THE UNIVERSITY OF CHICAGO

THE GENETIC HISTORY AND ADAPTATIONS OF HIGH ALTITUDE EAST ASIANS IN THE TIBETAN PLATEAU

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF HUMAN GENETICS

BY

CHOONGWON JEONG

CHICAGO, ILLINOIS

AUGUST 2016

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SUPPLEMENTARY FIGURES

LIST OF SUPPLEMENTARY TABLES

LIST OF SUPPLEMENTARY TEXT

ACKNOWLEDGEMENTS

First, I would like to thank my advisor, Anna Di Rienzo, for her wonderful guidance and support of my thesis work. Throughout my thesis work, I have been fortunate enough to enjoy her thoughtful mentoring, unlimited curiosity, and rigorous scientific reasoning. It has been a truly amazing experience to have such a positive role model as a young scientist. I also would like to thank the members of my thesis committee, Carole Ober, John Novembre and Matthew Stephens, for their helpful discussions and insights on pushing forward my thesis project, as well as their consistent support and encouragement throughout my time in Chicago. Current and former members of the Di Rienzo lab and the department of Human Genetics have given me invaluable support for my scientific training as well as dearest friendship. Especially, it has been my great pleasure to work together with long-term lab mates, Silvia Kariuki and Shigeki Nakagome.

I have been unbelievably fortunate to be able to work with amazing collaborators outside the University of Chicago. Dr. Cynthia Beall first introduced me to the field of high altitude adaptations and provided an essential link between our genetics laboratory and study participants far away in the remote regions of the world. Drs. Mark Aldenderfer and Christina Warinner made a perfect team for conducting an ancient DNA project.

Finally, I would like to dedicate this dissertation to my family. They have shown unchanging support and patience for my pursuit of scientific career. My wife, Heesun Choi, so much enriched my PhD student life in Chicago that I dare not to look back how it used to be before I married.

ABSTRACT

In this dissertation, I present population genetic studies of indigenous people of the Tibetan plateau, collectively called “Tibetans”. In Chapter 2, I describe population structure of Tibetans and a demographic model explaining it, by analyzing genome-wide variation data of Tibetan populations, including Sherpa people from Nepal, in conjunction with global populations. Specifically, I provide evidence that Tibetan populations are structured due to varying degree of admixture with lowland East Asian gene pool. Among the sampled Tibetan cohorts, the Sherpa are the least, if any, affected by this admixture, and thus are the best representatives of this “high altitude” East Asian gene pool. Assuming this high altitude ancestry contains variants adaptive to the high altitude, I prioritize genetic loci with high proportion of this ancestry in Tibetans as candidates of adaptively evolving genes. In Chapter 3, I investigate the genetic profile of eight ancient individuals from the Himalayas with their whole genome sequences. I show that all of these individuals, ranging from 2,500 to 1,300 years old, are most closely related to contemporary Tibetans, thus making them the first whole ancient genomes of genetic East Asians. The remarkable stability of genetic composition in this region across time, in contrast to diverse cultural connections with various parts of the outside world, highlights the role of high altitudes as a strong barrier to gene flow. I propose that topographic differences made human migration from East Asia to the Tibetan plateau easier than that from either South or Central Asia. Especially, areas with intermediate altitude in the gradually ascending topography may have provided a base camp for ancestral Tibetans to accumulate genetic and cultural adaptations to cope with the high altitude. I also report that the adaptive Tibetan haplotype in the *EPAS1* gene seems to have increased in frequency much later than the *EGLN1* haplotype in this region.

Considering evidence of early Holocene or late Upper Pleistocene divergence of the Tibetan gene pool from that of lowlanders, it is possible that there may be complicating factors such as population structure, multiple selective pressures or epistatic fitness effect in the evolution of these variants. Lastly, in Chapter 4, I explore the genetic history of the Ainu, an indigenous hunter-gatherer group in northern Japan, inspired by the long-held hypothesis of their sharing of ancestry with Tibetans based on Y haplogroup D-M174. I find that the Ainu represent an even deeper branch of East Asian ancestry, rather than sharing a common ancestry with Tibetans. Also, I find evidence for additional gene flow between the Ainu and lowland East Asians. The Ainu are unique among East Asians in being genetically closer to northeast Siberians than to Central Siberians. This may be due to either gene flow or common ancestry, the latter of which raises the possibility that the Ainu represent an early branch of the first East Asian migrants into Siberia, who eventually became ancestors of Native Americans. Ancient genomes from northeast Asia and Siberia will be invaluable to disentangle the evolutionary history of eastern Eurasia in high resolution. Together, I expect that additional population genomic studies using both contemporary and ancient samples will reveal our evolutionary past in East Asia and help us understand how our ancestors could adapt to such a diverse array of environments, ranging from tropical rainforests to freezing dark winter in the Arctic Circle.

CHAPTER 1: INTRODUCTION

Genetic adaptations to local environments in humans

Following their initial movement out of Africa around 50-100 thousand years ago (kya), anatomically modern humans (AMH) colonized the rest of the world at a remarkable pace (Henn et al. 2012; Scally and Durbin 2012). Remains of AMH, such as skeletons or stone tools, are found in Europe, Southeast Asia, Australasia and arctic Siberia as early as 42-49 kya (Summerhayes et al. 2010; Demeter et al. 2012; Hublin 2012; Pitulko et al. 2016). Such a rapid range expansion becomes even more surprising when we consider accompanying drastic changes in the environments. For example, in the Siberian Arctic Circle with long and dark winter, well-established archaeological sites date back to 27 kya (Pitulko et al. 2004) and there are evidences of human activity dating back to 45 kya (Pitulko et al. 2016). Even in the Tibetan plateau, so called "the third pole" of the world, people have occupied this rugged terrain for at least over the last 15,000 years (Aldenderfer 2011; Brantingham et al. 2013). Human activities also induced major changes in human environments. For example, a switch from hunting and gathering to farming accompanied qualitative changes in diets, pathogen loads, and social structures (Mira et al. 2006; Eshed et al. 2010). Archaeological records suggest a substantial health impact of these changes, such as increase in prevalence of infectious diseases and decrease in health status (Cohen and Armelagos 1984; Eshed et al. 2010; Mummert et al. 2011). Such new environments, extremely divergent from thse of sub-Saharan Africa, presumably posed strong selective pressures, which in turn caused biological and cultural responses in the form of new or shifted phenotypes adaptive to such local environments. Indeed, phenotypes with unusual differentiation among populations, such as skin pigmentation (Jablonski and Chaplin 2000, 2010) and lactase

persistence (Swallow 2003), have attracted much attention from geneticists as potential outcomes of such a process.

Adaptations to specific local features in heterogeneous environments, or local adaptations, have been widely observed and studied by evolutionary biologists and ecologists for decades (Savolainen et al. 2013). Many studies focused on the role of local adaptations as a force for population differentiation and speciation (Levene 1953; Gavrilets 2003). That is, as locally adaptive features accumulate in an indigenous population, it becomes increasingly difficult for migrants without such adaptations to compete with the natives, and therefore gene flow between heterogeneous environments may decrease over time (Blanquart et al. 2013). In ecology, the "common garden" experiment is a standard process to test local adaptations, in which organisms are set up in a series of environments, both native and foreign, and fitness outcome is compared between groups in each environment (Kawecki and Ebert 2004; Blanquart et al. 2013). If each population or species outperforms foreign ones in its indigenous environment, it is interpreted as an evidence of local adaptation (Kawecki and Ebert 2004; Blanquart et al. 2013; Savolainen et al. 2013). Local adaptations can occur in widely varying spatial scales, ranging from differing climate zones (Huey et al. 2000) to patchy soils in a landscape (Wright et al. 2013), and to different species of host plants growing side by side (Imo et al. 2013).

In most cases, neither environmental heterogeneity nor local adaptation functions an absolute barrier to gene flow. Occasionally, a new variant introduced by gene flow into a new environment confers a fitness advantage to its bearers and by doing so quickly spreads into the target population. This phenomenon, called "adaptive introgression" especially if it transgresses a species boundary, has been recently highlighted as a new mode of introducing adaptive variants into a population (Hedrick 2013). One may commonly assume that the introgressed

genetic variant is adaptive in its native environment, but this is not a necessary condition for introgression: i.e. a variant may exert a fitness advantage in its new environment while its evolution in the original environment is either neutral or even under purifying selection.

Although phenomena of local adaptation and adaptive introgression have been studied for decades, it used to be extremely difficult to identify the underlying genetic variants and selective pressures. Recent advances in genomics now provide powerful ways to tackle these questions. Especially, high density data of genetic variation across whole genomes are leveraged to find genomic regions showing either a statistical association with candidate adaptive phenotypes or an unusual pattern of genetic variation deviant from neutral expectations (Savolainen et al. 2013). The former, a phenotype-centered approach, is represented by genome-wide association studies (GWAS). GWAS, by investigating hundreds of thousands of genetic markers along the genome for their statistical association with the adaptive phenotype, provide a comprehensive search for genetic loci affecting a phenotype. Over the years, GWAS have proven to be extremely fruitful, currently providing more than 21,000 genotype-phenotype associations for a wide range of traits (Welter et al. 2014), at least a portion of which are likely to have had significant fitness consequences in our evolutionary past. However, GWAS require *a priori* choice of relevant phenotypes and a large sample size, typically in the thousands, to detect genetic associations with small effect size (Lettre et al. 2007; Bush and Moore 2012). In contrast, many phenotype-agnostic methods can be applied to a cohort of moderate size (tens or hundreds) without phenotype information (Voight et al. 2006; Sabeti et al. 2007; Pickrell et al. 2009; Yi et al. 2010). Ideally, this group of methods detects genes evolving under positive natural selection in a population, evidenced by either unusually fast allele frequency differentiation (Weir and Cockerham 1984; Yi et al. 2010) or extended linkage disequilibrium (LD) around a focal variant

(Voight et al. 2006; Sabeti et al. 2007), which in turn may point to the underlying biological functions and, in some cases, the environmental factors. For example, multiple genetic variants around the *LCT* (lactase) gene were found to be under strong positive selection in populations with dairy culture in Europe, Africa and Middle East (Enattah et al. 2002; Ingram et al. 2007; Tishkoff et al. 2007; Enattah et al. 2008). However, it is often difficult to connect gene sets discovered to specific biological functions or selection pressures. Also, it now seems like that strong positive selection on *de novo* mutations is a rare event in human genetic history, and other types of selections are harder to distinguish from neutrally evolving loci (Pritchard and Di Rienzo 2010; Pritchard et al. 2010; Hernandez et al. 2011; Field et al. 2016). Recently, new methods for testing natural selection on polygenic phenotypes have been developed (Daub et al. 2013; Berg and Coop 2014). These methods work by integrating relatively weak information, such as magnitude and direction of allele frequency differentiation, across ascertained sets of variants or genes. Therefore, such methods are expected to be used more frequently as both genetic association and population genetic resources accumulate in a population of interest.

High altitude adaptations in indigenous human populations

High altitudes, typically defined as 2,500 m.a.s.l. (meters above sea level), are characterized by a multitude of environmental factors challenging to human subsistence, such as low oxygen supply due to reduced barometric pressure (hypobaric hypoxia), high ultraviolet (UV) radiation, limited biological resources and low productivity (Aldenderfer 2006; Barton 2016). Among them, hypobaric hypoxia has attracted a lot of attention as one of the most important selection pressures acting in high altitude environments due to several reasons. First, the oxygen delivery system is an essential part of physiology in human and other complex animals, which

obligatorily depend on oxygen supply for their survival. Indeed, most of animal phyla share a conserved network of genes, the HIF (hypoxia inducible factor) pathway, used for sensing and responding to oxygen availability, controlled by oxygen sensing proteins called PHDs (prolyl hydroxylase domain-containing proteins) and by transcription factors called HIFs (Majmundar et al. 2010; Taylor and McElwain 2010; Semenza 2012). The molecular functions of these proteins have been studied intensively in the context of a role of hypoxia in cancer, inflammation or development (Kaelin Jr and Ratcliffe 2008; Semenza 2012). Second, unlike most of other environmental factors, hypoxia is experienced by all people in the same way irrelevant to sex, age, wealth or cultural context, because oxygen partial pressure varies only as a function of altitude (Beall 2007). This implies that biological responses may have disproportionately contributed to altitude adaptations, in contrast to adaptations to other harsh environments, such as high latitude with extreme coldness, where cultural and behavioral changes are likely to have played a substantial role (Goebel 2002; Osborn 2014). Therefore, genetic study of human adaptations to high altitudes has a potential for enhancing our understanding of how this ancient biological pathway operates in human physiology.

Another interesting feature of human high altitude adaptations is that several populations from different ethnic backgrounds independently colonized high elevations, including the Tibetan plateau, the Andean altiplano, and the Ethiopian highlands (Beall 2006). Surprisingly, indigenous populations of these high altitude regions show phenotypes distinct from their lowland neighbors as well as from each other (Beall 2006, 2007). When placed into high altitudes, human physiology deploys a battery of acclimatization responses as a response to hypoxia, including acceleration of erythropoiesis and increase in respiration and basal metabolism (Beall 2007). As one stays longer at altitude, basal metabolic rate and respiration

gradually go down while red blood cell counts and hemoglobin concentration stay high (Beall 2007). Aymara and Quechua, Native American populations indigenous to the Andean altiplano, were the first indigenous group of high altitudes whose biology was studied for altitude adaptations. Over the past century, morphological and physiological studies of high altitude populations showed that many of their phenotypes are similar to responses shown in lowland sojourners, such as increased hemoglobin level and lowered arterial blood oxygen saturation (the proportion of hemoglobin molecules associated with oxygen molecules in arterial blood). However, they also show unique phenotypes such as thick and broad chest morphology (Beall 1982; Stinson 1985). Surprisingly, studies of East Asian highlanders in the Tibetan plateau and the Himalayas revealed a totally different set of physiological traits in Tibetans. For example, Tibetans show no increase in their blood hemoglobin level up to 4,000 m.a.s.l., and have much lower hemoglobin level compared to that of Han Chinese or Andean highlanders even above 4,000 m altitude, although hemoglobin level do increase in that extreme range of altitude (Beall and Reichsman 1984; Beall and Goldstein 1987; Beall et al. 1998). Arterial oxygen saturation level is slightly lower in Tibetans than in Andeans (Beall et al. 1999), although no protein-coding variation was found in hemoglobin genes of either population. Combining two features, Tibetans have arterial oxygen content (the amount of oxygen carried by per volume of blood) substantially lower than the sea level mean, while Andeans actually have a value higher than that expected for their high hemoglobin level (Beall 2007). Additional Tibetan-specific phenotypes, such as high level of exhaled nitric oxide (Beall et al. 2001) and elevated resting ventilation and hypoxic ventilatory response (Beall et al. 1997), seem to contribute Tibetan adaptations to high altitude, together with other yet-to-be-discovered adaptive phenotypes. Studies of Amhara people in the Ethiopian highlands add further complications: they seem to show hemoglobin and oxygen

saturation levels indistinguishable from those of European Americans at the sea level, even though they reside at 3,530 m.a.s.l. (Beall et al. 2002). Therefore, they share with Tibetans the hemoglobin phenotype, but not the oxygen saturation one.

Population genomic studies of indigenous high altitude populations mostly adopted a phenotype-independent approach, focusing on discovering loci under recent positive selection (Beall et al. 2010; Bigham et al. 2010; Simonson et al. 2010; Yi et al. 2010; Peng, Yang, et al. 2011; Wang et al. 2011; Xu et al. 2011; Scheinfeldt et al. 2012; Huerta-Sánchez et al. 2013), with an exception of a single study on Ethiopians which adopted both genomic scans of positive selection and GWAS on multiple phenotypes with a moderate sample size (Alkorta-Aranburu et al. 2012). Tibetans have been most highlighted in population genomics, mainly due to the initial findings of two genes, *EGLN1* (egl nine homolog 1) and *EPAS1* (endothelial PAS domain protein 1), which harbor strong signatures of positive selection as well as statistical association with hemoglobin level (Beall et al. 2010; Simonson et al. 2010; Yi et al. 2010; Xiang et al. 2013). These genes also are essential components of the HIF pathway: the *EGLN1* gene encodes one of three PHD proteins (PHD2) and the *EPAS1* gene encodes HIF2α protein, which constitute alpha subunit of HIF2 transcription factor (Kaelin Jr and Ratcliffe 2008). Under normoxic condition, PHD2 hydroxylates a proline residue of HIF2α, which then induces ubiquitination and breakdown (Kaelin Jr and Ratcliffe 2008). Also, it is well known that the *EPO* (erythropoietin) gene, a master regulator of red blood cell generation, is a target of the HIF1 and HIF2 transcription factors in kidney and liver, respectively (Yoon et al. 2011). Interestingly, Tibetan haplotypes in the *EGLN1* and *EPAS1* genes were reported to be associated with lower hemoglobin level (Simonson et al. 2010; Yi et al. 2010; Xiang et al. 2013). Functional studies of two non-synonymous *EGLN1* SNPs (single nucleotide polymorphisms), D4E (rs186996510) and

C127S (rs12097901), reported a blunted response to low oxygen in double mutants, although details of the biochemical mechanism is still unclear (Lorenzo et al. 2014; Petousi et al. 2014). Together with no increase of hemoglobin level to altitudes up to 4,000 m in Tibetans, these results seem to suggest that the increase of hemoglobin level at altitudes due to acclimatization may be maladaptive. High incidence of chronic mountain sickness (CMS), defined by extremely high concentration of hemoglobin, is observed in Andean highlanders (5%), while Tibetans and Ethiopian Amhara, who both have unelevated hemoglobin level up to 3,500 to 4,000 m.a.s.l., show much lower (1%) or negligible level of CMS prevalence, respectively (León-Velarde et al. 2014). Also, many Andean women suffer from preeclampsia during their pregnancies, while it affects Tibetans to a much lesser degree (Niermeyer 2014). Such health concerns suggest that increased hemoglobin level in a long term or during particularly susceptible periods of life are detrimental to human physiology.

