1948. Les mathématiques de l'hérédité: Barnéoud frères.

Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, Chennagiri N, Nordenfelt S, Tandon A et al. . 2016. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature 538(7624):201-206.

Manica A, Amos W, Balloux F, and Hanihara T. 2007. The effect of ancient population bottlenecks on human phenotypic variation. Nature 448(7151):346-348.

Maples BK, Gravel S, Kenny EE, and Bustamante CD. 2013. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. American journal of human genetics 93(2):278-288.

Markus B, Alshafee I, and Birk OS. 2014. Deciphering the fine-structure of tribal admixture in the Bedouin population using genomic data. Heredity 112(2):182-189.

Marrero AR, Bravi C, Stuart S, Long JC, Pereira das Neves Leite F, Kommers T, Carvalho CM, Pena SD, Ruiz-Linares A, Salzano FM et al. . 2007. Pre- and post-

Columbian gene and cultural continuity: the case of the Gaucho from southern Brazil. Hum Hered 64(3):160-171.

Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, Daly MJ, Bustamante CD, and Kenny EE. 2017. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. American journal of human genetics 100(4):635-649.

Mathias RA, Taub MA, Gignoux CR, Fu W, Musharoff S, O'Connor TD, Vergara C, Torgerson DG, Pino-Yanes M, Shringarpure SS et al. . 2016. A continuum of admixture in the Western Hemisphere revealed by the African Diaspora genome. Nature communications 7:12522.

McEvoy BP, and Visscher PM. 2009. Genetics of human height. Economics & Human Biology 7(3):294-306.

McKeigue PM. 1998. Mapping genes that underlie ethnic differences in disease risk: methods for detecting linkage in admixed populations, by conditioning on parental admixture. American journal of human genetics 63(1):241-251.

McVean G. 2009. A Genealogical Interpretation of Principal Components Analysis. PLoS genetics 5(10):e1000686.

Meade TA. 2016. A history of modern Latin America : 1800 to the present. Chichester, West Sussex, UK: Wiley Blackwell. xviii, 388 pages : illustrations ; 325 cm. p.

Meltzer DJ. 2009. First peoples in a new world : colonizing ice age America. Berkeley ; London: University of California Press. xviii, 446 416 of plates : ill. (some col.), maps ; 427 cm. p.

Menozzi P, Piazza A, and Cavalli-Sforza L. 1978. Synthetic maps of human gene frequencies in Europeans. Science 201(4358):786-792.

Meyer D, VR CA, Bitarello BD, DY CB, and Nunes K. 2017. A genomic perspective on HLA evolution. Immunogenetics.

Milligan BG. 2003. Maximum-likelihood estimation of relatedness. Genetics 163(3):1153-1167.

Montinaro F, Busby GB, Pascali VL, Myers S, Hellenthal G, and Capelli C. 2015. Unravelling the hidden ancestry of American admixed populations. Nature communications 6:6596.

Moorjani P, Patterson N, Hirschhorn JN, Keinan A, Hao L, Atzmon G, Burns E, Ostrer H, Price AL, and Reich D. 2011. The history of African gene flow into Southern Europeans, Levantines, and Jews. PLoS genetics 7(4):e1001373.

Moreno-Estrada A, Gignoux CR, Fernandez-Lopez JC, Zakharia F, Sikora M, Contreras AV, Acuna-Alonzo V, Sandoval K, Eng C, Romero-Hidalgo S et al. . 2014. Human genetics. The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. Science 344(6189):1280-1285.

Moreno-Estrada A, Gravel S, Zakharia F, McCauley JL, Byrnes JK, Gignoux CR, Ortiz-Tello PA, Martinez RJ, Hedges DJ, Morris RW et al. . 2013. Reconstructing the population genetic history of the Caribbean. PLoS genetics 9(11):e1003925.

Morner M. 1967. Race mixture in the history of Latin America. Boston,: Little. xii, 178 p. p.

Mountain JL, and Risch N. 2004. Assessing genetic contributions to phenotypic differences among 'racial' and 'ethnic' groups. Nature genetics 36:S48.

Mullineaux LG, Castellano TM, Shaw J, Axell L, Wood ME, Diab S, Klein C, Sitarik M, Deffenbaugh AM, and Graw SL. 2003. Identification of germline 185delAG

BRCA1 mutations in non-Jewish Americans of Spanish ancestry from the San Luis Valley, Colorado. Cancer 98(3):597-602.

Naftali S, Rosenfeld M, Wolf M, and Elad D. 2005. The air-conditioning capacity of the human nose. Annals of biomedical engineering 33(4):545-553.

Negus VE. 1954. Introduction to the Comparative Anatomy of the Nose and Paranasal Sinuses: Hunterian Lecture delivered at the Royal College of Surgeons of England on 20th May 1954. Annals of The Royal College of Surgeons of England 15(3):141-173.

Norton HL, Kittles RA, Parra E, McKeigue P, Mao X, Cheng K, Canfield VA, Bradley DG, McEvoy B, and Shriver MD. 2007. Genetic evidence for the convergent evolution of light skin in Europeans and East Asians. Mol Biol Evol 24(3):710-722.

Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR et al. . 2008. Genes mirror geography within Europe. Nature 456:98.

Novembre J, and Peter BM. 2016. Recent advances in the study of fine-scale population structure in humans. Curr Opin Genet Dev 41:98-105.

Novembre J, and Ramachandran S. 2011. Perspectives on human population structure at the cusp of the sequencing era. Annual review of genomics and human genetics 12:245-274.

Novembre J, and Stephens M. 2008. Interpreting principal component analyses of spatial population genetic variation. Nature genetics 40(5):646-649.

O'Connell J, Gurdasani D, Delaneau O, Pirastu N, Ulivi S, Cocca M, Traglia M, Huang J, Huffman JE, Rudan I et al. . 2014. A general approach for haplotype phasing across the full spectrum of relatedness. PLoS genetics 10(4):e1004234.