Understanding genetic history of Tibetans and their altitude adaptations

The genetic history of Tibetans has long attracted population geneticists, even prior to the recent genomic studies of Tibetan altitude adaptations. Especially, Tibetans have a high proportion of Y haplogroup D-M174 which has coalescence time estimates of 50 kya or older (Shi et al. 2008; Qi et al. 2013). The geographic distribution of this haplogroup is also unusual: it is common in Tibetans, in Japanese, especially in the Ainu, an indigenous population of the northern part of Japan who are thought to be descendants of prehistoric Japanese hunter-gatherers with Jomon culture (Hammer et al. 2006), and in Andamanese islanders (Thangaraj et al. 2005). Otherwise, it is practically absent in Eurasia (Chiaroni et al. 2009). Such a patchy geographic distribution inspired an idea that Tibetans, together with ancient Jomon people in

Japan, may represent the first wave of AMH migrants into East Asia, who were largely replaced by the following waves of migrants in the other regions (Stoneking and Delfin 2010). However, hypotheses on the genetic history of Tibetans were rarely tested using genome-wide variation data. One study proposed a very recent split of 2,750 years ago (ya) between Tibetans and Han Chinese, based on site frequency spectrum data obtained from exome sequencing of 50 Tibetans (Yi et al. 2010). However, Neolithic agricultural societies were widespread in the Tibetan plateau by at least 3,600 ya, evidenced by many well-dated archaeological remains (Guedes et al. 2014; Chen, Dong, et al. 2015). Also, early Holocene or even late Upper Paleolithic sites found in the Tibetan plateau suggest an even earlier presence of hunter-gatherer populations (Aldenderfer 2011), although it is an open question if they contributed to the gene pool of contemporary Tibetans. Another interesting finding is that the Tibetan *EPAS1* haplotype with adaptive signatures embeds a 50-kb region introgressed from an archaic hominin group related to Denisovans (Huerta-Sánchez et al. 2014). It is not known if Denisovans lived or adapted to high altitudes. Also, it is unclear how this haplotype was transferred from an archaic hominin group to ancestors of contemporary Tibetans.

The focus of this dissertation study was to improve our understanding of the genetic history of Tibetans, who are underrepresented in large-scale genomic resources of human diversity, by generating and analyzing population-scale genome-wide variation data. Understanding population structure not only has been a central focus of population genetics for decades (Rosenberg et al. 2002), but also has been shown to be an important factor to account for in GWAS (Marchini et al. 2004; Price et al. 2006). Also, a proper model of population history is a key to interpret signatures of adaptive evolution (Jensen et al. 2016). For these reasons, I explored the following three specific questions in this dissertation study.

In my first project described in Chapter 2, I show that varying degree of gene flow with lowland East Asians is a key force of genetic heterogeneity among Tibetans. For this, I focused on the Sherpa population from the high altitude Himalayan arc in Solukhumbu district, Nepal. Sherpa, meaning "eastern people" in Tibetan, are an ethnic group well known to Westerners for their superb performance in mountaineering, with their historical records suggesting a migration from eastern Tibet ("Kham") around 4-6 centuries ago to avoid political disturbance (Oppitz 1974). They also share similar cultural aspects with other Tibetan groups and speak a Tibetic language. Finally, they share the Tibetan pattern of physiological adaptations to high altitudes (Gilbert-Kawai et al. 2014). Therefore, it is natural to classify the Sherpa as a subset of the broad ethnic Tibetan group of populations. By jointly analyzing Sherpa and other Tibetan genotype data from multiple locations in the Tibetan plateau, I found that Tibetan populations form a genetic cline due to admixture with lowland East Asians: Sherpa were the key of this finding because they were at the end of this cline among the sampled cohorts, meaning that other Tibetan cohorts are genetically closer to lowland East Asians than Sherpa are to them. Therefore, I conclude that Sherpa are the best representatives of the ancestral Tibetan gene pool among the contemporary Tibetans, although it is possible that Sherpa have substantial amount of gene flow from lowlanders. When changes in the effective population size (N_e) across time was inferred from whole genome sequence data, Sherpa showed an early divergence from lowland East Asians, such as Han Chinese or Dai, going back as early as 20-40 kya. Combining this with evidences from uniparental markers, I propose that Tibetans split from the rest of East Asians in the late Pleistocene or early Holocene, implying a role of climate changes during the ice age. Lastly, I tried to leverage admixture in Tibetans for finding genes conferring fitness advantages in high altitudes. In short, if the ancestral high altitude gene pool accumulated many adaptive

variants across genome, having an allele of high altitude ancestry in these loci would have been advantageous, while alleles from both ancestries were equally likely to propagate to the offspring at neutral loci. Therefore, positive selection on adaptive variants following admixture may result in excess high altitude ancestry at these loci, when compared to neutrally evolving ones. By performing local ancestry deconvolution with Sherpa and lowland East Asians as representatives of the ancestral high and low altitude gene pools, respectively, I prioritized several loci with high level of Sherpa ancestry, including *EGLN1* and *EPAS1*, as candidates of adaptive evolution in Tibetans.

Reconstruction of genomes of ancient origin, by sequencing DNA (deoxyribonucleic acid) molecules extracted from biological remains of individuals living in the past, is a powerful technique to observe a snapshot of our evolutionary past. For the last few years, ancient DNA (aDNA) studies revolutionized our understanding of the evolutionary history of AMH (Rasmussen et al. 2010; Rasmussen et al. 2011; Keller et al. 2012; Skoglund et al. 2012; Fu et al. 2014; Lazaridis et al. 2014; Raghavan et al. 2014), as well as that of archaic hominins such as Neandertals and Denisovans (Green et al. 2010; Reich et al. 2010; Meyer et al. 2012; Prüfer et al. 2014). In Chapter 3, I studied ancient genomes from the high traverse valleys of the Himalayas in the Annapurna Conservation Area (ACA) in Central Nepal, dating from 2,500 to 1,300 years before present (BP). This harsh environment was one of the last places colonized by modern humans, which historically provided a vital link between Tibet and South Asia, two regions of different genetic, linguistic and cultural profiles (Gayden et al. 2013). Therefore, there have been many hypotheses about who migrated into this region, when and from where (Singh 1999; Alt et al. 2003; Peng, Palanichamy, et al. 2011; Aldenderfer 2013). By generating and analyzing whole genome sequences of eight ancient individuals from this region, I found that populations of this

region have been of genetic East Asian origin, particularly a high altitude East Asian one, from the first well-established permanent settlement. Therefore, it is likely that genetic and cultural adaptations to high altitudes provided a huge advantage in the colonization of this region from the plateau, while gene flow from the lowlands was mostly limited. This makes a sharp contrast with dense cultural connections between this region and the world outside, evidenced by rich grave goods of East, South, Central or West Asian origin (Knörzer 2000; Alt et al. 2003). Another interesting finding was the evolutionary trajectories of two adaptive haplotypes in the *EGLN1* and *EPAS1* genes. While the derived alleles of the *EGLN1* non-synonymous variant D4E (rs186996510) were observed in all samples, derived alleles of the *EPAS1* SNPs appeared only in the most recent samples, dating 1,300 BP. Therefore, this study suggests that the Tibetan gene pool was already on its way of divergence from the lowland East Asians at least 2,500 BP and that natural selection on the two well-known adaptive variants occurred at different times in this region.

The Japanese archipelago has a long history of human settlement, showing one of the oldest potteries in the world used by a sedentary "Jomon" culture of hunter-gatherers going back to at least 16,500 BP. Genetic data clearly support a "dual origin" of the contemporary Japanese, which means that modern Japanese are admixed descendants of indigenous Jomon hunter-gatherers and paddy-field rice farmers ("Yayoi" people) who began to migrate from the mainland East Asia into Japan around 2,300 BP (Hanihara 1991; Habu 2004; Jinam et al. 2012; Nakagome et al. 2015). The Ainu people are an indigenous hunter-gatherer group of Hokkaido and Sakhalin, northern islands of the Japanese archipelago and Russian Far East, who are thought to be a direct descendant of Jomon hunter-gatherers (Hanihara 1991; Habu 2004). While the admixture process and the resulting population structure of modern Japanese are well

understood in genetics and archaeology, the origin of Jomon and Yayoi people and their relationship with surrounding East Asian populations remain unclear. In Chapter 4, I investigated the genetic relationship of the Ainu with populations outside the archipelago to understand the origin of Jomon hunter-gatherers. This work was partially inspired by the long-held hypothesis of a genetic connection between Tibetans and the Ainu, based on their sharing of Y haplogroup D-M174 (Hammer et al. 2006). Instead of finding evidence for shared ancestry between the Ainu and Tibetans, I discovered that the Ainu split first from the rest of East Asians including Tibetans, thus forming an outgroup to all East Asian farmer groups. The Ainu-related genetic ancestry is present in substantial proportions in nearby populations of northeast Asia, suggesting a possibility that Ainu-related populations were once distributed in the mainland East Asia. Interestingly, the Ainu are genetically closer to northeast Siberians than to central Siberians, contrary to all other East Asians, who show the opposite pattern. I consider both post-divergence gene flow around the sea of Okhotsk and shared deep ancestry between the Ainu and northeast Siberians as a potential explanation. If the latter is true, the Ainu (and the prehistoric Jomon people) may represent a branch split off from the earliest stream of migrants into Siberia who eventually became ancestors of contemporary Native Americans.

CHAPTER 2: ADMIXTURE FACILITATES GENETIC ADAPTATIONS TO HIGH ALTITUDE IN TIBET[1]

2.1: Abstract

Admixture is recognized as a widespread feature of human populations, renewing interest in the possibility that genetic exchange can facilitate adaptations to new environments. Studies of Tibetans revealed candidates for high-altitude adaptations in the *EGLN1* and *EPAS1* genes, associated with lower hemoglobin concentration. However, the history of these variants or that of Tibetans remains poorly understood. Here, we analyze genotype data for the Nepalese Sherpa, and find that Tibetans are a mixture of ancestral populations related to the Sherpa and Han Chinese. *EGLN1* and *EPAS1* genes show a striking enrichment of high-altitude ancestry in the Tibetan genome, indicating that migrants from low altitude acquired adaptive alleles from the highlanders. Accordingly, the Sherpa and Tibetans share adaptive hemoglobin traits. This admixture-mediated adaptation shares important features with adaptive introgression. Therefore, we identify a novel mechanism, beyond selection on new mutations or on standing variation, through which populations can adapt to local environments.

[1] Citation for chapter: Jeong C et al. 2014. "Admixture facilitates genetic adaptations to high altitude in Tibet." Nat Commun 5:3281.

2.2: Introduction

The environments and indigenous populations of high-altitude (≥ 2,500 m in altitude) are an ideal study system for understanding the genetic basis of adaptive traits (Storz et al. 2010). Low barometric pressure and consequent physiological hypoxia constitute a strong selective pressure (Beall et al. 2004; Beall et al. 2010; Simonson et al. 2010; Yi et al. 2010), which is unavoidable and invariant across individuals at a given altitude because it cannot be influenced by behavioral or cultural practices (Storz et al. 2010). A distinctive set of physiological traits found in Tibetan highlanders, including unelevated hemoglobin concentrations up to 4,000 m altitude, are clearly linked to O_2 delivery (Beall et al. 2010). In Tibetans, variants in the *EGLN1* (egl nine homolog 1) and *EPAS1* (endothelial PAS-domain containing protein 1) genes harbor signals of adaptive allele frequency divergence relative to low-altitude East Asian populations as well as association signals with hemoglobin concentration (Beall et al. 2010; Simonson et al. 2010; Yi et al. 2010). These genes are major components of the HIF (hypoxia inducible factor) pathway, which senses and reacts to changes in O_2 supply (Kaelin Jr and Ratcliffe 2008). Despite these recent insights, the evolutionary history of these adaptive alleles remains poorly understood.

The genetic history of East Asian populations includes complex patterns of ancient admixture (Patterson et al. 2012; Lipson et al. 2013; Loh et al. 2013), but little is known about the genetic relationship of Tibetans with other East Asian populations. The dramatic growth of low-altitude East Asian populations in the past 10,000-30,000 years inferred based on genome sequence data (Li and Durbin 2011; Meyer et al. 2012) is likely to have created intense demographic pressure, possibly leading to expansion into the Tibetan plateau and genetic exchange with resident populations. This mixing of populations from different local

environments, in turn, would create the potential for transfer of alleles advantageous at high-altitude to the gene pool of migrants.

To resolve the genetic history of Tibetans and of their adaptation to high-altitude, we obtained genetic and phenotypic data for a sample of 69 Sherpa, a population famous for their superb performance in mountaineering and an example of successful adaptation to high-altitude environments. All sampled individuals were born and raised at ≥ 3,000 m altitude in the Himalayas. Genotypes of 96 unrelated Tibetan individuals from three previous studies (Beall et al. 2010; Simonson et al. 2010; Wang et al. 2011) were also analyzed. These individuals were sampled in three different high-altitude regions of the plateau: the Tibet Autonomous Region (near Lhasa) (Wang et al. 2011), Yunnan (Beall et al. 2010) and Qinghai (Simonson et al. 2010) provinces in China (3,200-4,350 m altitude). We merged the genotype data of the Sherpa and Tibetans with the International HapMap phase 3 (HapMap3) dataset (Altshuler et al. 2010) using imputation for non-overlapping variants. For some analyses, this data set was combined with additional genotype data for the following populations: worldwide populations in the HGDP (Human Genome Diversity Panel) (Li et al. 2008; Patterson et al. 2012), Indian and Central Asian populations (Xing et al. 2010) and two Siberian populations (Hancock et al. 2011).

Here, we show that Tibetans are the admixed descendants of ancestral populations related to contemporary Sherpa and Han Chinese. We also show that high-altitude adaptive variants originated in an ancestral population (represented by present-day Sherpa) and that they preferentially propagated in the Tibetan gene pool after admixture. Our results provide a clear example of transfer of adaptive alleles between human populations, which is supported by ancestry-based tests, population genetic signatures of local adaptations and by adaptive phenotype data.

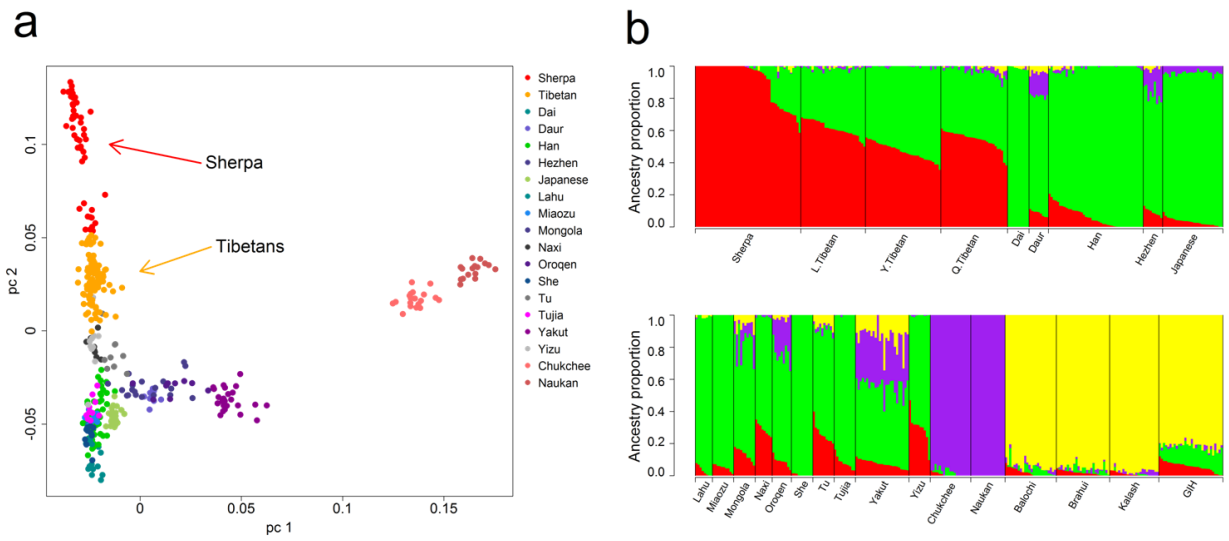
2.3: Results

The admixture origin of Tibetans

We first conducted descriptive analyses to assess the population structure of the Sherpa and Tibetans within the context of other Asians. Interestingly, the Sherpa and Tibetans form a major axis of genetic variation in principal component analysis (PCA) (Patterson et al. 2006), in which Tibetans are located between the Sherpa and other East Asians (PC2 in **Figure 2.1a**). Although the pattern observed in the PCA plot can result from several demographic processes (e.g. strong genetic drift in the Sherpa), it is also consistent with a history of admixture in Tibetans between ancestral populations closely related to the contemporary Sherpa and low-altitude East Asians. Unsupervised clustering analysis using ADMIXTURE (Alexander et al. 2009) also infers Tibetans as a mixture of two genetic components: one is highly enriched in the Sherpa (but rare in lowlander populations) and will be referred to as the "high-altitude component", and the other in low-altitude East Asians and will be referred to as the "low-altitude component" (**Figure 2.1b**). The inclusion of a broader range of Asian populations shows that the high-altitude component is not due to shared ancestry with South or Central Asians (**Supplementary Figures 2.1-2.3**). The Sherpa also show evidence of admixture with East Asians (**Supplementary Table 2.1**) and marked inter-individual variation in ancestry proportions, but they are unique in harboring individuals with 100% inferred high-altitude component (**Figure 2.1, Supplementary Figures 2.1 and 2.2**). The date of this East Asian admixture into the Sherpa was estimated as 23.4 generations ago, based on the decay of linkage disequilibrium (LD) (Loh et al. 2013) (see Methods; **Supplementary Figures 2.4 and 2.5, and Supplementary Table 2.2**). This date is in close agreement with the historical record of a

Sherpa migration out of their ancestral homeland in Eastern Tibet 400-600 years ago to their current place, Solu-Khumbu (Oppitz 1974). In subsequent analyses, we considered the subset of 21 individuals with 100% high-altitude component (referred to as HA-proxy sample; "HA" for high-altitude) to be the descendants of an ancient high-altitude population with a broad geographic distribution across the plateau. Taken together, these results strongly suggest that Tibetans are admixed descendants of two populations currently represented by the HA-proxy and low-altitude East Asians (such as Han Chinese), while the Sherpa more recently experienced admixture with nearby East Asian populations, most likely Tibetans (**Supplementary Figure 2.5**).