O'Fallon BD, and Fehren-Schmitz L. 2011. Native Americans experienced a strong population bottleneck coincident with European contact. Proceedings of the National Academy of Sciences of the United States of America 108(51):20444-20448.

Ohlrogge S. 2008. Anthropological atlas of male facial features. Frankfurt: Verlag fur Polizeiwissenschaft.

Olivieri A, Sidore C, Achilli A, Angius A, Posth C, Furtwängler A, Brandini S, Capodiferro MR, Gandini F, Zoledziewska M et al. . 2017. Mitogenome Diversity in Sardinians: A Genetic Window onto an Island's Past. Molecular Biology and Evolution 34(5):1230-1239.

Olsson U. 1979. Maximum likelihood estimation of the polychoric correlation coefficient. Psychometrika 44(4):443-460.

Pagani L, Kivisild T, Tarekegn A, Ekong R, Plaster C, Gallego Romero I, Ayub Q, Mehdi SQ, Thomas MG, Luiselli D et al. . 2012. Ethiopian genetic diversity reveals linguistic stratification and complex influences on the Ethiopian gene pool. American journal of human genetics 91(1):83-96.

Parra EJ, Kittles RA, and Shriver MD. 2004. Implications of correlations between skin color and genetic ancestry for biomedical research. Nature genetics 36(11 Suppl):S54-60.

Parra EJ, Marcini A, Akey J, Martinson J, Batzer MA, Cooper R, Forrester T, Allison DB, Deka R, Ferrell RE et al. . 1998. Estimating African American Admixture Proportions by Use of Population-Specific Alleles. The American Journal of Human Genetics 63(6):1839-1851.

Parra FC, Amado RC, Lambertucci JR, Rocha J, Antunes CM, and Pena SD. 2003. Color and genomic ancestry in Brazilians. Proceedings of the National Academy of Sciences of the United States of America 100(1):177-182.

Parsons JJ. 1968. Antioqueño colonization in Western Colombia: University of California Press.

Pasaniuc B, Zaitlen N, Lettre G, Chen GK, Tandon A, Kao WH, Ruczinski I, Fornage M, Siscovick DS, Zhu X et al. . 2011. Enhanced statistical tests for GWAS in admixed populations: assessment using African Americans from CARe and a Breast Cancer Consortium. PLoS genetics 7(4):e1001371.

Paternoster L, Standl M, Waage J, Baurecht H, Hotze M, Strachan DP, Curtin JA, Bonnelykke K, Tian C, Takahashi A et al. . 2015. Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis. Nature genetics 47(12):1449-1456.

Paternoster L, Zhurov AI, Toma AM, Kemp JP, St Pourcain B, Timpson NJ, McMahon G, McArdle W, Ring SM, Smith GD et al. . 2012. Genome-wide association study of three-dimensional facial morphology identifies a variant in PAX3 associated with nasion position. American journal of human genetics 90(3):478-485.

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, and Reich D. 2012. Ancient admixture in human history. Genetics 192(3):1065-1093.

Patterson N, Price AL, and Reich D. 2006. Population structure and eigenanalysis. PLoS genetics 2(12):e190.

Pedersen MW, Ruter A, Schweger C, Friebe H, Staff RA, Kjeldsen KK, Mendoza ML, Beaudoin AB, Zutter C, Larsen NK et al. . 2016. Postglacial viability and colonization in North America's ice-free corridor. Nature 537(7618):45-49.

Pena SD, Di Pietro G, Fuchshuber-Moraes M, Genro JP, Hutz MH, Kehdy Fde S, Kohlrausch F, Magno LA, Montenegro RC, Moraes MO et al. . 2011. The genomic ancestry of individuals from different geographical regions of Brazil is more uniform than expected. PloS one 6(2):e17063.

Pfaff CL, Parra EJ, Bonilla C, Hiester K, McKeigue PM, Kamboh MI, Hutchinson RG, Ferrell RE, Boerwinkle E, and Shriver MD. 2001. Population Structure in Admixed Populations: Effect of Admixture Dynamics on the Pattern of Linkage Disequilibrium. The American Journal of Human Genetics 68(1):198-207.

Phillips C. 2015. Forensic genetic analysis of bio-geographical ancestry. Forensic science international Genetics 18:49-65.

Phillips C, Parson W, Lundsberg B, Santos C, Freire-Aradas A, Torres M, Eduardoff M, Borsting C, Johansen P, Fondevila M et al. . 2014. Building a forensic ancestry panel from the ground up: The EUROFORGEN Global AIM-SNP set. Forensic science international Genetics 11:13-25.

Pickrell JK, Patterson N, Loh P-R, Lipson M, Berger B, Stoneking M, Pakendorf B, and Reich D. 2014. Ancient west Eurasian ancestry in southern and eastern Africa. Proceedings of the National Academy of Sciences 111(7):2632-2637.

Pickrell JK, and Reich D. 2014. Toward a new history and geography of human genes informed by ancient DNA. Trends in Genetics 30(9):377-389.

Pompanon F, Bonin A, Bellemain E, and Taberlet P. 2005. Genotyping errors: causes, consequences and solutions. Nature Reviews Genetics 6:847.

Pompanon F, Bonin A, and SpringerLink (Online service). 2012. Data Production and Analysis in Population Genomics Methods and Protocols. Methods in Molecular

Biology, Methods and Protocols,. Totowa, NJ: Humana Press : Imprint: Humana Press,. p XI, 337 p. 367 illus., 316 illus. in color.

Popejoy AB, and Fullerton SM. 2016. Genomics is failing on diversity. Nature 538(7624):161-164.

Price AL, Patterson N, Yu F, Cox DR, Waliszewska A, McDonald GJ, Tandon A, Schirmer C, Neubauer J, Bedoya G et al. . 2007. A genomewide admixture map for Latino populations. American journal of human genetics 80(6):1024-1036.

Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski I, Beaty TH, Mathias R, Reich D, and Myers S. 2009. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS genetics 5(6):e1000519.

Pritchard JK, Stephens M, and 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 MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ et al. . 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics 81(3):559-575.

Qin H, and Zhu X. 2012. Power comparison of admixture mapping and direct association analysis in genome-wide association studies. Genetic epidemiology 36(3):235-243.

Quinto-Sanchez M, Adhikari K, Acuna-Alonzo V, Cintas C, Silva de Cerqueira CC, Ramallo V, Castillo L, Farrera A, Jaramillo C, Arias W et al. . 2015. Facial asymmetry and genetic ancestry in Latin American admixed populations. American journal of physical anthropology 157(1):58-70.

R-Core-Team. 2013. R: A language and environment for statistical computing.

Raghavan M, Steinrucken M, Harris K, Schiffels S, Rasmussen S, DeGiorgio M, Albrechtsen A, Valdiosera C, Avila-Arcos MC, Malaspinas AS et al. . 2015. POPULATION GENETICS. Genomic evidence for the Pleistocene and recent population history of Native Americans. Science 349(6250):aab3884.

Rasmussen M, Anzick SL, Waters MR, Skoglund P, DeGiorgio M, Stafford TW, Jr., Rasmussen S, Moltke I, Albrechtsen A, Doyle SM et al. . 2014. The genome of a Late Pleistocene human from a Clovis burial site in western Montana. Nature 506(7487):225-229.

Reich D, Patterson N, Campbell D, Tandon A, Mazieres S, Ray N, Parra MV, Rojas W, Duque C, Mesa N et al. . 2012. Reconstructing Native American population history. Nature 488(7411):370-374.

Reich D, Price AL, and Patterson N. 2008. Principal component analysis of genetic data. Nature genetics 40(5):491-492.

Reich D, Thangaraj K, Patterson N, Price AL, and Singh L. 2009. Reconstructing Indian population history. Nature 461(7263):489-494.

Reichel-Dolmatoff G, Oyuela-Caycedo A, and Raymond JS. 1998. Recent Advances in the Archaeology of the Northern Andes: In Memory of Gerardo Reichel-Dolmatoff: Institute of Archaeology, University of California, Los Angeles.

Relethford JH. 2002. Apportionment of global human genetic diversity based on craniometrics and skin color. American journal of physical anthropology 118(4):393-398.

Relethford JH. 2009. Race and global patterns of phenotypic variation. American journal of physical anthropology 139(1):16-22.

Reyes-Centeno H, Ghirotto S, Detroit F, Grimaud-Herve D, Barbujani G, and Harvati K. 2014. Genomic and cranial phenotype data support multiple modern human dispersals from Africa and a southern route into Asia. Proceedings of the National Academy of Sciences of the United States of America 111(20):7248-7253.

Risch N. 2006. Dissecting Racial and Ethnic Differences. New England Journal of Medicine 354(4):408-411.

Risch N, Burchard E, Ziv E, and Tang H. 2002. Categorization of humans in biomedical research: genes, race and disease. Genome biology 3(7):comment2007.2001-comment2007.2012.

Ritz-Timme S, Gabriel P, Tutkuviene J, Poppa P, Obertova Z, Gibelli D, De Angelis D, Ratnayake M, Rizgeliene R, Barkus A et al. . 2011. Metric and morphological assessment of facial features: a study on three European populations. Forensic science international 207(1-3):239.e231-238.

Rivet P. 1943. Les origines de l'homme américain. Montréal, Canada: Éditions de l'Arbre. 132 : ill. ; 120 cm. p.

Rodríguez-Varela R, Günther T, Krzewińska M, Storå J, Gillingwater TH, MacCallum M, Arsuaga JL, Dobney K, Valdiosera C, Jakobsson M et al. . 2017. Genomic Analyses of Pre-European Conquest Human Remains from the Canary Islands Reveal Close Affinity to Modern North Africans. Current Biology 27(21):3396-3402.e3395.

Romero-Hidalgo S, Ochoa-Leyva A, Garciarrubio A, Acuna-Alonzo V, Antunez-Arguelles E, Balcazar-Quintero M, Barquera-Lozano R, Carnevale A, Cornejo-Granados F, Fernandez-Lopez JC et al. . 2017. Demographic history and biologically relevant genetic variation of Native Mexicans inferred from whole-genome sequencing. Nature communications 8(1):1005.

Romero RC. 2010. The Chinese in Mexico, 1882-1940: University of Arizona Press.

Rosenberg NA, Huang L, Jewett EM, Szpiech ZA, Jankovic I, and Boehnke M. 2010. Genome-wide association studies in diverse populations. Nature reviews Genetics 11(5):356-366.

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, and Feldman MW. 2002. Genetic structure of human populations. Science 298(5602):2381-2385.

Royal CD, Novembre J, Fullerton SM, Goldstein DB, Long JC, Bamshad MJ, and Clark AG. 2010. Inferring genetic ancestry: opportunities, challenges, and implications. American journal of human genetics 86(5):661-673.

Ruiz-Linares A. 2014. How Genes Have Illuminated the History of Early Americans and Latino Americans. Cold Spring Harb Perspect Biol.

Ruiz-Linares A, Adhikari K, Acuna-Alonzo V, Quinto-Sanchez M, Jaramillo C, Arias W, Fuentes M, Pizarro M, Everardo P, de Avila F et al. . 2014. Admixture in Latin America: geographic structure, phenotypic diversity and self-perception of ancestry based on 7,342 individuals. PLoS genetics 10(9):e1004572.

Sachar HM. 1994. Farewell España: the world of the Sephardim remembered: Knopf.