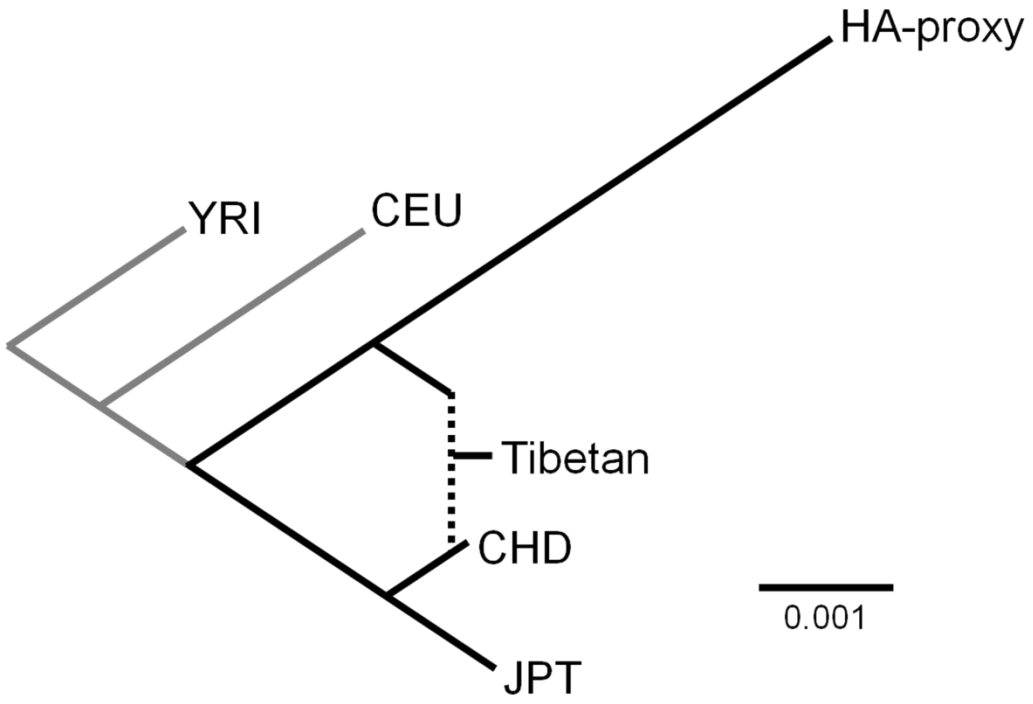
Figure 2.1: The genetic structure of Sherpa and Tibetans relative to other East Asian populations. (a) Principal component analysis (PCA) of Sherpa (49 unrelated individuals), Tibetans (n = 96), Maritime Chukchee (n = 19), Naukan Yup'ik (n = 16) and East Asian populations from the HGDP (n = 210). PC1 and PC2 explain 2.5% and 1.2% of total variation, respectively. (b) ADMIXTURE analysis with K = 4. Red and green colors represent the high-altitude and low-altitude components, respectively. Yellow and purple ancestries are mainly present in the Indian-Pakistani and Siberian populations, respectively. L.Tibetan = Lhasa Tibetan (Wang et al. 2011) (n = 30); Y.Tibetan = Yunnan Tibetan (Beall et al. 2010) (n = 35); Q.Tibetan = Qinghai Tibetan (Simonson et al. 2010) (n = 31); GIH = HapMap3 GIH (Gujarati Indians from Houston, Texas; n = 30). Balochi (n = 24), Brahui (n = 25) and Kalash (n = 23) are from the HGDP.



Formal tests of admixture give further support to the admixture hypothesis of Tibetan origin. We used the D-test and 3-population (f_3) test (Patterson et al. 2012), both of which exploit the pattern of shared genetic drift from the moments of allele frequency. The D-test asks if a set of four populations fits into a simple bifurcating tree. Under this null, two pairs of populations with non-overlapping drift paths are expected; thus, there is no shared genetic drift and the expectation of the D statistic is zero. Admixture breaks up the bifurcating tree topology and generates an overlap of drift paths, making the D statistic deviate from zero. The 3-population test uses information about the shared genetic drift between a target population and each of two reference populations. Under the null of a bifurcating tree, the shared genetic drift is the drift occurred on the branch leading to the target population and the f_3 statistic is expected to take a markedly positive value. Therefore, a negative value is strong evidence for a deviation from the null. We used HapMap3 YRI (Yoruba in Ibadan, Nigeria) as the outgroup, CHD (Chinese in Metropolitan Denver, Colorado) as representative of the ancestral low-altitude East Asians, and the HA-proxy sample as representative of the ancestral high-altitude population. D-test results were highly significant for all three Tibetan populations (D = -5.1 to -9.8 standard deviation (SD); see Methods; **Supplementary Table 2.3**). The 3-population test results were also negative (f_3 = -1.5 to -3.5 SD; see Methods; **Supplementary Table 2.3**). Tests with the other two HapMap3 East Asian populations, CHB (Han Chinese in Beijing, China) and JPT (Japanese in Tokyo, Japan), showed similar results (**Supplementary Table 2.3**). Furthermore, all the Tibetan samples were consistently positioned in the population tree as a mixture of branches leading to the HA-proxy and CHD, using either *MixMapper* (Lipson et al. 2013) (**Figure 2.2 and Supplementary Table 2.4**) or *TreeMix* (Pickrell and Pritchard 2012) (**Supplementary Figure 2.6 and Supplementary Table 2.5**). An LD decay method (Loh et al. 2013) that tests for admixture and

estimates the admixture timing did not detect a signal in Tibetans and, hence, could not be used to estimate the admixture timing. This method is known to lose power with increasing time since admixture more quickly than the 3-population test (Patterson et al. 2012) and, possibly, other methods based on genetic drift. Given the significant admixture signals from the 3-population test, the D-test, *MixMapper*, and *TreeMix* as well as the patterns observed in the ADMIXTURE and PCA plots, we speculate that the results of the LD decay method reflect an old admixture date (see Methods; **Supplementary Figure 2.7 and Supplementary Table 2.2**).

Figure 2.2: Tibetans as a mixture of the HA-proxy and Han Chinese related ancestral populations in the scaffold tree. Grey branches are not drawn to proportion. The scale bar represents 0.001 in the drift unit. YRI = Yoruba in Ibadan, Nigeria; CEU = Utah residents with Northern and Western European ancestry from the CEPH collection; CHD = Chinese in Metropolis Denver, Colorado; JPT = Japanese in Tokyo, Japan.



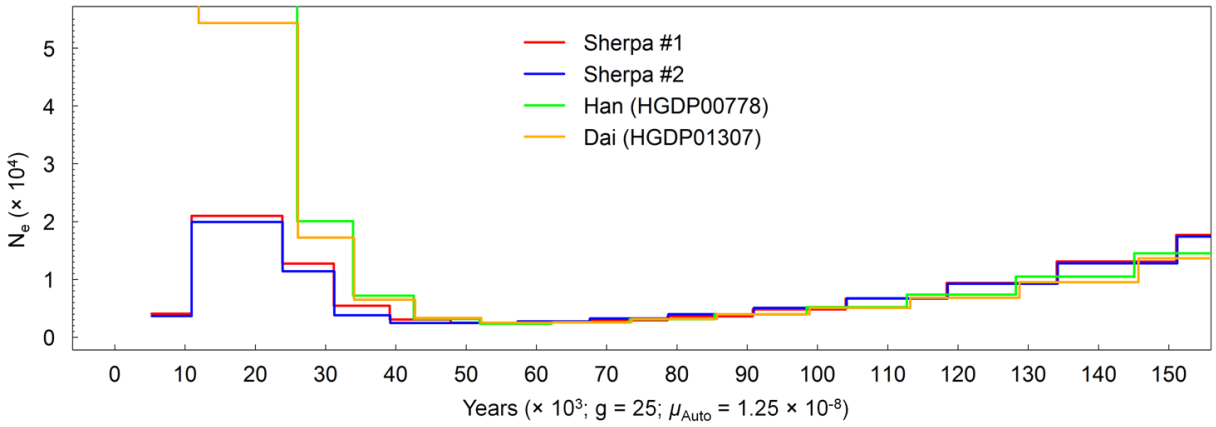
The demography of the ancestral high-altitude population

To learn about the genetic history of the ancestral population that contributed to the Tibetan gene pool, we sequenced the genomes of two HA-proxy males to high coverage (27-30x) and inferred the history of population size using the Pairwise Sequentially Markovian Coalescent (PSMC) model (Li and Durbin 2011). Interestingly, the high-altitude demographic profile begins to diverge from that of Han Chinese and Dai (Meyer et al. 2012) approximately 40,000 years ago and shows no signature of the dramatic exponential population growth characterizing low-altitude East Asians (see Methods; **Figure 2.3**). The same analysis with pairs of X chromosome sequences gives estimates of the split time between the high-altitude and the low-altitude populations of approximately 20,000 years ago (see Methods; **Supplementary Figure 2.8**). These results suggest that the ancestral high-altitude population diverged from other low-altitude East Asians in our sample during the upper Paleolithic. This is consistent with previous proposals based on archaeological, mitochondrial DNA (mtDNA) and Y chromosome data, pointing to an initial colonization of the Tibetan plateau around 30,000 years ago (Aldenderfer 2011; Qi et al. 2013). A second, more recent migration to Tibet was proposed based on mtDNA and Y chromosome data (Aldenderfer 2011; Qi et al. 2013), which is consistent not only with our admixture hypothesis but also with archaeological evidence (Aldenderfer 2011; Qi et al. 2013).

To learn more about the history of the newly detected high-altitude ancestry component, we analyzed the HA-proxy samples for admixture with archaic humans, i.e. Neanderthal and Denisovan. We tested if the HA-proxy or Tibetans experienced different levels of admixture with archaic hominins relative to other modern human populations. When projected onto the PC plane defined by the Chimpanzee, Neanderthal and Denisovan genotype data, the HA-proxy and Tibetans clustered with the other HGDP East Asian populations (**Supplementary Figure 2.9**).

The D-test results also show that the HA-proxy and Tibetans have Neanderthal ancestry similar to those of Eurasian populations but higher than those of African populations. As all other Eurasian populations, the HA-proxy and Tibetans have Denisovan ancestry lower than that of Papuans (**Supplementary Table 2.6**). These results suggest that the history of ancient admixture in the ancestral high altitude population does not differ substantially from that of other East Asian populations.

Figure 2.3: Whole genome sequence based inference on effective population size. The effective population sizes (N_e) of the ancestral high- and low-altitude populations are inferred from whole genome sequences of two Sherpa and two low-altitude East Asians (a Han and a Dai individual) (g = generation time in year; μ_{Auto} = autosomal neutral mutation rate per base per generation).



Selection and association signals in Sherpa and Tibetans

At the phenotypic level, the Sherpa, like Tibetans, have unelevated hemoglobin concentration at high-altitude (Beall et al. 1998) (**Supplementary Table 2.7**). To ask if the sharing of this adaptive phenotype is due to shared beneficial alleles, we conducted population genetic and genetic association analyses in our Sherpa data. The SNPs near *EGLN1* and *EPAS1* with the highest population branch statistic (PBS) (Yi et al. 2010) in Tibetans also had top PBS values in the HA-proxy (see Methods; **Supplementary Table 2.8**). We also replicated, in the Sherpa, associations between *EPAS1* SNPs and hemoglobin concentration previously reported in Tibetans (Beall et al. 2010). Specifically, 19 out of 26 significant SNPs in Tibetans were also associated with hemoglobin levels in the Sherpa (linear mixed model $p < 0.05$; N = 69), all with the same allelic direction reported in Tibetans (see Methods; **Supplementary Table 2.9**). This is a significant enrichment in low *p*-values as only 3 out of 1000 permutations showed greater or equal to 19 SNPs with $p < 0.05$ (**Supplementary Table 2.10**).

Adaptive excess of high-altitude ancestry in Tibetans

These findings indicate that Tibetans share genetic adaptations with the Sherpa despite a substantial amount of gene flow from low-altitude East Asians (**Figure 2.1**), leading to the hypothesis that alleles associated with lower hemoglobin levels preferentially propagated in the admixed gene pool. Consistent with this hypothesis, an analysis of inferred local ancestry of Tibetans using HAPMIX (Price et al. 2009) clearly showed that the *EGLN1* (3.59 SD) and *EPAS1* (3.74 SD) genes are highly enriched for high-altitude ancestry, representing respectively the second and the third strongest signals of excess high-altitude ancestry in the Tibetan genome (see Methods; **Figure 2.4 and Supplementary Figure 2.10**). In addition, SNPs with excess

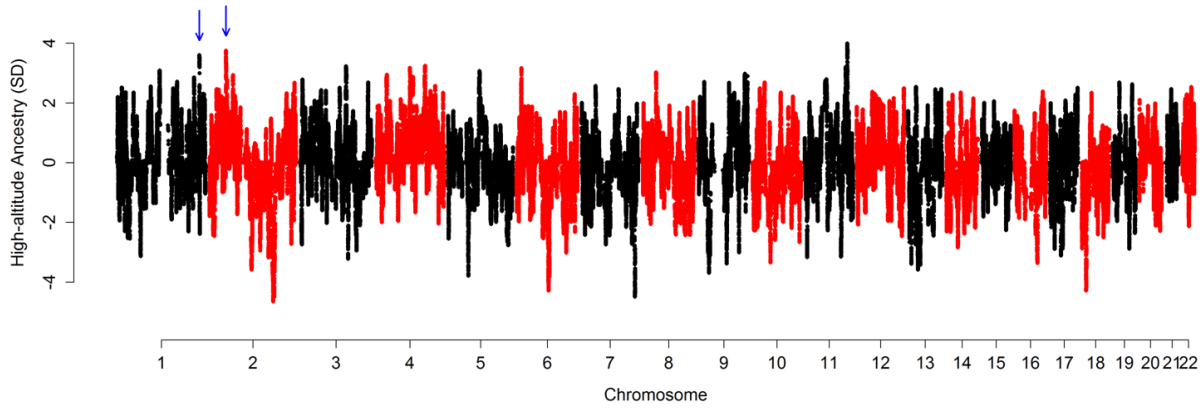
high-altitude ancestry are significantly enriched around genes involved in the HIF pathway (Reactome pathways gene set "Cellular response to hypoxia"; **Supplementary Table 2.11**) (Matthews et al. 2009), even after removing *EGLN1* and *EPAS1* genes (see Methods; **Supplementary Table 2.12**). We also modeled Tibetan allele frequencies as a linear combination of those of the 11 HapMap3 populations and the HA-proxy (see Methods) (Alkorta-Aranburu et al. 2012). SNPs with large residuals in the linear model are likely to be a departure from the average admixture proportion in the Tibetan genome. Using multiple regression and the linear coefficients estimated using all SNPs, we found that SNPs in and around *EGLN1* and *EPAS1* SNPs had extremely large residuals, further supporting the idea that alleles in these genes disproportionately increased in frequency in Tibetans after admixture (**Supplementary Figures 2.11 and 2.12**).

Interestingly, the SNP with the largest high-altitude ancestry proportion in the Tibetan genome (rs1003081; **Figure 2.4**) lies less than 1 kb away from a candidate gene for hypoxia-adaptations, hypoxia up-regulated protein 1 (*HYOU1*), which encodes a molecular chaperone induced in the endoplasmic reticulum under hypoxic condition (Chene et al. 2006; Sanson et al. 2008) (**Supplementary Figure 2.13**). The same SNP is also a *cis* expression quantitative trait locus for another nearby candidate gene, hydroxymethylbilane synthase (*HMBS*; **Supplementary Figure 2.13**), in liver (Schadt et al. 2008) and monocytes (Zeller et al. 2010). This gene encodes the third enzyme in the heme biosynthesis pathway (Gubin and Miller 2001); therefore, variation in this gene could have a direct effect on hemoglobin concentration. We tested SNPs in this peak of high-altitude ancestry for an association with hemoglobin levels in the 69 Sherpa and found that 19 of 64 SNPs have $p < 0.05$ (**Supplementary Table 2.13**; linear mixed model). Only 12 out of 1,000 permutations have 19 or more SNPs with $p < 0.05$,

indicating a significant enrichment of low association *p*-values in this region (**Supplementary Table 2.14**).

A local excess of high-altitude ancestry might also be generated under a simple branching model of population history without admixture if selection acted in parallel in the Sherpa and Tibetans on a variant that predated their split. However, given the strong evidence for admixture described above, the excess of high-altitude ancestry at known adaptive loci argues for selection on these variants in the admixed population.

Figure 2.4: The distribution of high-altitude ancestry proportions across the Tibetan genome. Blue arrows mark the positions of *EGLN1* (in chromosome 1) and *EPAS1* (in chromosome 2).



2.4: Discussion

Our results show that adaptations can be facilitated by admixture with locally adapted resident populations. Gene flow has been appreciated as a source of adaptive alleles by theoretical population geneticists for a long time, but few cases in animals have been documented at the empirical level and several of them have only partial support in the data (Wright 1931; Anderson 1949; Stebbins 1959; Lewontin and Birch 1966; Arnold 2004; Hedrick 2013). Here, we reported a case of adaptation facilitated by gene flow between modern human populations that is supported by ancestry-based analysis, phenotypic data and population genetic evidence. Given the prevalence of admixture in humans, our findings suggest that, this mode of adaptation may be common and that further examples may be found in other populations, e.g. Asian Indians (Reich et al. 2009), Europeans (Pinhasi et al. 2012) and Sub-Saharan Africans (Pickrell et al. 2012), that originated from the mixture of ancestral gene pools with different local adaptations. Interestingly, the evolutionary dynamics of these admixture-driven adaptations closely resembles that of adaptively introgressed alleles across a species boundary.

Our demographic model for high-altitude populations suggests a much older split between high- and lowlanders compared to a previous model based on exome sequence data (Yi et al. 2010). As a consequence, we estimated selection coefficients of 0.0004-0.0023 for the variants in *EGLN1* and *EPAS1* gene regions in the HA-proxy sample (see Methods; **Supplementary Table 2.15**). In contrast to a previous proposal (Yi et al. 2010), our estimates are an order of magnitude smaller than those for lactose tolerance alleles in Europe and Africa (Tishkoff et al. 2007). However, all these estimates depend on many model assumptions and are

typically associated with large uncertainties; therefore, further analyses are necessary to obtain accurate estimates of the selection coefficient.

The acquisition of adaptive variants through gene flow adds a new dimension to the ongoing debate on the relative importance of selective sweeps and polygenic adaptations in human evolution (Pritchard and Di Rienzo 2010). Unlike selective sweeps and polygenic adaptations, gene flow can introduce a suite of adaptive alleles that evolved in concert and that segregate at appreciable frequencies in the admixed population. Finally, our findings highlight the importance of sampling unique branches of the human population tree, such as the Sherpa, to detect admixture events and to elucidate the history of human adaptations.

2.5: Methods

Study subjects

Demographic, phenotypic, and genetic data were collected from 69 high-altitude native Sherpa residing in two villages at 3,800 m in the Khumbu region of Nepal during the summer of 2010 (**Supplementary Table 2.7**). The participants were healthy non-smoking men and women, born and raised above 3,000 m altitude, 17-54 years of age, who had not traveled to areas lower than 2,400 m or higher than 5,400 m in the previous six months, had normal body mass index, lung function and blood pressure, were not anemic, did not have fever and were not pregnant. All study participants provided written informed consent. This study was approved by the IRBs at Case Western Reserve University and University of Chicago, by the Oxford Tropical Research Ethics Committee and by the Nepal Health Research Council.

Phenotypic data collection

Standing height without shoes (GPM Anthropometer, Stuttgart, Switzerland) and weight in light clothing (Pelouze Mechanical Shipping Scale, P114S, Bridgeview, IL) were measured to the nearest mm and pound, respectively, according to standard protocol (Weiner and Lourie 1969). Blood hemoglobin concentration was measured in duplicate using the cyanmethemoglobin technique (Hemocue, Angelholm, Switzerland) immediately after venipuncture.