Salvatore RD, Coatsworth JH, and Challú AE. 2010. Living standards in Latin American history : height, welfare, and development, 1750-2000. Cambridge, Mass. ; London: Harvard University David Rockefeller Center for Latin American Studies. iii, 313 : ill. ; 323 cm. p.

Salzano FM. 2016. The role of natural selection in human evolution - insights from Latin America. Genetics and molecular biology 39(3):302-311.

Salzano FM, and Bortolini MC. 2002. The evolution and genetics of Latin American populations. Cambridge ; New York: Cambridge University Press. xvi, 512 p. p.

Salzano FM, and Sans M. 2014. Interethnic admixture and the evolution of Latin American populations. Genetics and molecular biology 37(1 Suppl):151-170.

Sanchez-Albornoz N. 1974. The population of Latin America: a history. Berkeley,: University of California Press. xv, 299 p. p.

Sánchez-Albornoz N. 1994. La población de América Latina : desde los tiempos precolombinos al año 2025. Madrid: Alianza Editorial.

Sandoval JR, Salazar-Granara A, Acosta O, Castillo-Herrera W, Fujita R, Pena SD, and Santos FR. 2013. Tracing the genomic ancestry of Peruvians reveals a major legacy of pre-Columbian ancestors. J Hum Genet 58(9):627-634.

Schlebusch CM, Skoglund P, Sjodin P, Gattepaille LM, Hernandez D, Jay F, Li S, De Jongh M, Singleton A, Blum MG et al. . 2012. Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. Science 338(6105):374-379.

Seldin MF, Pasaniuc B, and Price AL. 2011. New approaches to disease mapping in admixed populations. Nature reviews Genetics 12(8):523-528.

Shaffer JR, Orlova E, Lee MK, Leslie EJ, Raffensperger ZD, Heike CL, Cunningham ML, Hecht JT, Kau CH, Nidey NL et al. . 2016. Genome-Wide Association Study Reveals Multiple Loci Influencing Normal Human Facial Morphology. PLoS genetics 12(8):e1006149.

Shriver MD, Parra EJ, Dios S, Bonilla C, Norton H, Jovel C, Pfaff C, Jones C, Massac A, Cameron N et al. . 2003. Skin pigmentation, biogeographical ancestry and admixture mapping. Human genetics 112(4):387-399.

SIGMA-T2D-Consortium. 2013. Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. Nature 506:97.

Silva-Zolezzi I, Hidalgo-Miranda A, Estrada-Gil J, Fernandez-Lopez JC, Uribe-Figueroa L, Contreras A, Balam-Ortiz E, del Bosque-Plata L, Velazquez-Fernandez D, Lara C et al. . 2009. Analysis of genomic diversity in Mexican Mestizo populations to develop genomic medicine in Mexico. Proceedings of the National Academy of Sciences of the United States of America 106(21):8611-8616.

Skoglund P, Mallick S, Bortolini MC, Chennagiri N, Hunemeier T, Petzl-Erler ML, Salzano FM, Patterson N, and Reich D. 2015. Genetic evidence for two founding populations of the Americas. Nature 525(7567):104-108.

Skoglund P, and Reich D. 2016. A genomic view of the peopling of the Americas. Curr Opin Genet Dev 41:27-35.

Slatkin M. 2008. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. Nature reviews Genetics 9(6):477-485.

Soejima M, and Koda Y. 2007. Population differences of two coding SNPs in pigmentation-related genes SLC24A5 and SLC45A2. International journal of legal medicine 121(1):36-39.

Stepan N. 1991. "The Hour of Eugenics": Race, Gender, and Nation in Latin America: Cornell University Press.

Stokowski RP, Pant PV, Dadd T, Fereday A, Hinds DA, Jarman C, Filsell W, Ginger RS, Green MR, van der Ouderaa FJ et al. . 2007. A genomewide association study of skin pigmentation in a South Asian population. American journal of human genetics 81(6):1119-1132.

Stulp G, Barrett L, Tropf FC, and Mills M. 2015. Does natural selection favour taller stature among the tallest people on earth? Proceedings of the Royal Society B: Biological Sciences 282(1806).

Tang H, Choudhry S, Mei R, Morgan M, Rodriguez-Cintron W, Burchard EG, and Risch NJ. 2007. Recent genetic selection in the ancestral admixture of Puerto Ricans. American journal of human genetics 81(3):626-633.

Tang H, Jorgenson E, Gadde M, Kardia SL, Rao DC, Zhu X, Schork NJ, Hanis CL, and Risch N. 2006. Racial admixture and its impact on BMI and blood pressure in African and Mexican Americans. Human genetics 119(6):624-633.

Tang H, Quertermous T, Rodriguez B, Kardia SLR, Zhu X, Brown A, Pankow JS, Province MA, Hunt SC, Boerwinkle E et al. . 2005. Genetic Structure, Self-Identified Race/Ethnicity, and Confounding in Case-Control Association Studies. The American Journal of Human Genetics 76(2):268-275.

Thomas H. 1997. The slave trade : the history of the Atlantic slave trade, 1440-1870. New York: Simon & Schuster. 908 , 916 leaves of plates : ill, maps, ports. p.

Thompson EA. 2013. Identity by descent: variation in meiosis, across genomes, and in populations. Genetics 194(2):301-326.

Thornton R. 1987. American Indian holocaust and survival : a population history since 1492. Norman: University of Oklahoma Press. xx, 292 : ill., maps ; 224 cm. p.

Thornton TA, and Bermejo JL. 2014. Local and global ancestry inference and applications to genetic association analysis for admixed populations. Genetic epidemiology 38 Suppl 1:S5-S12.

Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, Hirbo JB, Awomoyi AA, Bodo JM, Doumbo O et al. . 2009. The genetic structure and history of Africans and African Americans. Science 324(5930):1035-1044.

Tishkoff SA, and Verrelli BC. 2003. Patterns of Human Genetic Diversity: Implications for Human Evolutionary History and Disease. Annual review of genomics and human genetics 4(1):293-340.

Torero A. 2005. Idiomas de los Andes : lingüística e historia. Lima: IFEA, Instituto Francés de Estudios Andinos : Editorial Horizonte. 565 : ill., maps ; 524 cm. p.

Tyler K. 2008. Ethnographic Approaches to Race, Genetics and Genealogy. Sociology Compass 2(6):1860-1877.

van Dorp L, Balding D, Myers S, Pagani L, Tyler-Smith C, Bekele E, Tarekegn A, Thomas MG, Bradman N, and Hellenthal G. 2015. Evidence for a Common Origin of Blacksmiths and Cultivators in the Ethiopian Ari within the Last 4500 Years: Lessons for Clustering-Based Inference. PLoS genetics 11(8):e1005397.

Velez C, Palamara PF, Guevara-Aguirre J, Hao L, Karafet T, Guevara-Aguirre M, Pearlman A, Oddoux C, Hammer M, Burns E et al. . 2012. The impact of Converso Jews on the genomes of modern Latin Americans. Human genetics 131(2):251-263.

Vergara C, Murray T, Rafaels N, Lewis R, Campbell M, Foster C, Gao L, Faruque M, Oliveira RR, Carvalho E et al. . 2013. African ancestry is a risk factor for asthma and high total IgE levels in African admixed populations. Genetic epidemiology 37(4):393-401.

Via M, Gignoux CR, Roth LA, Fejerman L, Galanter J, Choudhry S, Toro-Labrador G, Viera-Vera J, Oleksyk TK, Beckman K et al. . 2011. History shaped the geographic distribution of genomic admixture on the island of Puerto Rico. PloS one 6(1):e16513.

Wade L. 2017. On the trail of ancient mariners. Science 357(6351):542-545.

Wade P. 2009. Race and Sex in Latin America: Pluto Press.

Wade P, López Beltrán C, Restrepo E, and Santos RV. 2014. Mestizo genomics : race mixture, nation, and science in Latin America. Durham ; London: Duke University Press. xii, 304 pages : illustrations, map ; 324 cm. p.

Wang S, Lewis CM, Jakobsson M, Ramachandran S, Ray N, Bedoya G, Rojas W, Parra MV, Molina JA, Gallo C et al. . 2007. Genetic variation and population structure in native Americans. PLoS genetics 3(11):e185.

Wang S, Ray N, Rojas W, Parra MV, Bedoya G, Gallo C, Poletti G, Mazzotti G, Hill K, Hurtado AM et al. . 2008. Geographic patterns of genome admixture in Latin American Mestizos. PLoS genetics 4(3):e1000037.

Wang X, Zhu X, Qin H, Cooper RS, Ewens WJ, Li C, and Li M. 2011. Adjustment for local ancestry in genetic association analysis of admixed populations. Bioinformatics 27(5):670-677.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. Theoretical population biology 7(2):256-276.

Webster TH, and Wilson Sayres MA. 2016. Genomic signatures of sex-biased demography: progress and prospects. Curr Opin Genet Dev 41:62-71.

Weir BS, Anderson AD, and Hepler AB. 2006. Genetic relatedness analysis: modern data and new challenges. Nature Reviews Genetics 7:771.

Weir BS, and Cockerham CC. 1984. Estimating F-Statistics for the Analysis of Population Structure. Evolution 38(6):1358-1370.

Wilkins JF. 2006. Unraveling male and female histories from human genetic data. Curr Opin Genet Dev 16(6):611-617.

Williams R, and Wienroth M. 2017. Social and ethical aspects of forensic genetics: A critical review. Forensic science review 29(2):145-169.

Williams RC, Long JC, Hanson RL, Sievers ML, and Knowler WC. 2000. Individual Estimates of European Genetic Admixture Associated with Lower Body-Mass Index, Plasma Glucose, and Prevalence of Type 2 Diabetes in Pima Indians. American journal of human genetics 66(2):527-538.

Willing E-M, Dreyer C, and van Oosterhout C. 2012. Estimates of Genetic Differentiation Measured by F(ST) Do Not Necessarily Require Large Sample Sizes When Using Many SNP Markers. PloS one 7(8):e42649.

Wilson RT, Roff AN, Dai PJ, Fortugno T, Douds J, Chen G, Grove GL, Nikiforova SO, Barnholtz-Sloan J, Frudakis T et al. . 2011. Genetic Ancestry, Skin Reflectance and Pigmentation Genotypes in Association with Serum Vitamin D Metabolite Balance. Hormone molecular biology and clinical investigation 7(1):279-293.

Winkler CA, Nelson GW, and Smith MW. 2010. Admixture mapping comes of age. Annual review of genomics and human genetics 11:65-89.

Wojcik G, Graff M, Nishimura KK, Tao R, Haessler J, Gignoux CR, Highland HM, Patel YM, Sorokin EP, Avery CL et al. . 2017. Genetic Diversity Turns a New PAGE in Our Understanding of Complex Traits. bioRxiv.

Wright S. 1951. The genetical structure of populations. Ann Eugen 15(4):323-354.

Yudell M, Roberts D, DeSalle R, and Tishkoff S. 2016. Taking race out of human genetics. Science 351(6273):564-565.

Zaidi AA, Mattern BC, Claes P, McEcoy B, Hughes C, and Shriver MD. 2017. Investigating the case of human nose shape and climate adaptation. PLoS genetics 13(3):e1006616.

Zaitlen N, Pasaniuc B, Sankararaman S, Bhatia G, Zhang J, Gusev A, Young T, Tandon A, Pollack S, Vilhjálmsson BJ et al. . 2014. Leveraging population admixture to explain missing heritability of complex traits. Nature genetics 46(12):1356-1362.

Zhu JH, Lee HP, Lim KM, Lee SJ, and Wang de Y. 2011. Evaluation and comparison of nasal airway flow patterns among three subjects from Caucasian, Chinese and Indian ethnic groups using computational fluid dynamics simulation. Respiratory physiology & neurobiology 175(1):62-69.