Genotype data

All the Sherpa samples were genotyped using Illumina HumanOmni1-Quad beadchip in the Genomics Core Facility at Case Western Reserve University. We removed SNPs with call rate < 95%, minor allele frequency < 5% or with strand-ambiguity. Genetic relatives were inferred from a random set of 2,000 autosomal SNPs using Relpair v2.0.1 (Epstein et al. 2000). Twenty related individuals with closer relationships than first cousins were excluded in subsequent analyses except for the genotype-phenotype association test. We also obtained genotype data of 96 Tibetans from three previous studies, each including 30-35 individuals (**Supplementary Table 2.16**) (Beall et al. 2010; Simonson et al. 2010; Wang et al. 2011). Because all three data sets are from different regions of the Tibetan plateau and from different genotyping platforms (Illumina Human1M-Duo v3 for Lhasa Tibetans (Wang et al. 2011), Illumina Human610-Quad for Yunnan Tibetans (Beall et al. 2010) and Affymetrix Genome-wide Human SNP 6.0 for Qinghai Tibetans (Simonson et al. 2010)), most of the analyses were conducted on the three Tibetan samples separately. To maximize the overlap between data sets, we imputed the Sherpa and the three Tibetan genotype data sets using *IMPUTE2* (Howie et al.

2009) with the HapMap3 data set (**Supplementary Table 2.17**) (Altshuler et al. 2010) as a reference. Each of 4 data sets was imputed separately. For each individual and SNP, we called a genotype if it had posterior probability ≥ 0.9, otherwise, we treated it as missing data. We excluded SNPs with call rate < 96.5% in the Sherpa or in any of the three Tibetan samples (≥ 3 missing genotypes in the Sherpa or ≥ 2 missing genotypes in Tibetans). This process yielded 543,555 SNPs overlapping between the Sherpa, Tibetans, and HapMap3 data sets ("HM3"; n = 1,165); prior to imputation, 80,819 overlapping SNPs were identified. To analyze Sherpa and Tibetan genetic variation in a broader context, we overlapped this data set with each of three additional data sets. First, we overlapped it with data from HGDP (Li et al. 2008) and two Siberian populations, Naukan Yup'ik and maritime Chukchee (Hancock et al. 2011), to obtain 50,464 SNPs with genotype call rate ≥ 96.5% for all populations in the data set ("HM3-HGDP"; n = 2,138). Second, we overlapped the HM3 data set with HGDP individuals genotyped on Affymetrix Axiom® Genome-wide Human Origins 1 array, described by Patterson et al (Patterson et al. 2012). We obtained 58,756 SNPs ("HM3-HumanOrigin"; n = 2,107) after excluding strand-ambiguous or low quality SNPs (≥ 2 missing genotypes in any of the 53 HGDP populations). Last, we overlapped the HM3 data set with the genotype data of 14 Asian populations (Thai, Vietnamese, Cambodian, Iban, Buryat, Kyrgyzstani, Nepalese, Pakistani, Andhra Pradesh Brahmins, Andhra Pradesh Madiga, Andhra Pradesh Mala, Tamil Nadu Brahmins, Tamil Nadu Dalit and Irulas) from Xing et al (Xing et al. 2010). We obtained 62,541 SNPs after excluding strand-ambiguous or low quality (≥ 2 missing genotypes in any of the 14 Asian populations) SNPs ("HM3-Asian"; n = 1,418).

Analyses of admixture

We used EIGENSOFT 4.2 (Patterson et al. 2006) to perform principal component analysis (PCA). For unsupervised clustering analysis, we used ADMIXTURE v1.22 (Alexander et al. 2009) with 5-fold cross validation to find the optimal number of clusters. When using the HM3-Asian data set for ADMIXTURE analysis, we generated linkage disequilibrium (LD) trimmed SNP sets by removing one SNP from each pair of SNPs with $r^2 > 0.2$ in 50 SNP blocks using PLINK v1.07 (Purcell et al. 2007). The D-test and 3-population test were performed as implemented in ADMIXTOOLS v1.1 (Patterson et al. 2012). We used ALDER (Loh et al. 2013) to date the admixture event using admixture LD decay. To reduce the noise introduced by genetic drift specific to the HA-proxy sample and CHD, we used SNP loadings of a principal component (PC) representing the Sherpa-Tibetan axis of genetic variation as a weight vector instead of allele frequency difference between two reference populations. To obtain this weight vector, we ran a PCA with the Sherpa, Tibetans, and HapMap3 CHD (Chinese in Metropolitan Denver, Colorado), CHB (Han Chinese in Beijing, China), JPT (Japanese in Tokyo, Japan) and GIH (Gujarati Indians in Huston, Texas) (**Supplementary Figure 2.14**). SNP loadings of the second PC were used as a weight vector. A bin size of 0.25 centiMorgan (cM) was used for all analyses. When fitting the exponential curve, we excluded SNP pairs within distance bins of 0.5 cM or smaller to remove those in LD in the ancestral populations. Significance of the estimates was assessed by block jackknifing of one chromosome at a time.

Whole-genome sequence analysis

Two Sherpa males with 100% high-altitude ancestry component were chosen for 100-bp paired-end whole genome sequencing using Illumina HiSeq 2000. We followed a standard

Illumina TruSeq DNA sample preparation protocol for paired-end sequencing to construct sequencing libraries. Each sample was tagged with a unique adapter sequence. We mixed an equal amount of library DNA from two samples and sequenced the mixed library in a flow cell. After separating raw reads using adapter sequences, we mapped reads onto the human genome reference sequence (hg19) using BWA (Li and Durbin 2009). We further adjusted the alignment by conducting local realignment around indels and base quality recalibration steps using Genome Analysis Tool Kit (McKenna et al. 2010). After removing duplicates using Picard, we called consensus genome sequences of each individual using Samtools v0.1.19 (Li et al. 2009). We applied the Pairwise Sequentially Markovian Coalescent (PSMC) model (Li and Durbin 2011) to autosomal consensus sequence of each sample separately, using the following options: -N15 -t15 -r5 -p "2*2+50*1+4+6". To compare them with low-altitude East Asians, we downloaded the aligned sequence reads (in BAM file format) for a Han Chinese and a Dai individual from Meyer et al (Meyer et al. 2012). These sequence data were generated using essentially the same protocol as for our sequence data: sequencing libraries were prepared following the standard Illumina TruSeq DNA sample preparation protocol for paired-end sequencing of 101 bp reads and 200-400 bp insert size, sequencing was performed on an Illumina HiSeq 2000, reads were mapped to the human genome reference sequence (hg19) using BWA (Li and Durbin 2009), and the coverage was 27.7x and 28.3x for a Han and a Dai individual, respectively. In addition, to minimize any bias introduced by differences in post-alignment processing, we processed the Han and Dai reads in the same way as we did for the Sherpa samples. To convert the estimates of population parameters into the effective population size (N_e) and time in year, we used autosomal neutral mutation rate $\mu_{Auto} = 1.25 \times 10^{-8}$ per base-pair per generation and 25 years per generation (Scally and Durbin 2012). We also applied the PSMC model to composite diploid X

chromosome sequences, composed of two haploid X chromosome sequences from two males. For this analysis, we called the genotypes of X chromosome sequences using the same pipeline applied to the autosomal sequences and called a heterozygote if the haploid genotype calls of two individuals are different at a site. Sites with a missing genotype in either of two individuals were coded as missing data. We ran the PSMC with the options -N25 -t15 -r5 -p "6+2*4+3+13*2+3+2*4+6". Time bins were sliced in a coarser way than those for the autosomal data, considering the lower resolution of X chromosome data. We adjusted the neutral mutation rate for X chromosome (μ_X) by applying the ratio of male-to-female mutation rate $\alpha = 2$ (Miyata et al. 1987) and the formula $\mu_X = \mu_{Auto} \times [2(2+\alpha)] / [3(1+\alpha)] = 1.11\times10^{-8}$ (Li and Durbin 2011). All raw sequences generated in this study have been deposited into the Sequence Read Archive (NCBI) with the accession numbers SRS520217 and SRS520218.

Local ancestry estimation

We estimated local ancestry across the genome of three Tibetan samples using HAPMIX (Price et al. 2009), using the 21 HA-proxy and the 21 CHD individuals with the highest proportion of the low-altitude component as the reference populations (**Figure 2.1b**). Phased genotypes of CHD individuals were obtained from the HapMap3 website. The Sherpa genotype data were phased by using fastPHASE (Scheet and Stephens 2006). We estimated the local ancestry of each of the three Tibetan samples separately, using 50% admixture proportion, 80 generations since admixture and population recombination parameter $\rho = 600$ for both reference populations. To summarize the local ancestry estimates across the three samples, we first individually centered local ancestry estimates by subtracting individual mean local ancestry. Then, we averaged local ancestry of each SNP across all individuals and standardized these

mean values of local ancestry. We repeated this step for each of three Tibetan samples separately. To guard against the effect of genotyping error in the estimated local ancestry, we additionally ran HAPMIX with only half of the SNPs in Tibetans (odd and even numbered SNPs, respectively). False local ancestry peaks inferred from genotyping errors at a SNP will disappear in only one of these two sets. Therefore, we defined a penalty for the mean local ancestry value (MLA) of each SNP, where the penalty is the absolute difference between MLAs from the odd-numbered and even-numbered SNP sets. We adjusted the estimated mean local ancestry by

$$MLA_{adjusted} = MLA_{all} - \frac{MLA_{all}}{|MLA_{all}|} \times |MLA_{odd} - MLA_{even}|$$

So, this adjustment reduces the peak height toward zero in proportion to the difference between odd/even SNP sets. If this adjustment changes the sign of MLA, we set MLA to zero. After adjustment, we standardized MLA (**Figure 2.4 and Supplementary Figure 2.10**).

Gene set enrichment analysis

We tested for an enrichment of genes involved in the response to hypoxia in the regions with excess high-altitude ancestry across the Tibetan genome. We focused on the Reactome pathway gene set "Cellular response to hypoxia" (25 genes) and its subset "Oxygen-dependent proline hydroxylation of hypoxia-inducible factor alpha" (18 genes; **Supplementary Table 2.11**) (Matthews et al. 2009). To determine whether there was an excess of SNPs with high high-altitude ancestry in the gene sets, we calculated top 0.5, 1.0 or 5.0 percentile of the high-altitude local ancestry of all SNPs within 10 kb of all genes except for the hypoxia genes. Then, we calculated the proportion of SNPs with higher high-altitude local ancestry than the above percentiles within 10 kb of genes in the gene set of interest. We repeated this process 1,000 times by bootstrapping the whole genome and counted the number of bootstrap replicates for which the

proportion of the top high-altitude ancestry SNPs in the hypoxia genes is higher than the corresponding percentile. We repeated the enrichment analysis after removing *EGLN1* and *EPAS1* genes from the gene set.

Population branch statistic analysis

To detect SNPs showing high divergence of allele frequency in the Sherpa and Tibetans from low-altitude East Asians, we calculated the population branch statistic (PBS) of each SNP in the HA-proxy and each of three Tibetan samples. To maximize the amount of the genome covered by SNPs, we used a lenient SNP filtering by allowing up to 10% missing genotypes for the HA-proxy and for each of three Tibetan samples, resulting in 879,434 SNPs ("extended HM3"). We also merged all three Tibetan samples to increase accuracy in allele frequency estimates. We chose HapMap3 CHD and CEU (Utah residents with Northern and Western European ancestry from the CEPH collection) to represent a low-altitude East Asian population and the outgroup, respectively. We followed Weir and Cockerham (Weir and Cockerham 1984) to calculate pair-wise F_{ST} and Yi et al (Yi et al. 2010) to calculate PBS. We retrieved SNPs with the highest rank around *EGLN1* and *EPAS1* genes (within 300 kb) in Tibetans and checked their ranks in the HA-proxy. All analyses were performed using R v2.15.1 (R Core Team 2012).

Multiple regression analysis

To identify SNPs with extreme frequency divergence in Tibetans while taking into account their history of admixture, we used a multiple regression (MR) analysis described in Alkorta-Aranburu et al (Alkorta-Aranburu et al. 2012). Briefly, we modeled Tibetan allele frequencies in the extended HM3 data set as a linear combination of those of the 11 HapMap3

populations (**Supplementary Table 2.17**) and the HA-proxy. Alleles were polarized based on the global allele frequency. We obtained the standardized and squared residuals ("MR score") from the above MR model. We took the rank of each SNP, divided it by the total number of the SNPs and minus $\log_{10}$ transformed to get the transformed rank.

Genetic association with hemoglobin levels in the Sherpa

For the *EPAS1* gene region on chromosome 2, SNP genotypes in a 5 Mb region were imputed using *IMPUTE2* (Howie et al. 2009) with 1000 Genomes phase 1 integrated variant set (The 1000 Genomes Project Consortium 2012) as a reference panel. SNPs with information metric ≥ 0.9 were chosen for downstream analysis. Twenty-six SNPs passed this threshold among those reported to be associated with hemoglobin concentration in two Tibetan cohorts[3]. For the *HYOU1*/*HMBS* gene region on chromosome 11, we selected the 64 genotyped SNPs with high-altitude ancestry ≥ 3.7 SD (**Supplementary Figure 2.13**). We tested for a genetic association between mean posterior genotype of these SNPs and hemoglobin concentration in all 69 Sherpa individuals. We took relatedness among samples into account through a linear mixed model (LMM) scheme using GEMMA (Zhou and Stephens 2012) (**Supplementary Tables 2.9 and 2.13**). Genetic relatedness between individuals was estimated using the genome-wide genotype data. We included sex as a covariate. One thousand permutations of concentration across individuals were performed to test if the results were significantly enriched for low *p*-values (**Supplementary Tables 2.10 and 2.14**).

Mapping of admixed populations on phylogenetic trees

We used *MixMapper* (Lipson et al. 2013) and *TreeMix* (Pickrell and Pritchard 2012) to infer the ancestral populations contributing to the Tibetan gene pool and the proportion of admixture in a single step. We ran *MixMapper* with a scaffold tree of five populations (the HA-proxy, HapMap3 YRI, CEU, CHD and JPT) using the HM3 data set. These five populations were chosen to cover major branches of human populations and to reduce computational load. We also ran *TreeMix*, with each of the three Tibetan populations and the above five populations. We allowed three admixture events for each of these six-population sets. In both *MixMapper* and *TreeMix* analyses, we performed 500 bootstrap replicates with 50 SNPs as a block for bootstrap sampling.

Archaic human admixture in the Sherpa and Tibetans

We ran PCA for Chimpanzee, Neanderthal and Denisovan genotype data (HM3-HumanOrigin data set) and obtained the projection of modern human individuals on these PCs. Population means of PC1 and PC2 were plotted for the 53 HGDP populations, the HA-proxy and three Tibetan samples. We also performed the D-test. The D statistic was calculated for each of $((H_1, H_2), (A, \text{Chimpanzee}))$ quartets with the HM3-HumanOrigin data set: H_1 = the 53 HGDP populations; H_2 = the HA-proxy / Tibetans; A = Neanderthal or Denisovan.

Estimating the selection coefficient

Given the sharing of adaptive variants in the *EGLN1* and *EPAS1* gene regions between Tibetans and the HA-proxy, we estimated selection coefficients of these variants in the HA-

proxy because its demography is simpler than that of Tibetans. Here we applied a simple deterministic model of selective sweep with additive genetic effects, using the following formula:

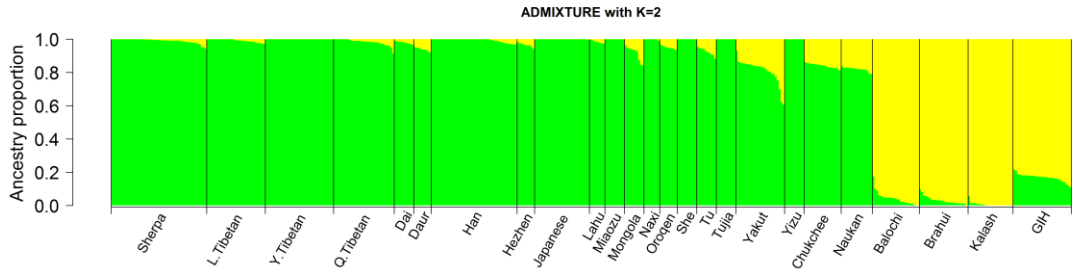
$$s = \frac{1}{t} \log \frac{p_t(1-p_0)}{p_0(1-p_t)}$$

We estimated the initial allele frequency (p_0) as the mean allele frequency of the three HapMap3 East Asian populations: CHD, CHB and JPT. The current allele frequency (p_t) was estimated as the allele frequency of the HA-proxy. We chose 8 SNPs (5 in *EGLN1* and 3 in *EPAS1* region; **Supplementary Table 2.15**), which have top PBS signals in the HA-proxy and Tibetans and are not in complete LD with each other. Based on our estimate of the split time of 20,000 years, we used 800 generations (with 25 years per generation; **Supplementary Figure 2.8**) for the onset of selection (t) in the ancestral high-altitude population, assuming that selection began right after the population split.

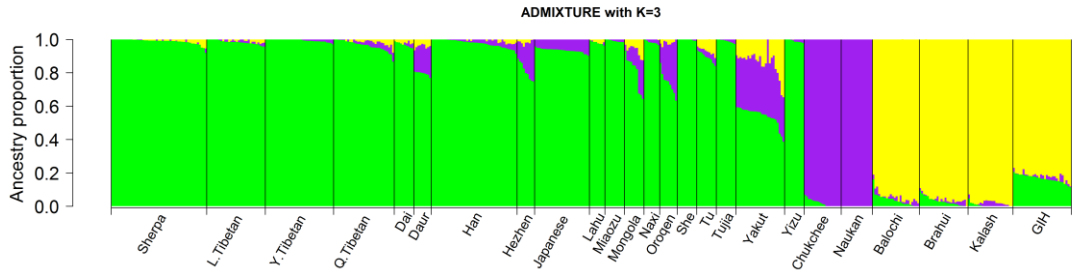
2.6: Appendix: Supplementary Materials

Supplementary Figure 2.1: ADMIXTURE analysis of the Sherpa, Tibetans and 21 additional Asian populations (a subset of HM3-HGDP data set) with K = 2 to 5. Cross validation (CV) error was lowest at K = 4. (a) K = 2 shows East Asian (green) and South Asian (yellow) components. (b) K = 3 separates the Siberian component (purple) from the East Asian one. (c) K = 4 separates the high-altitude component (red) from the low-altitude East Asian one (a repeat of Fig. 1B). (d) K = 5 separates Yakut population (skyblue).

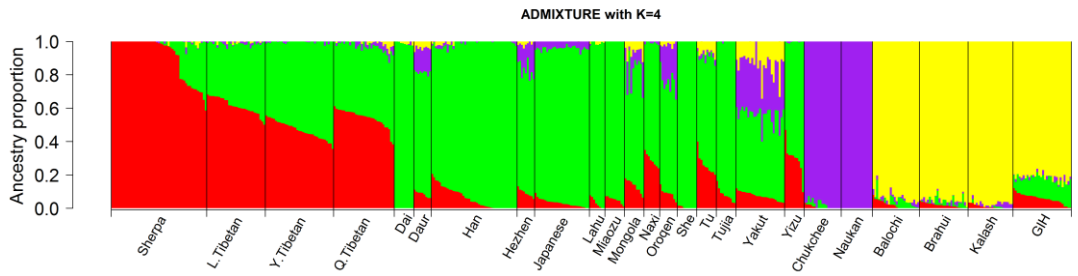
(a)



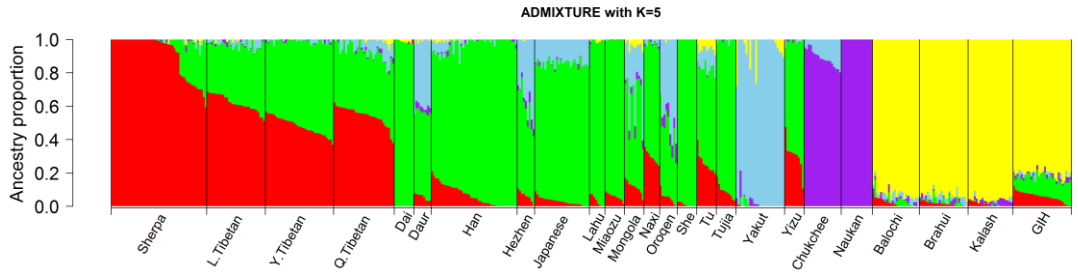
(b)



(c)

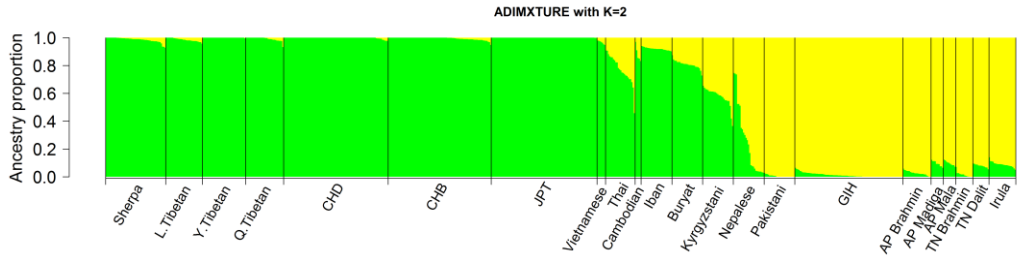


(d)

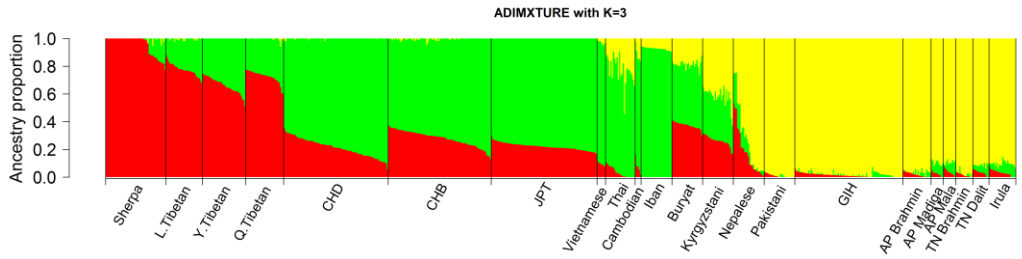


Supplementary Figure 2.2: ADMIXTURE analysis of the Sherpa, Tibetans and 18 Asian populations (a subset of HM3-Asian data set) with K = 2 to 5. The Sherpa ancestry is not related to Indian ancestry. CV error was lowest at K = 5. (a) K = 2 shows East Asian (green) and South Asian (yellow) components. (b) K = 3 separates the high-altitude component (red) from the low-altitude East Asian one. (c) K = 4 separates the Southern Indian ancestry (orange) from the Northern Indian one (yellow). (d) K = 5 separates the Northern low-altitude East Asian ancestry (dark green) from the Southern low-altitude East Asian one (green). AP = Andhra Pradesh; TN = Tamil Nadu.

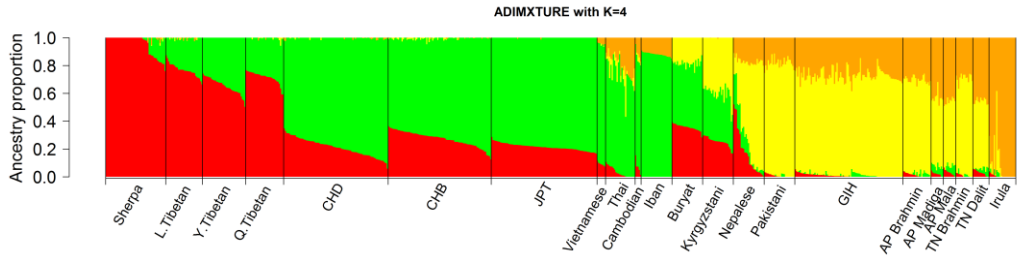
(a)



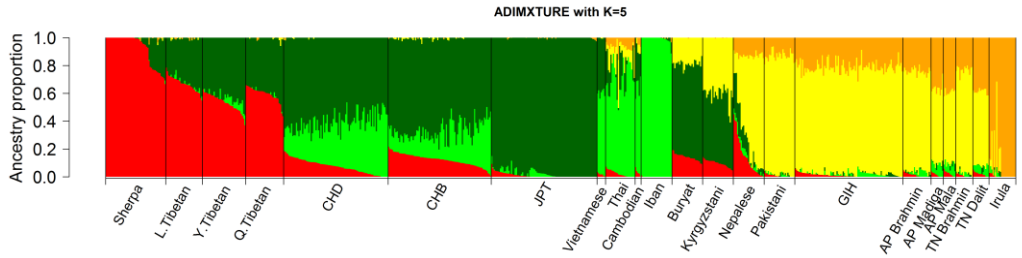
(b)



(c)

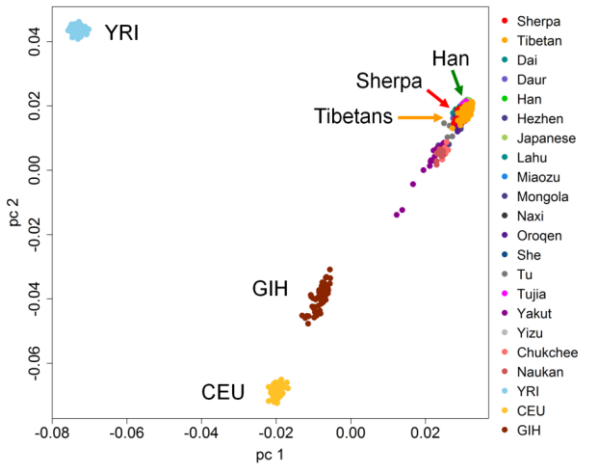


(d)

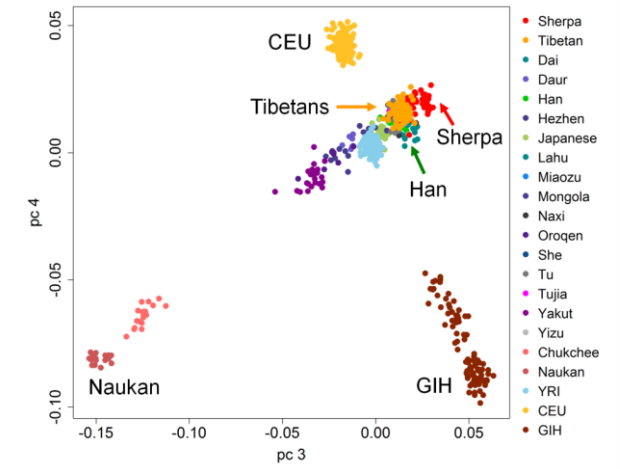


Supplementary Figure 2.3: PCA of 19 East Asian populations and three non East Asian HapMap3 populations. (a) PC1 and PC2 define a triangle with Africans, Europeans and East Asians at each vertex. Indian and Northern East Asians (including the Siberian populations) distribute across the European – East Asian cline. PC1 and PC2 explain 9.7% and 4.7% of total variation, respectively. (b) PC3 and PC4 define a triangle with Europeans, Indians and Siberian populations at each vertex. Northern East Asians (Daur, Hezhen, Mongola, Oroqen and Yakut) are pulled out in the direction of the Siberian populations. PC3 and PC4 explain 0.93% and 0.88% of total variation, respectively. YRI = Yoruba in Ibadan, Nigeria (n = 115); CEU = Utah residents with Northern and Western European ancestry from the CEPH collection (n = 115); GIH = Gujarati Indians in Huston, Texas (n = 88)

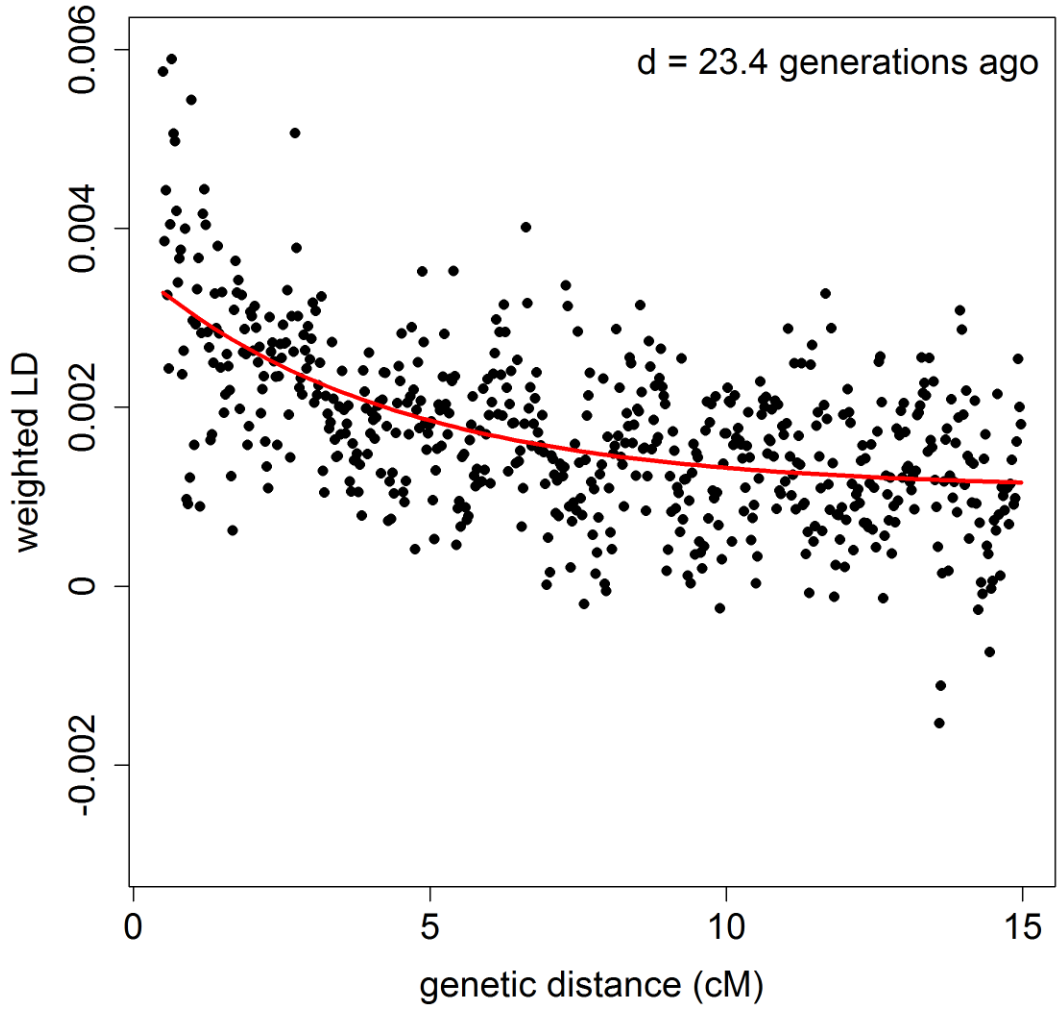
(a) PC1 vs. PC2



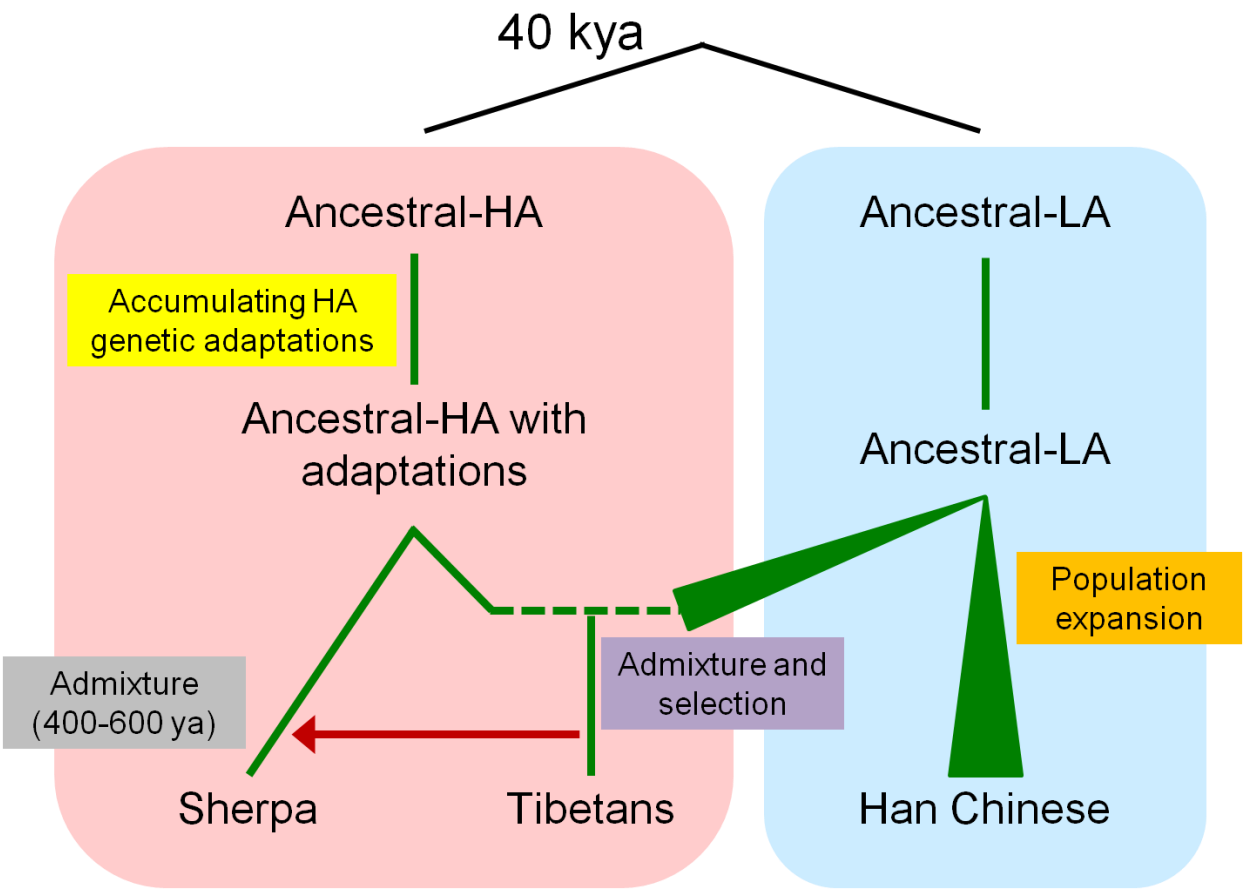
(b) PC3 vs. PC4



Supplementary Figure 2.4: LD decay in 49 unrelated Sherpa individuals after removing 20 individuals with closer relationships than first cousins. The red curve shows the exponential fit to the data.

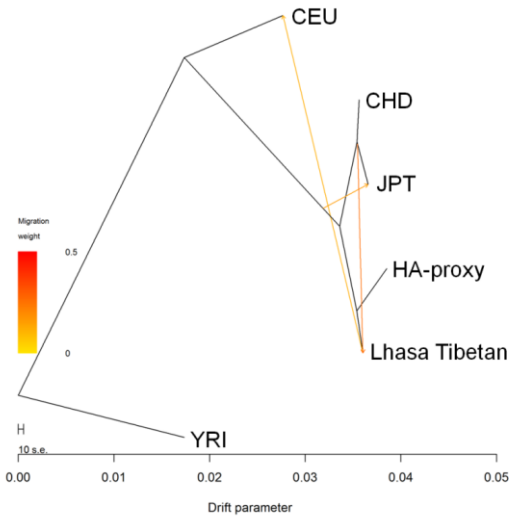


Supplementary Figure 2.5: A model of the genetic history of the Sherpa, Tibetans and Han Chinese.

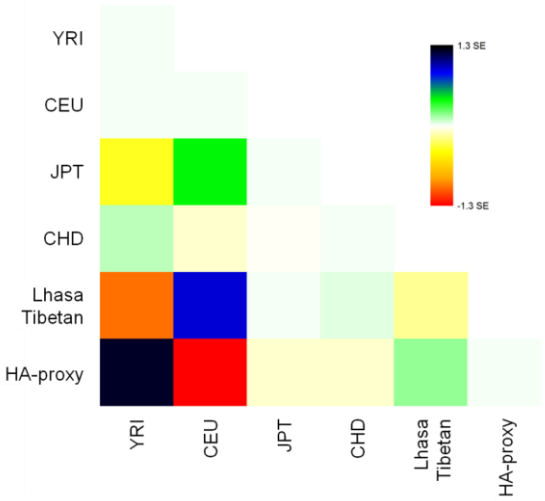


Supplementary Figure 2.6: Population trees and admixture events inferred by *TreeMix*, with each of the three Tibetan populations in addition to five other populations (the HA-proxy and HapMap3 YRI, CEU, CHD and JPT). (a, c, e) Maximum likelihood population trees and admixture events inferred for Lhasa (a), Yunnan (c), and Qinghai (e) Tibetans, respectively. We allowed three admixture events. Admixture events in Tibetans were inferred as the first (in Lhasa and Yunnan Tibetans) and the third (in Qinghai Tibetan) admixture event. (b, d, f) Residual fits from the maximum likelihood trees in (a, c, e) respectively.