Appendix. Description of the 129 clusters generated by fineSTRUCTURE and associated analyses.

See explanatory notes at the bottom of the table.

fS Clust	Contains	N	Decision	Explanation of decision	Donor/*Surrogate*	Surrog	Additional notes
1	South.Sudan(1/8)	1	Donor	Single sample cluster	Out.SouthSudan		
2	Ethiopia(3/3)+South.Sudan(7/8)	10	Surrogate		*EastAfrica1*	*1*	
3	Kenya(35/73)	35	Surrogate (Merged)	Similar according to TVD and tree distance	*EastAfrica2*	*2*	No clear assign-ment
4	Kenya(38/73)	38					
5	Namibia.3(1/9)	1	Donor	Single sample cluster	Out.Namibia.3		
6	Namibia.3(1/9)*	1	Donor	Single sample cluster	Out.Namibia.3		
7	Namibia.3(6/9)	6	Surrogate		*Namibia*	*3*	
8	Namibia.2(1/14)+Namibia.3(1/9)	2	Donor	Similar to **7**, no contribu-tion	Out.Namibia.2 Out.Namibia.3		
9	South.Africa.3(1/19)	1	Donor	Single sample cluster	Out.South.Africa.3		
10	Namibia.2(1/14)	1	Donor	Single sample cluster	Out.Namibia.2		
11	Namibia.2(6/14)	6	Donor	No contribution	Out.Namibia.2		
12	Namibia.2(5/14)	5	Donor	No contribution	Out.Namibia.2		
13	South.Africa.3(10/19)	10	Surrogate (Merged)	Similar according to TVD and tree distance	*SouthAfrica*	*4*	
14	South.Africa.3(8/19)	8					
15	Gambia(3/111)+Sierra.Leone(1/69)	4	Donor	Similar to 18, small	Out.Gambia Out.SierraLeone		
16	Gambia(10/111)	10	Donor	Similar to 18, small	Out.Gambia		
17	Gambia(18/111)	18	Donor	Similar to 18, small	Out.Gambia		
18	Gambia(29/111)	29	Surrogate (Merged)	Similar according to TVD and tree distance	*WestAfrica1*	*5*	
19	Gambia(22/111)	22					
20	Gambia(29/111)*	29	Donor	Similar to 18, small	Out.Gambia		
21	Sierra.Leone(68/69)	68	Surrogate		*WestAfrica2*	*6*	

22	Nigeria.1(31/101)+Nigeria.2(1/95)	32	Surrogate (Merged)	Similar according to TVD and tree distance	*WestAfrica3* Out.Nigeria.1 Out.Nigeria.2	*7*	2 Nigeria.2 and 1 inconsistent ind excluded
23	Nigeria.1(69/101)+Nigeria.2(1/95)	70					
24	Nigeria.1(1/101)+Nigeria.2(93/95)	94	Donor	Similar to **23**	Out.Nigeria.1 Out.Nigeria.2		
25	Botswana(1/14)	1	Donor	Single sample cluster	Out.Botswana		
26	Botswana(1/14)*	1	Donor	Single sample cluster	Out.Botswana		
27	Botswana(1/14)**	1	Donor	Single sample cluster	Out.Botswana		
28	Botswana(3/14)	3	Donor	No contribution	Out.Botswana		
29	South.Africa.1(1/3)	1	Donor	Single sample cluster	Out.South.Africa.1		
30	South.Africa.2(1/4)	1	Donor	Single sample cluster	Out.South.Africa.2		
31	Botswana(3/14)*	3	Donor	No contribution	Out.Botswana		
32	Botswana(5/14)	5	Donor	No contribution	Out.Botswana		
33	South.Africa.1(2/3)+South.Africa.2 (3/4)	5	Donor	No contribution	Out.South.Africa.1 Out.South.Africa.2		
34	Angola(1/19)+Namibia.1(1/15)	2	Donor	No contribution	Out.Angola Out.Namibia.1		
35	Angola(8/19)	8	Donor	No contribution	Out.Angola		
36	Angola(10/19)+Namibia.2(1/14)	11	Donor	No contribution	Out.Angola Out.Namibia.2		
37	Namibia.1(14/15)	14	Donor	No contribution	Out.Namibia.1		
38	Jordan.1(1/15)	1	Donor	Single sample cluster	Out.Jordan.1		
39	Israel.1(2/2)+Jordan.1(2/15)	4	Donor	Similar to **41**	Out.Israel.1 Out.Jordan.1		
40	Jordan.1(2/15)	2	Donor	Small cluster, similar **41**	Out.Jordan.1		
41	Jordan.1(7/15)+Yemen(2/2)	9	Surrogate		*EastMediterranean1*	*8*	
42	Jordan.1(1/15)+Jordan.2(3/3)+Palestine(3/3)	7	Surrogate		*EastMediterranean2*	*9*	
43	Turkey.2(2/2)	2	Donor	Complex genetic profile	Out.Turkey.2		