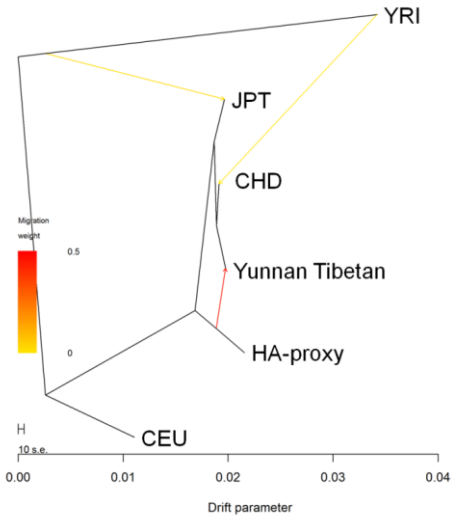
(a)



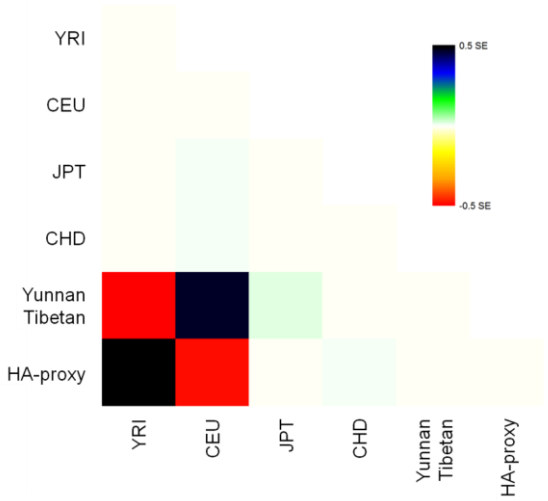
(b)



(c)

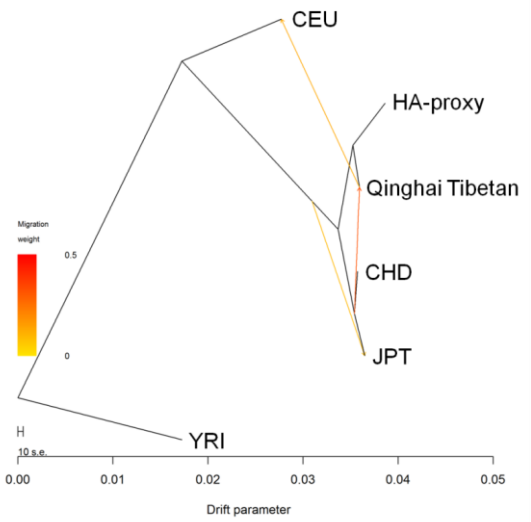


(d)

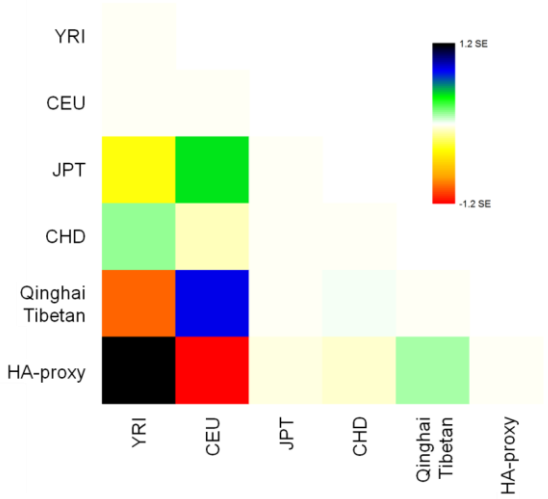


Supplementary Figure 2.6 – Continued.

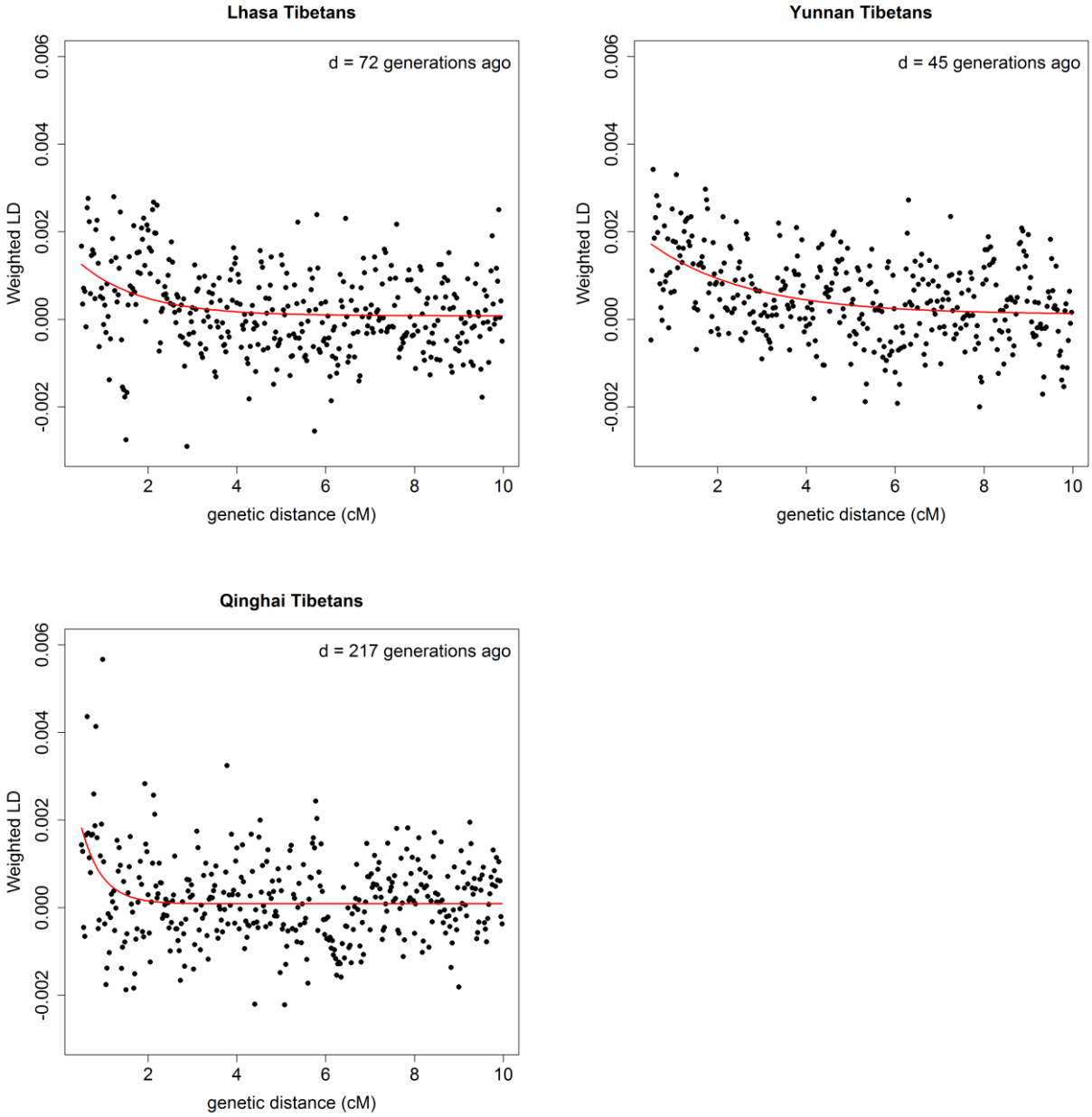
(e)



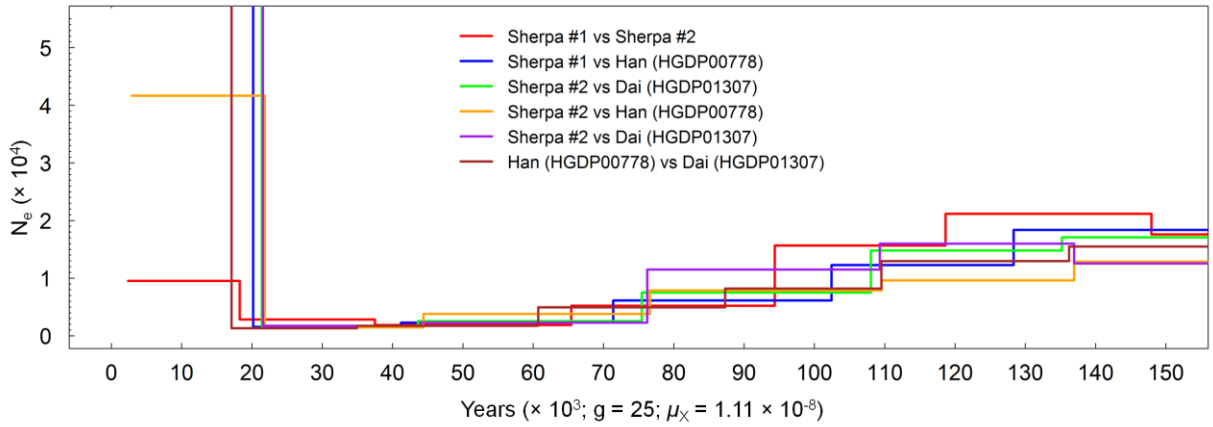
(f)



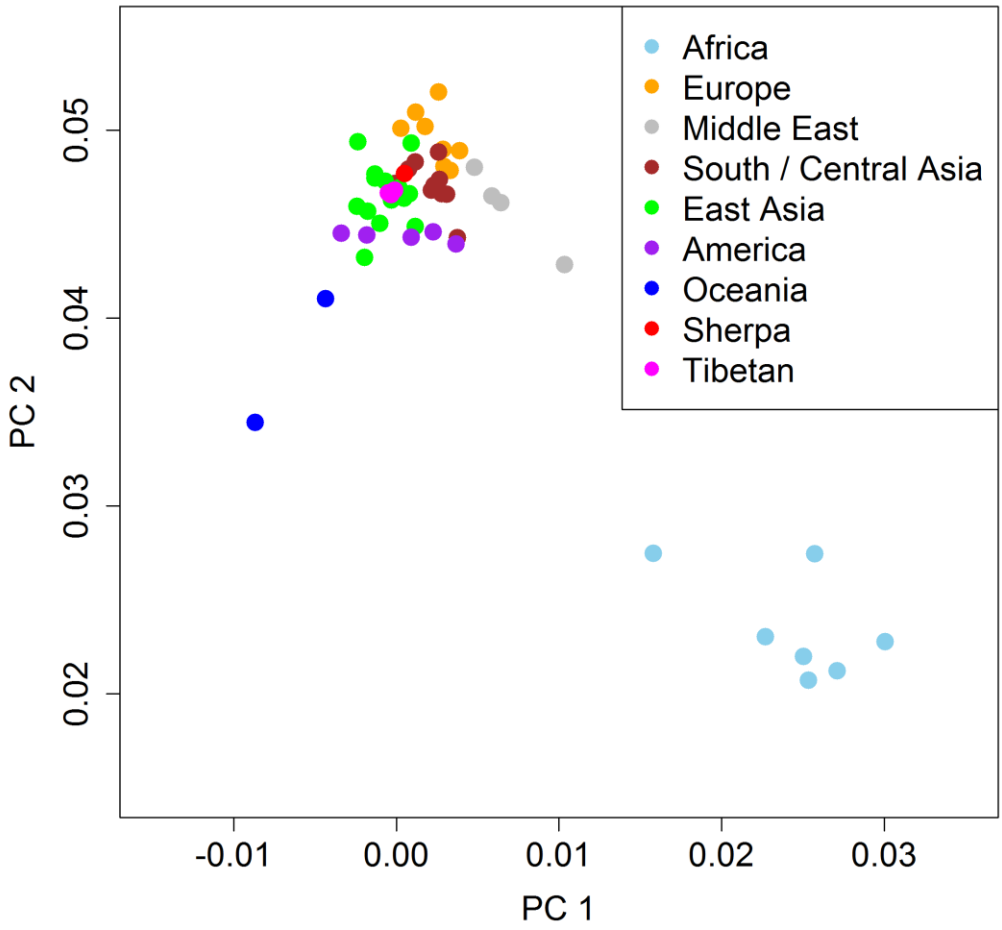
Supplementary Figure 2.7: LD decay in three Tibetan populations. Red curves show the exponential fit to the data.



Supplementary Figure 2.8: Effective population size (N_e) inferred from composite diploid X chromosome sequences (an X chromosome sequence from each of two individuals). An abrupt increase in N_e implies a split between two populations from which the X chromosomes are sampled. (g = generation time in year; μ_X = X chromosomal neutral mutation rate per base per generation)

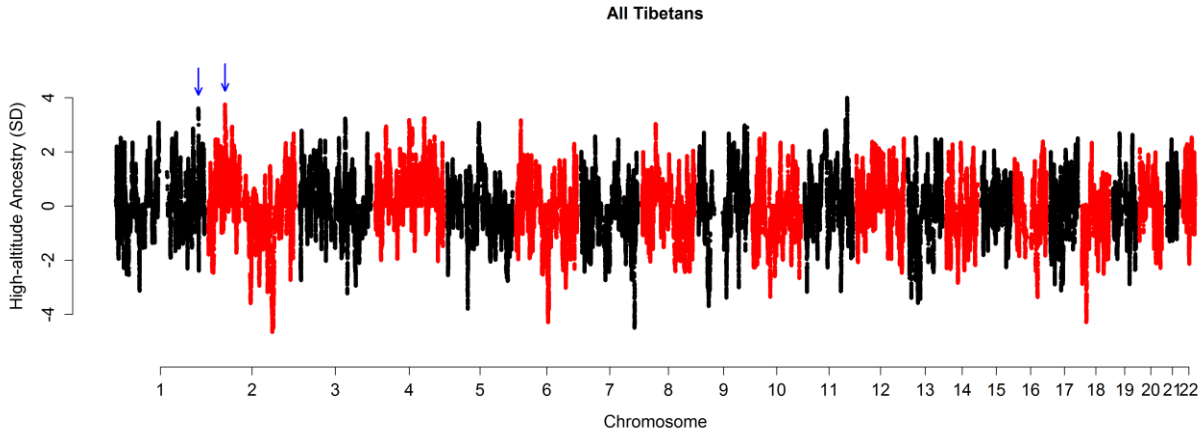


Supplementary Figure 2.9: A projection of modern human populations onto the PC plain defined by a chimpanzee, a Neanderthal and a Denisovan allele. PC 1 separates chimpanzee and archaic humans (chimpanzee > archaic humans) and PC 2 splits Neanderthal from Denisovan (Neanderthal > Denisovan). Population mean PC values of each modern human population were projected onto this plain. The HA-proxy (red) and Tibetans (magenta) cluster with other Eurasian populations. As expected, Africans show a deficit of archaic human admixture and Papuans show a higher affinity with Denisovan.

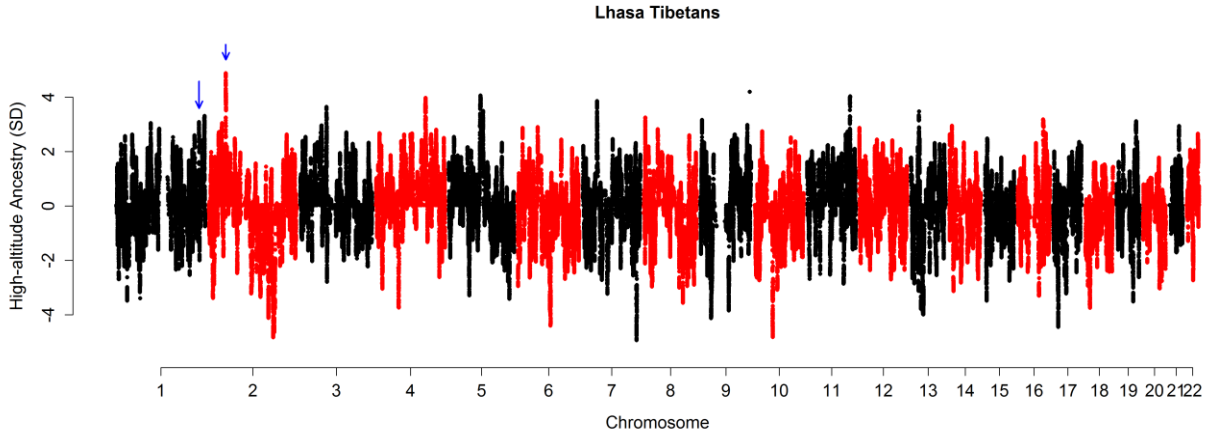


Supplementary Figure 2.10: Local ancestry estimates for Tibetan populations. (a) A merged data of all three Tibetans (a repeat of **Figure 2.4**). (b) Lhasa Tibetans. (c) Yunnan Tibetans. (d) Qinghai Tibetans. Blue arrows mark the positions of *EGLN1* (in chromosome 1) and *EPAS1* (in chromosome 2).

(a)

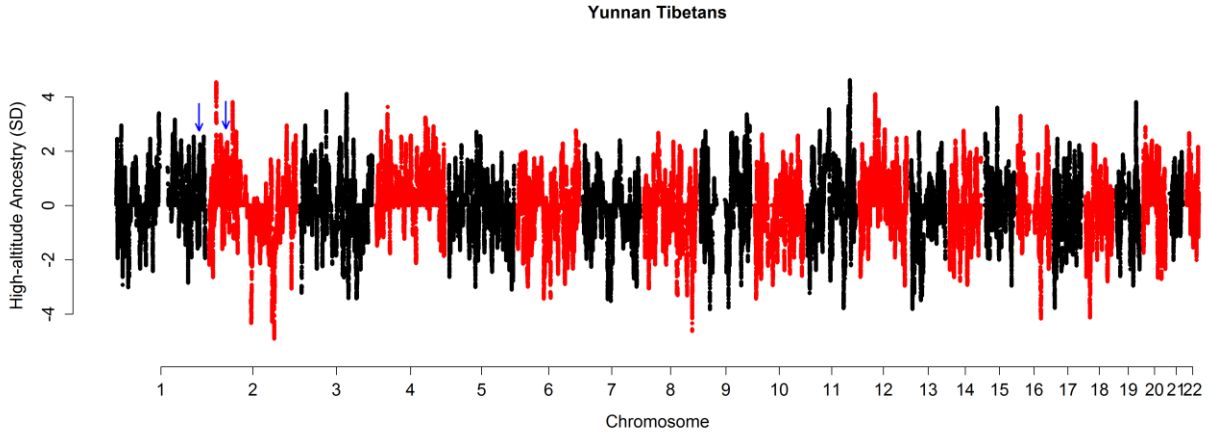


(b)

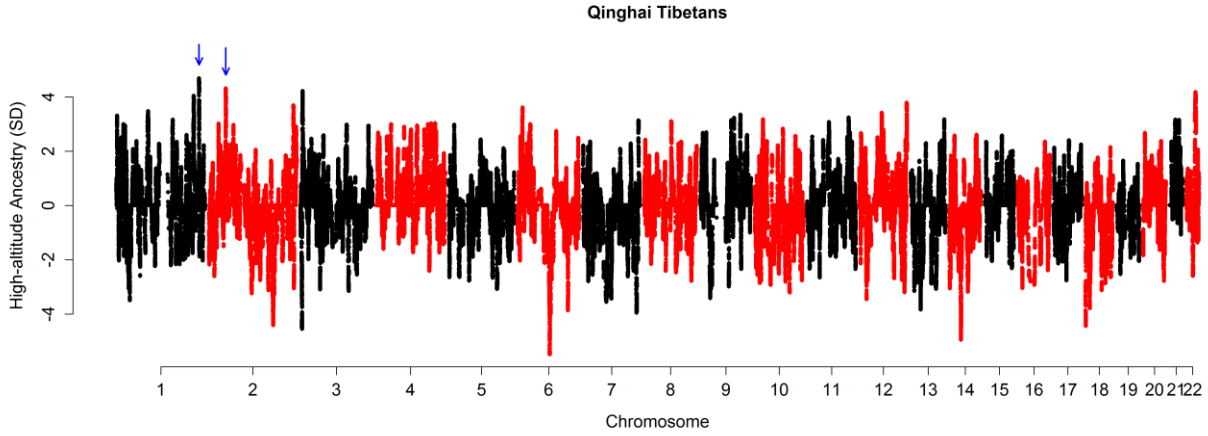


Supplementary Figure 2.10 – Continued.

(c)

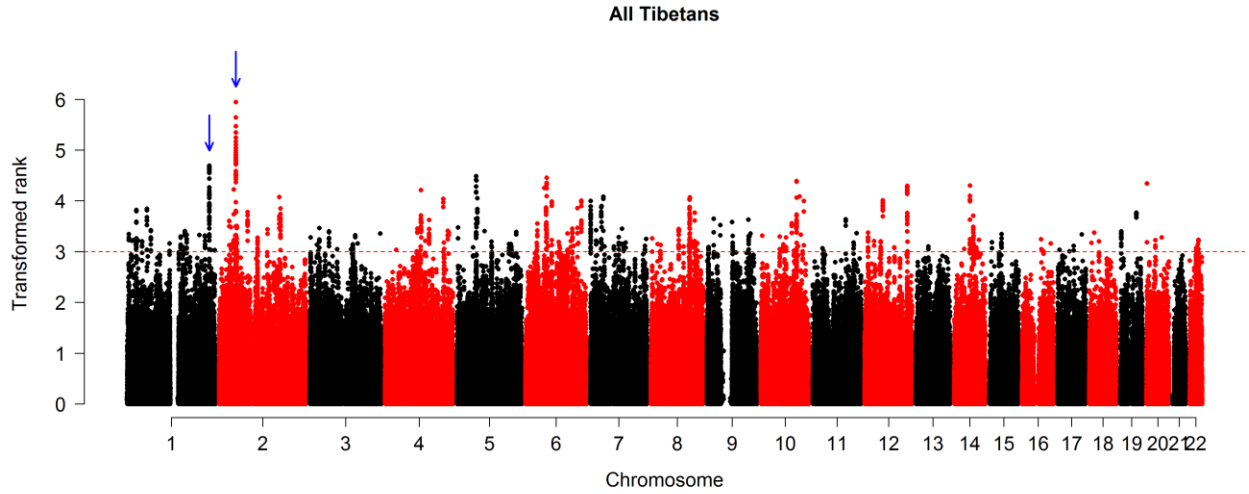


(d)

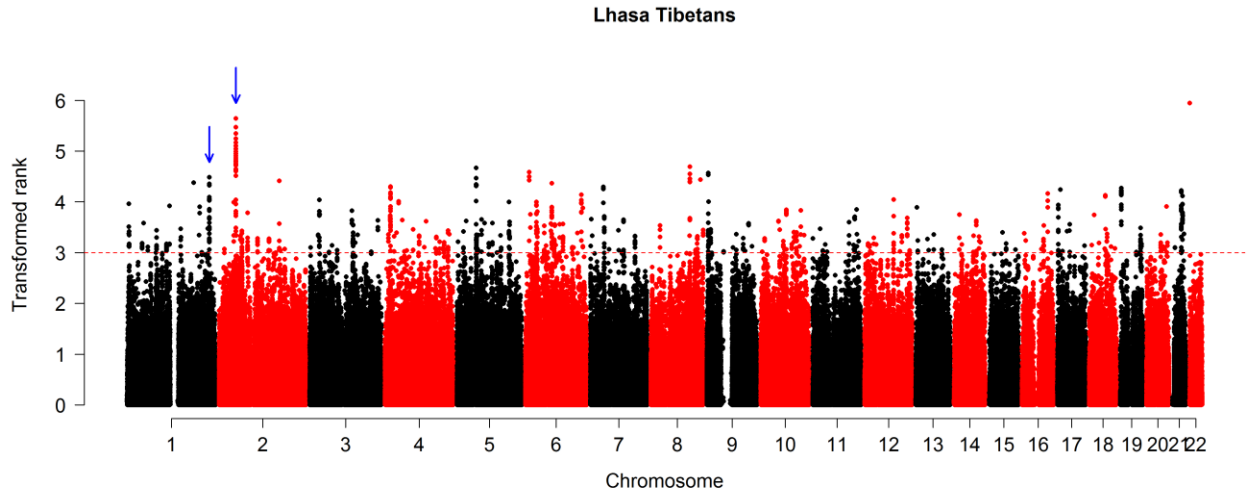


Supplementary Figure 2.11: Manhattan plots of the transformed ranks of MR scores in the Tibetan samples across the genome. (a) A merged sample of the three Tibetan samples. (b) Lhasa Tibetans. (c) Yunnan Tibetans. (d) Qinghai Tibetans. Blue arrows mark the positions of *EGLN1* (in chromosome 1) and *EPAS1* (in chromosome 2). Dotted red lines show top 0.1% cutoff.

(a)

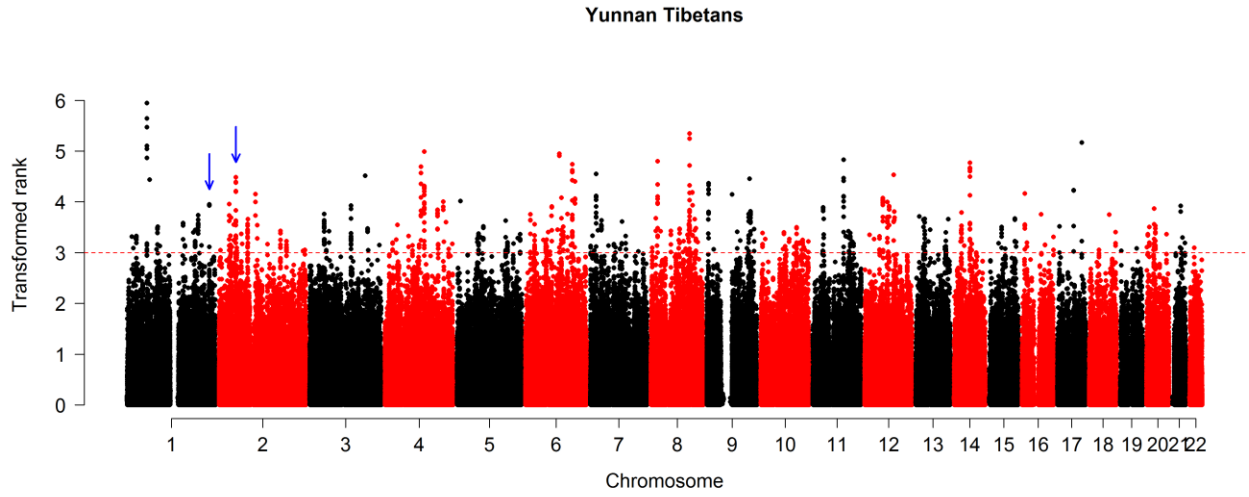


(b)

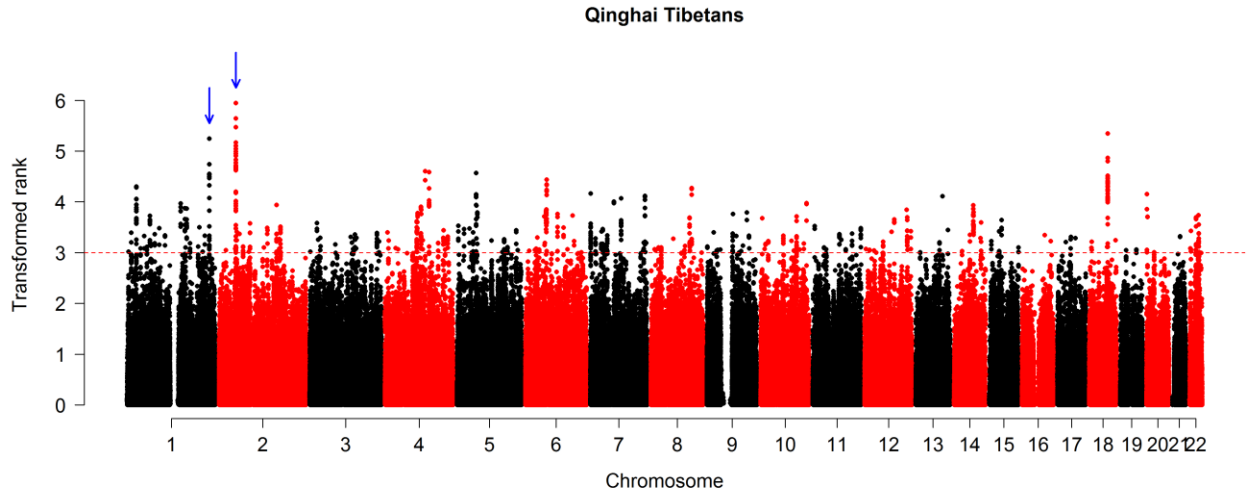


Supplementary Figure 2.11 – continued.

(c)

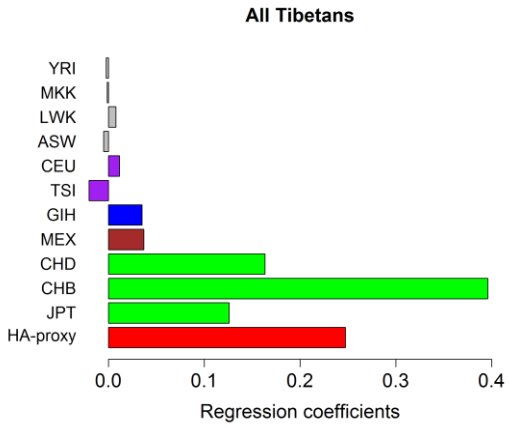


(d)

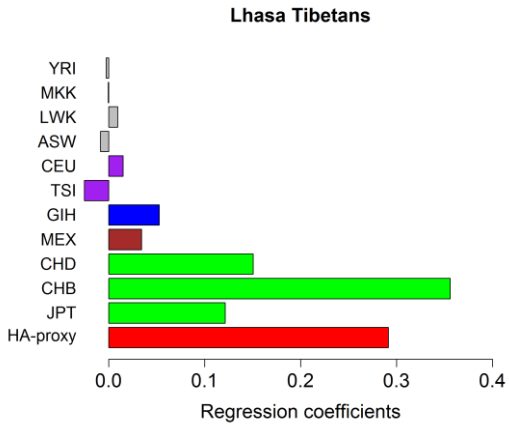


Supplementary Figure 2.12: Regression coefficients for each of the 11 HapMap3 populations and the HA-proxy in MR analysis. (a) A merged sample of the three Tibetan samples. (b) Lhasa Tibetans. (c) Yunnan Tibetans. (d) Qinghai Tibetans.

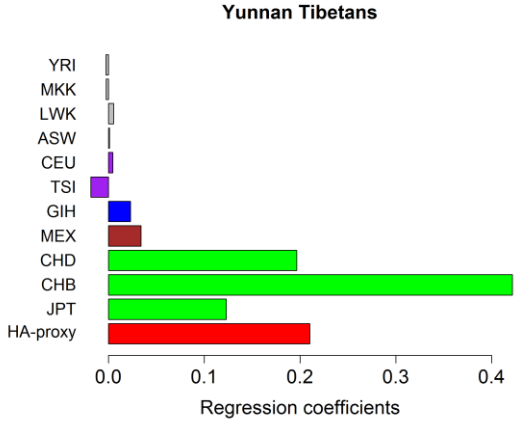
(a)



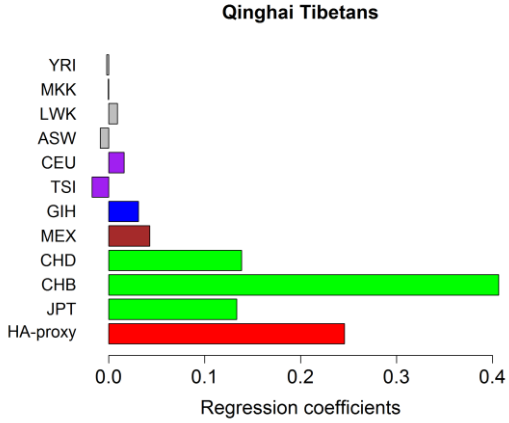
(b)



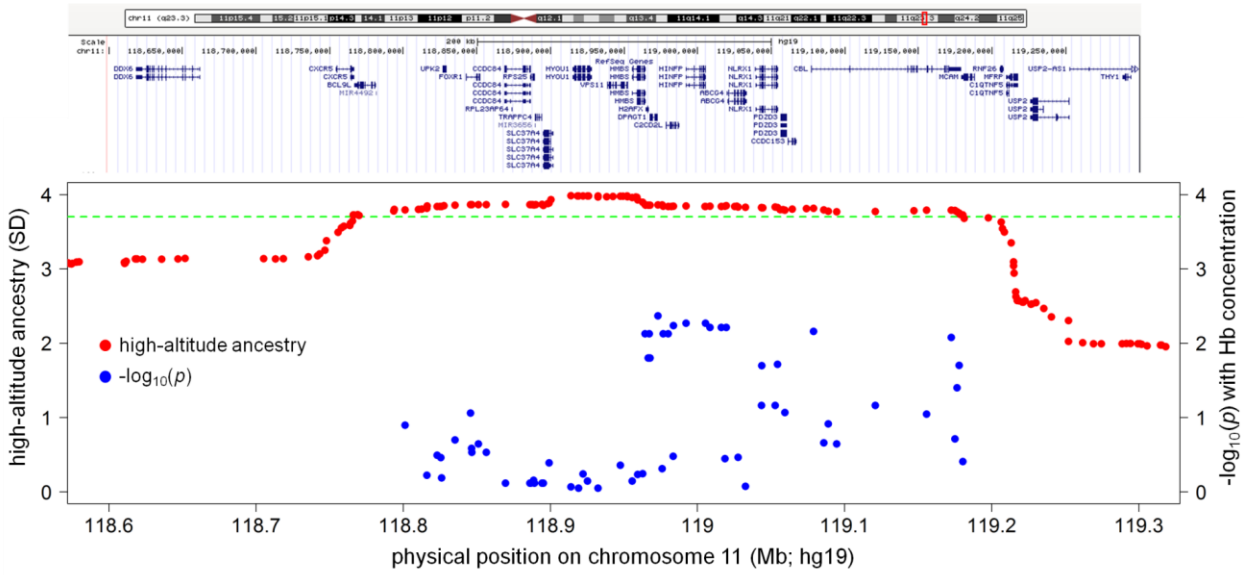
(c)



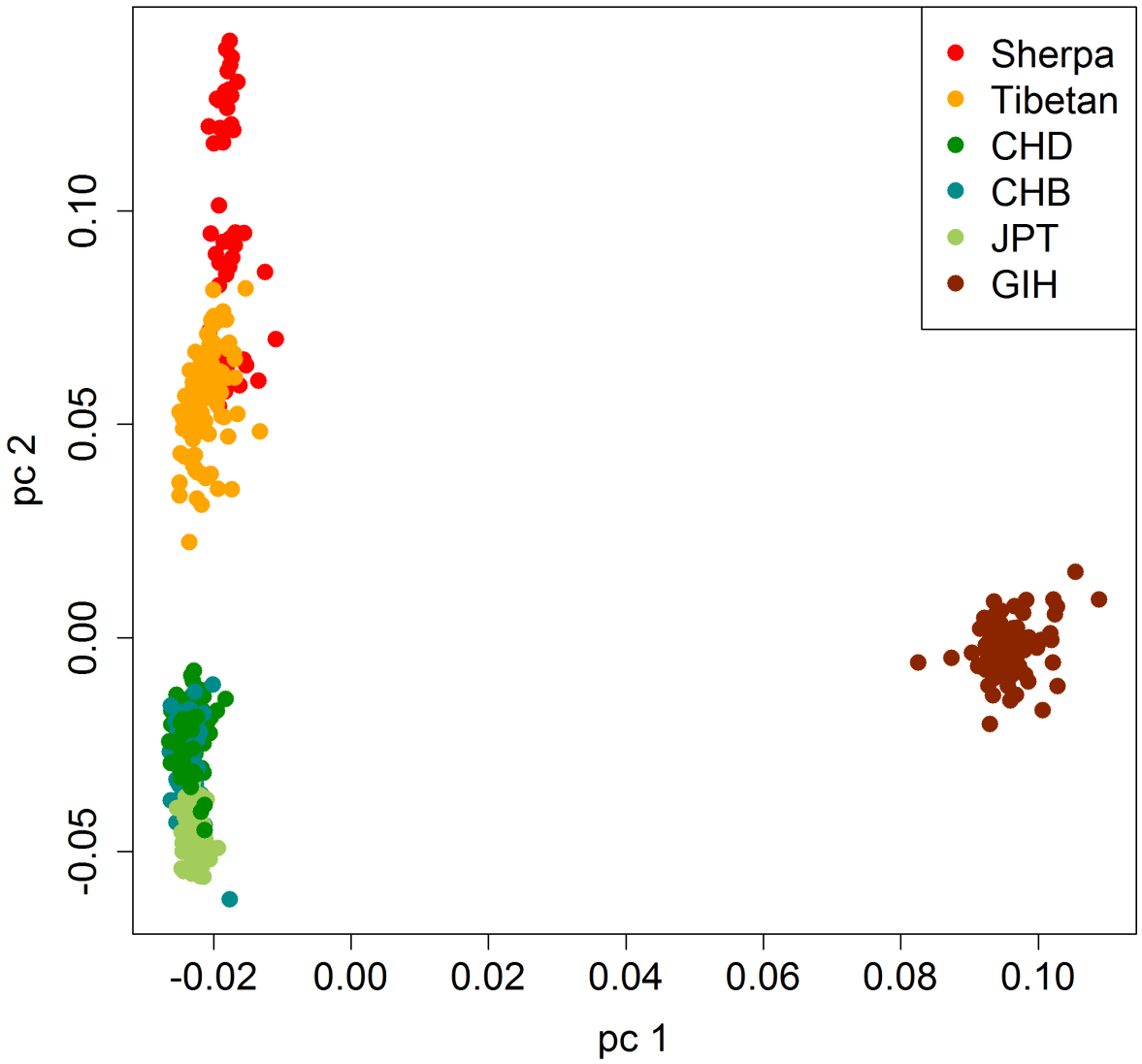
(d)



Supplementary Figure 2.13: The distribution of the high-altitude ancestry and the evidence for association between SNPs in the *HYOU1*/*HMBS* region and hemoglobin concentration. The positions of the reference sequence genes are shown in the top part of the figure. $-\log_{10}(p\text{-value})$ for the association with hemoglobin concentration of the 64 SNPs in the region with the high-altitude ancestry ≥ 3.7 SD are shown with blue dots. The green dashed line marks 3.7 SD of the high-altitude ancestry.