44	Geor-gia(2/2)+Greece.1(1/2)+Greece.2(2/2)	5	Donor	Complex genetic profile	Out.Georgia Out.Greece.1 Out.Greece.2		
45	Iraq(2/2)+Israel.2(2/2)+Jordan.1(2/15)	6	Donor	Complex genetic profile	Out.Iraq, Out.Israel.2 Out.Jordan.1		
46	Morocco.2(7/7)	7	Surrogate		*Sephardic3*	*10*	
47	Libya.2(1/7)+Turkey.1(7/7)	8	Surrogate		*Sephardic1*	*11*	
48	Tunisia.2(4/6)	4	Surrogate (Merged)	Similar according to TVD and tree distance	*Sephardic2*	*12*	
49	Libya.2(6/7)+Tunisia.2(2/6)	8					
50	Libya.1(1/14)+Tunisia.1(2/14)	3	Surrogate (Merged)	Similar according to TVD and tree distance	*SouthMediterranean1*	*13*	
51	Libya.1(11/14)+Tunisia.1(3/14)	14					
52	Libya.1(2/14)+Tunisia.1(9/14)	11					
53	Morocco.1(3/11)	3	Surrogate (Merged)	Similar according to TVD and tree distance	*SouthMediterranean2*	*14*	
54	Morocco.1(8/11)	8					
55	Spain.4(2/15)	2	Donor	Similar to **56**, small	Out.Spain.4		
56	Spain.10(4/6)+Spain.11(5/8)+Spain.12(3/14)+Spain.14(1/7)+Spain.2(1/14)+Spain.4(13/15)+Spain.5(4/4)+Spain.6(4/8)+Spain.7(4/7)+Spain.9(5/12)	44	Surrogate		*CentralSouthSpain* Out.Spain.5 Out.Spain.10	*15*	2 inds excluded - inconsistent assignment
57	Spain.10(2/6)+Spain.12(6/14)+Spain.13(5/6)+Spain.17(6/6)+Spain.8(3/15)	22	Surrogate		*CentralNorthSpain* Out.Spain.8 Out.Spain.12 Out.Spain.17	*16*	4 inds excluded - inconsistent assignment
58	Spain.8(12/15)	12	Donor	Drifted, no contribution	Out.Spain.8		
59	Italy.2(3/3)	3	Donor	Complex genetic profile	Out.Italy.2		
60	Spain.1(2/8)+Spain.11(1/8)+Spain.12(5/14)+Spain.13(1/6)+Spain.14(6/7)+Spain.15(10/10)+Spain.16(7/8)+Spain.7(3/7)+Spain.9(1/12)	36	Surrogate		*Catalonia* Out.Spain.1 Out.Spain.11 Out.Spain.12	*17*	3 inds relocated to **56**, 4 inds excluded - inconsistent

61	Portu-gal.2(1/31)+Spain.11(2/8)+Spain.2(13/14)+Spain.3(2/2)+Spain.6(1/8)	19	Surrogate		*CanaryIslands*	*18*	1 ind relocated to **62**
62	Portugal.1(18/18)+Portu-gal.2(30/31)+Spain.1(6/8)+Spain.16(1/8)+Spain.6(3/8)+Spain.9(6/12)	64	Surrogate		*Portugal/WestSpain* Out.Spain.1 Out.Spain.9 Out.Spain.6 Out.Spain.16	*19*	3 inds relocated to **56**, 9 inds ex-cluded - incon-sistent assign-ment
63	France.1(1/2)+Spain.18(1/14)+Spain.19(8/8)	10	Surrogate (Merged)	Similar according to TVD and tree distance	*Basque*	*20*	
64	France.1(1/2)+Spain.18(13/14)	14					
65	Bulgaria(2/2)+Greece.1(1/2)+Italy.1(2/2)+Italy.5(15/15)	20	Surrogate		*Italy1* Out.Greece.1	*21*	1 Greece.1 ind removed
66	Italy.3(31/106)	31	Surrogate		*Italy2*	*22*	
67	Italy.3(75/106)+Italy.4(2/2)	77	Donor	No contribution	Out.Italy.3,Out.Italy.4		
68	UK.2(28/29)	28	Donor	Similar to **69**, no contrib.	Out.UK.2		
69	France.2(1/3)+NW.Europe(74/91)+UK.1(31/31)+UK.2(1/29)+UK.3(1/1)+UK.4(3/3)	111	Surrogate		*NorthWestEurope2* Out.UK.4 Out.NW.Europe Out.France.2	*23*	10 inds excluded - inconsistent as-signment
70	Germany(6/37)+Hun-gary(1/2)+NW.Europe(10/91)	17	Donor	Similar to **69**, small con-tribution to CANDELA	Out.Germany Out.Hungary Out.NW.Europe		
71	UK.5(2/2)+UK.6(21/21)	23	Donor	No contribution	Out.UK.5, Out.UK.6		
72	France.2(2/3)+Germany(31/37)+Hun-gary(1/2)+NW.Europe(7/91)	41	Surrogate		*NorthWestEurope1* Out.Hungary Out.France.2 Out.NW.Europe	*24*	Non-Germany in-divudals removed
73	Russia(2/2)	2	Surrogate		*NorthEastEurope1*	*25*	
74	Estonia(2/2)+Finland(7/99)	9	Surrogate		*NorthEastEurope2*	*26*	
75	Finland(29/99)	29	Surrogate		*NorthEastEurope3*	*27*	

76	Finland(41/99)	41	(Merged)	Similar according to TVD and tree distance			
77	Finland(22/99)	22					
78	China.1(2/82)	2	Donor	Similar to **79**, small	Out.China.1		
79	China.1(72/82)	72	Surrogate		*China/Vietnam1*	*28*	
80	China.1(7/82)	7	Donor	Similar to **79**, small	Out.China.1		
81	Vietnam(91/95)	91	Surrogate		*China/Vietnam2*	*29*	
82	Japan(1/104)+Korea(1/2)	2	Donor	Samples represented by different clusters	Out.Japan Out.Korea		
83	China.3(1/31)+China.4(64/101)+Korea(1/2)	66	Surrogate		*ChinaHan* Out.China.3 Out.Korea	*30*	2 non China.4 inds removed
84	China.2(26/66)+China.3(2/31)+China.4(3/101)+Vietnam(4/95)	35	Donor	Similar to **84**, complex genetic background	Out.China.2 Out.China.3 Out.China.4 Out.Vietnam		
85	China.1(1/82)+China.2(40/66)+China.3(2/31)+China.4(29/101)	72	Donor	Contains several populations present in other clusters	Out.China.1 Out.China.2 Out.China.3, Out.China.4		
86	China.3(26/31)+China.4(5/101)	31	Donor	No contribution to CANDELA	Out.CHB Out.CHS.Fu.Jian		
87	Japan(72/104)	72	Surrogate (Merged)		*Japan*	*31*	
88	Japan(31/104)	31					
89	Chile.3(2/65)	2	Donor	Similar to **90**, small	Out.Chile.3		
90	Bolivia.2(6/12)+Chile.1(1/3)+Chile.3(27/65)	34	Surrogate (Merged)		*Quechua2* Out.Chile.3	*32*	3 inds excluded - inconsistent assignment
91	Bolivia.2(2/12)+Chile.3(23/65)	25					
92	Peru.3(5/5)	5	Removed	Removed as donor and recipient, because high drift			

93	Bolivia.1(10/12)+Chile.1(1/3)+Chile.3(1/65)+Peru.2(2/3)+Peru.4(6/17)	20	Surrogate		*Aymara*, Out.Chile.1, Out.Chile.3, Out.Bolivia.1	*33*	4 inds excluded - inconsistent assignment
94	Bolivia.1(2/12)+Bolivia.2(4/12)+Chile.1(1/3)+Chile.3(10/66)+Peru.4(1/17)	18	Donor	Whole cluster has inconsistent assignment	Out.Bolivia.1 Out.Bolivia.2 Out.Chile.1 Out.Chile.3 Out.Peru.4		
95	Argentina.1(10/19)+Chile.3(1/67)	11	Surrogate		*Colla*, Out.Chile.3	*34*	Chile.3 removed
96	Argentina.1(9/19)	9	Donor	Similar to **95**, no contrib.	Out.Argentina.1		
97	Peru.2(1/3)+Peru.4(8/17)	9	Surrogate		*Quechua1*	*35*	
98	Colombia.1(2/16)+Colombia.2(1/3)	3	Surrogate		*ChibchaPaez3*	*36*	
99	Costa.Rica.2(3/3)	3	Surrogate		*ChibchaPaez2*	*37*	
100	Costa.Rica.1(4/4)	4	Surrogate		*ChibchaPaez1*	*38*	
101	Colombia.5(4/4)	4	Surrogate		*ChibchaPaez5*	*39*	
102	Colombia.3(2/2)	2	Surrogate		*ChibchaPaez6*	*40*	
103	Colombia.4(4/4)	4	Removed	Removed as donor and recipient, because high drift			
104	Colombia.1(2/16)	2	Donor	Similar to **105**, drifted	Out.Colombia.1		
105	Colombia.1(11/16)	11	Surrogate, Merged with 106		*ChibchaPaez4*	*41*	
106	Colombia.1(1/16)+Colombia.2(2/3)	3	Surrogate, Merged with 105		*ChibchaPaez4*	*41*	
107	Peru.1(1/13)+Peru.4(2/16)	3	Surrogate		*AndesPiedmont*	*42*	
108	Argentina.2(2/2)+Chile.2(2/2)+Chile.3(1/65)	5	Surrogate		*Mapuche*	*43*	
109	Guatemala(5/5)+Mexico.9(2/2)	7	Surrogate		*Mayan*	*44*	
110	Brazil.1(1/3)	1	Removed	Single sample cluster, removed as donor and recipient, because high drift			
111	Brazil.1(2/3)	2	Removed	Removed as donor and recipient, because high drift			
112	Brazil.2(2/2)	2	Removed	Removed as donor and recipient, because high drift			
113	Paraguay(4/4)	4	Surrogate		*Amazon3*	*45*	
114	Colombia.7(3/3)	3	Removed	Removed as donor and recipient, because high drift			

115	Colombia.6(2/2)	2	Surrogate		*Amazon1*	*46*	
116	Peru.1(6/13)	6	Donor	Similar to **117**, no contrib	Out.Peru1	NA	
117	Peru.1(6/13)*	6	Surrogate		*Amazon2*	*47*	
118	Argentina.6(3/5)+Argentina.7(2/2)	5	Surrogate		*Chaco1*	*48*	
119	Argentina.6(2/5)	2	Donor	Similar to **118**, no contrib	Out.Argentina.6	NA	
120	Mexico.1(2/2)	2	Surrogate		*Pima*	*49*	
121	Mexico.10(8/22)+Mexico.2(2/20)	10	Surrogate		*Nahua1* Out.Mexico.10	*50*	1 ind excluded - inconsistent
122	Mexico.6(7/8)	7	Surrogate		*SouthMexico3*	*51*	
123	Mexico.8(6/8)	6	Surrogate		*SouthMexico2*	*52*	
124	Mexico.10(13/22)+Mexico.6(1/8)+Mexico.8(2/8)	16	Surrogate		*SouthMexico1*	*53*	
125	Mexico.10(1/22)+Mexico.2(18/20)	19	Surrogate		*Nahua2*	*54*	
126	Mexico.3(2/2)+Mexico.4(16/16)	18	Surrogate + Remove	Highly drifted population excluded (Mexico.4)	*Mixe* (Only Mexico.3)	*55*	
127	Argentina.3(1/13)+Argentina.5(3/3)	4	Surrogate (Merged)	Similar according to TVD and tree distance	*Chaco2*	*56*	
128	Argentina.3(5/13)	5					
129	Argentina.3(7/13)+Argentina.4(2/2)	9					

fS Clust: Cluster assigned by fineSTRUCTURE
Decision: Some references samples were used only as “Donors” for the subsequent sub-continental ancestry inference. Some are also used as “Surrogates” for the ancestral populations in SOURCEFIND and NNLS analyses. Some were “Removed” from the reference set.
Donor/*Surrogate*: This is the final grouping used for generating the “copying vectors” used for the sub-continental ancestry analysis. Groups in *Italics* are the ones that were selected as surrogates and are further described in Chapter 3 (Table 3.4).