Supplementary Figure 2.14: PCA of the Sherpa, three Tibetan samples and 4 HapMap3 Asian populations. PC2 represents the Sherpa – Tibetan axis of genetic variation.



Supplementary Table 2.1: The 3-population (f_3) test results for the admixed Sherpa samples with Tibetan or HapMap3 East Asians (CHD, CHB or JPT) as a reference population. The admixed Sherpa individuals (n = 28) are used as a target population and the HA-proxy and Tibetans or HapMap3 East Asian populations are used as reference populations. Units are in standard deviation (SD). CHD = Chinese in Metropolis Denver, Colorado; CHB = Han Chinese in Beijing, China; JPT = Japanese in Tokyo, Japan.

Test	Reference 1	f_3 statistic (SD)
f_3(Admixed Sherpa; Reference 1, HA-proxy)	CHD	-20.449
	CHB	-21.009
	JPT	-18.455
	Lhasa Tibetan	-21.810
	Yunnan Tibetan	-20.796
	Qinghai Tibetan	-20.934

Supplementary Table 2.2: Estimates of time since admixture in three Tibetan samples and in 49 unrelated Sherpa.

Population	SNP pair distance cutoff (cM)	# generation (SD)	Z-score
Lhasa Tibetans	0.50	72.0 (51.2)	1.41
	0.60	76.8 (55.2)	1.39
	0.70	69.7 (59.2)	1.18
Yunnan Tibetans	0.50	45.0 (18.2)	2.47
	0.60	47.4 (30.2)	1.57
	0.70	39.9 (36.4)	1.10
Qinghai Tibetans	0.50	216.8 (91.9)	2.36
	0.60	382.3 (244.6)	1.56
	0.70	441.6 (not available)	0.00
Sherpa	0.50	23.4 (7.7)	3.06
	0.60	21.9 (7.8)	2.81
	0.70	20.2 (7.9)	2.57

Supplementary Table 2.3: The D-test and 3-population (f_3) test results for three Tibetan samples with HapMap3 CHD, CHB or JPT as a reference East Asian (EA) population. D statistics are from ((HA-proxy, Tibetan), (EA, YRI)). Negative D statistics suggest a gene flow from EA to Tibetans. In the 3-population test, Tibetans are used as a target population and the HA-proxy and EA are used as reference populations. Units are in standard deviation (SD).

		Tibetan sampling location		
Test	EA population	Lhasa	Yunnan	Qinghai
D(HA-proxy, Tibetan; EA, YRI))	CHD	-5.132	-9.764	-5.907
	CHB	-5.195	-9.983	-6.111
	JPT	-5.271	-9.489	-6.076
f_3(Tibetan; EA, HA-proxy)	CHD	-2.283	-3.492	-1.519
	CHB	-0.545	-1.395	-0.215
	JPT	-0.867	-0.310	-0.193

Supplementary Table 2.4: Estimates of admixture parameters for the three Tibetan populations from *MixMapper*. The scaffold tree for fitting admixture includes the HA-proxy and four HapMap3 populations: YRI, CEU, CHD and JPT.

Admixed	Ref. 1	Ref. 2	# rep.	High-altitude ancestry proportion[1)]	Drift on the HA-proxy branch (before the split / total)[1), 2)]	Drift on the CHD branch (before the split / total)[1), 3)]
Lhasa	HA-proxy	CHD	500	0.762 (0.660-0.863)	0.0017 / 0.0048 (0.0013-0.0022)	0.0006 / 0.0006 (0.0006-0.0006)
Yunnan	HA-proxy	CHD	500	0.424 (0.342-0.510)	0.0027 / 0.0048 (0.0021-0.0036)	0.0005 / 0.0006 (0.0004-0.0006)
Qinghai	HA-proxy	CHD	500	0.718 (0.640-0.812)	0.0014 / 0.0048 (0.0012-0.0017)	0.0006 / 0.0006 (0.0005-0.0006)

1) Median and 95% bootstrap confidence interval (CI) of the parameters
2) The total length of the HA-proxy branch and its length before the split with the true ancestral population of Tibetans (in drift unit)
3) The total length of the CHD branch and its length before the split with the true ancestral population of Tibetans (in drift unit)

Supplementary Table 2.5: Admixture events in Tibetan populations inferred by *TreeMix*. We inferred admixture events for each of six population sets, each including a different Tibetan population and five other populations (the HA-proxy and HapMap3 YRI, CEU, CHD and JPT). Here we summarized the results of 500 bootstrap replicates. Results for robust signals of admixture are presented: population trees and source populations with median admixture proportion within 5-95% and at least 25 (5%) bootstrap replicates.

Tibetan	Population tree[1)]	Source	# rep.[2)]	Admixture proportion[3)]
Lhasa	(YRI, (CEU, ((CHD, JPT), (HA-proxy, Tibetan))))	CHD	365	0.263 (0.192-0.322)
		JPT	51	0.299 (0.228-0.377)
		(CHD, JPT)	76	0.282 (0.217-0.337)
Yunnan	(YRI, (CEU, (HA-proxy, (JPT, (CHD, Tibetan)))))	HA-proxy	397	0.436 (0.383-0.486)
	(YRI, (CEU, (HA-proxy, (Tibetan, (CHD, JPT)))))	HA-proxy	74	0.240 (0.197-0.466)
Qinghai	(YRI, (CEU, ((CHD, JPT), (HA-proxy, Tibetan))))	CHD	321	0.315 (0.233-0.383)
		JPT	61	0.348 (0.296-0.406)
		(CHD, JPT)	115	0.342 (0.289-0.399)

1) The topology of the inferred population tree
2) The number of bootstrap replicates (among 500 replicates) with the corresponding population tree and source population of admixture
3) The proportion of admixture from the source population

Supplementary Table 2.6: A comparison of the level of admixture with archaic humans (Neanderthal and Denisovan) in the HA-proxy and Tibetans to that of other modern human populations. D < -3 SD indicates significantly more archaic admixture in the HA-proxy / Tibetans than in a HGDP population.

		HA-proxy		Lhasa Tibetan	
Continent	H_1	Neanderthal	Denisovan	Neanderthal	Denisovan
Africa	Bantu Kenya	**-6.472**	**-3.339**	**-6.823**	**-3.688**
	Biaka Pygmy	**-7.232**	**-3.468**	**-7.553**	**-3.853**
	Mandenka	**-6.614**	-2.475	**-6.908**	-2.804
	Mbuti Pygmy	**-6.138**	-2.207	**-6.387**	-2.499
	San	**-3.479**	-1.014	**-3.668**	-1.250
	Yoruba	**-7.078**	-2.810	**-7.379**	**-3.163**
	Bantu South Africa	**-5.690**	-2.023	**-5.963**	-2.344
Europe	Adygei	-0.040	-1.061	-0.281	-1.526
	Basque	-0.624	-0.945	-0.889	-1.373
	French	0.051	-0.568	-0.188	-0.993
	Orcadian	0.374	-0.455	0.180	-0.844
	Russian	0.525	-0.130	0.317	-0.544
	Sardinian	-0.486	-0.506	-0.750	-0.897
	Tuscan	-0.374	-0.674	-0.606	-1.072
	Italian	-0.600	-0.634	-0.870	-1.063
Middle East	Bedouin	-1.610	-1.384	-1.960	-1.853
	Druze	-0.990	-1.122	-1.330	-1.563
	Mozabite	-2.852	-1.941	**-3.295**	-2.431
	Palestinian	-1.511	-1.219	-1.867	-1.706
South / Central Asia	Balochi	-0.266	-0.706	-0.551	-1.179
	Brahui	-0.471	-0.337	-0.775	-0.797
	Burusho	-0.686	-0.576	-1.047	-1.103
	Hazara	0.307	0.037	-0.024	-0.574
	Kalash	0.109	-0.245	-0.115	-0.649
	Makrani	-1.200	-0.343	-1.573	-0.798
	Pathan	-0.373	-0.298	-0.670	-0.772
	Sindhi	-0.671	-0.407	-1.009	-0.896
	Uygur	-0.633	-0.586	-1.079	-1.233
	Xibo	0.487	0.556	0.165	0.008

Supplementary Table 2.6 – Continued.

		HA-proxy		Lhasa Tibetan	
Continent	H_1	Neanderthal	Denisovan	Neanderthal	Denisovan
East Asia	Cambodian	0.144	0.162	-0.202	-0.388
	Dai	0.631	1.808	0.380	1.652
	Daur	0.877	1.567	0.641	1.215
	Han	0.029	0.457	-0.516	-0.187
	Northern Han	0.304	-0.407	-0.074	-1.273
	Hezhen	0.138	0.391	-0.238	-0.142
	Japanese	-0.011	0.218	-0.515	-0.449
	Lahu	0.470	1.175	0.223	0.827
	Miao	1.867	0.966	2.009	0.558
	Mongola	1.313	1.681	1.256	1.434
	Naxi	0.522	0.668	0.194	0.122
	Oroqen	0.606	0.699	0.348	0.222
	She	0.310	0.412	-0.015	-0.070
	Tu	-0.471	0.482	-1.083	-0.133
	Tujia	0.391	0.802	0.054	0.318
	Yakut	1.036	0.912	0.868	0.470
	Yi	1.326	0.806	1.234	0.302
America	Colombian	-0.330	0.278	-0.568	-0.042
	Karitiana	0.394	0.950	0.213	0.680
	Maya	-1.281	-0.417	-1.657	-0.888
	Pima	-0.754	0.028	-1.020	-0.313
	Surui	0.579	1.253	0.436	1.031
Oceania	Melanesian	0.637	2.088	0.478	1.858
	Papuan	1.168	**3.631**	1.038	**3.467**

Supplementary Table 2.6 – Continued.

Continent	H_1	Yunnan Tibetan		Qinghai Tibetan	
		Neanderthal	Denisovan	Neanderthal	Denisovan
Africa	Bantu Kenya	**-6.889**	**-3.698**	**-6.811**	**-3.687**
	Biaka Pygmy	**-7.599**	**-3.865**	**-7.506**	**-3.832**
	Mandenka	**-6.972**	-2.822	**-6.890**	-2.793
	Mbuti Pygmy	**-6.448**	-2.510	**-6.383**	-2.489
	San	**-3.714**	-1.283	**-3.686**	-1.264
	Yoruba	**-7.460**	**-3.175**	**-7.356**	**-3.154**
	Bantu South Africa	**-6.033**	-2.365	**-5.959**	-2.333
Europe	Adygei	-0.344	-1.587	-0.331	-1.586
	Basque	-0.950	-1.434	-0.936	-1.424
	French	-0.250	-1.065	-0.237	-1.057
	Orcadian	0.123	-0.910	0.134	-0.897
	Russian	0.256	-0.624	0.266	-0.607
	Sardinian	-0.811	-0.963	-0.790	-0.946
	Tuscan	-0.668	-1.142	-0.650	-1.124
	Italian	-0.936	-1.128	-0.912	-1.117
Middle East	Bedouin	-2.037	-1.912	-2.017	-1.904
	Druze	-1.390	-1.620	-1.365	-1.608
	Mozabite	**-3.386**	-2.504	**-3.328**	-2.459
	Palestinian	-1.938	-1.763	-1.907	-1.738
South / Central Asia	Balochi	-0.621	-1.254	-0.601	-1.226
	Brahui	-0.839	-0.876	-0.819	-0.850
	Burusho	-1.124	-1.185	-1.104	-1.175
	Hazara	-0.127	-0.707	-0.105	-0.659
	Kalash	-0.176	-0.734	-0.163	-0.711
South / Central Asia (continued)	Makrani	-1.625	-0.870	-1.595	-0.848
	Pathan	-0.743	-0.859	-0.723	-0.836
	Sindhi	-1.077	-0.971	-1.054	-0.948
	Uygur	-1.194	-1.346	-1.149	-1.325
	Xibo	0.055	-0.146	0.078	-0.092

Supplementary Table 2.6 – Continued.

Continent	H_1	Yunnan Tibetan		Qinghai Tibetan	
		Neanderthal	Denisovan	Neanderthal	Denisovan
East Asia	Cambodian	-0.311	-0.546	-0.270	-0.457
	Dai	0.307	1.611	0.301	1.555
	Daur	0.610	1.211	0.599	1.214
	Han	-0.727	-0.401	-0.621	-0.304
	Northern Han	-0.203	-1.459	-0.167	-1.327
	Hezhen	-0.347	-0.288	-0.328	-0.242
	Japanese	-0.719	-0.676	-0.629	-0.575
	Lahu	0.143	0.736	0.151	0.733
	Miao	1.975	0.423	1.814	0.446
	Mongola	1.165	1.307	1.161	1.338
	Naxi	0.083	-0.031	0.097	0.020
	Oroqen	0.257	0.096	0.265	0.135
	She	-0.128	-0.210	-0.095	-0.150
	Tu	-1.309	-0.301	-1.186	-0.243
	Tujia	-0.060	0.174	-0.035	0.211
	Yakut	0.804	0.343	0.796	0.390
	Yi	1.222	0.152	1.138	0.194
America	Colombian	-0.627	-0.115	-0.611	-0.090
	Karitiana	0.163	0.614	0.174	0.640
	Maya	-1.692	-0.973	-1.716	-0.947
	Pima	-1.057	-0.386	-1.056	-0.362
	Surui	0.385	0.961	0.397	0.994
Oceania	Melanesian	0.431	1.791	0.425	1.798
	Papuan	1.015	**3.453**	0.996	**3.424**

Supplementary Table 2.7: The Sherpa participants

Traits	Male[†]	Female[†]
Sample size	29	40
Age (year)	34.9 (11.0)	33.0 (9.3)
Weight (kg)	59.3 (4.8)	56.3 (7.2)
Height (m)	1.66 (0.06)	1.55 (0.06)
BMI (kg / m^2)	21.6 (1.7)	23.5 (2.7)
Hemoglobin (g / dl)	17.2 (1.4)	15.2 (1.0)

[†] The reported values are the mean and the standard deviation (in parenthesis)

Supplementary Table 2.8: SNPs around *EGLN1* and *EPAS1* genes with top PBS signals in Tibetans.

		Tibetan		HA-proxy		Allele frequency		
SNPs	Genes	PBS	Rank*	PBS	Rank*	HA-proxy	Tibetan	CHD
rs6679627	*EGLN1*	0.578	4	0.955	45	0.738	0.676	0.253
rs982414	*EPAS1*	0.754	1	1.676	1	0.619	0.532	0.006

*rank out of 879,434 SNPs

Supplementary Table 2.9: Association test results of 26 SNPs in *EPAS1* gene region in the Sherpa with hemoglobin concentration (g/dL)

SNP	Position[1)]	A_1	A_2[2)]	Freq (A_2)	beta[3)]	p-value	Direction[4)]	Deviation in Local ancestry (SD)[5)]
rs2121266	46535924	A	C	0.162	0.313	0.190	same	3.170
rs17034950	46538794	G	A	0.162	0.313	0.190	same	-
rs9973653	46548109	G	T	0.368	0.239	0.278	same	3.293
rs4953342	46552047	A	G	0.118	0.439	0.180	same	3.399
rs4953353	**46567276**	**G**	**T**	**0.140**	**0.635**	**0.026**	**same**	**3.644**
rs4953354	46575388	A	G	0.055	0.121	0.794	same	-
rs6715787	**46576172**	**C**	**G**	**0.119**	**0.651**	**0.011**	**same**	-
rs17035010	**46576488**	**C**	**T**	**0.118**	**0.650**	**0.011**	**same**	-
rs17035013	**46576573**	**T**	**C**	**0.119**	**0.650**	**0.011**	**same**	-
rs6544887	**46577212**	**T**	**C**	**0.169**	**0.595**	**0.012**	**same**	-
rs6756667	**46579409**	**A**	**G**	**0.169**	**0.595**	**0.012**	**same**	**3.697**
rs6712143	**46582891**	**A**	**G**	**0.118**	**0.650**	**0.011**	**same**	-
rs7583554	**46587097**	**T**	**C**	**0.119**	**0.653**	**0.011**	**same**	-
rs10206434	**46593536**	**A**	**G**	**0.213**	**0.450**	**0.041**	**same**	-
rs1374749	**46596433**	**G**	**A**	**0.184**	**0.455**	**0.044**	**same**	**3.717**
rs1992846	46597581	C	T	0.180	0.371	0.128	same	-
rs7565341	**46599030**	**G**	**A**	**0.187**	**0.443**	**0.049**	**same**	-
rs12467821	**46600894**	**C**	**T**	**0.226**	**0.425**	**0.044**	**same**	-
rs11675441	**46601496**	**T**	**C**	**0.227**	**0.424**	**0.044**	**same**	-
rs7583088	**46603165**	**G**	**A**	**0.228**	**0.420**	**0.046**	**same**	-
rs11678465	**46603260**	**T**	**C**	**0.228**	**0.420**	**0.046**	**same**	-
rs7557402	46603671	G	C	0.205	0.378	0.086	same	-
rs7594278	**46604593**	**T**	**G**	**0.218**	**0.449**	**0.032**	**same**	-
rs7571218	**46605659**	**G**	**A**	**0.199**	**0.473**	**0.029**	**same**	**3.740**
rs13006131	**46608542**	**G**	**C**	**0.212**	**0.428**	**0.041**	**same**	-
rs7590087	**46610680**	**C**	**A**	**0.198**	**0.473**	**0.029**	**same**	-

1) Physical position of SNPs in hg19 (chromosome 2)
2) Minor allele in 69 Sherpa samples
3) Per allele effect size of the minor allele (A_2) in comparison to the major allele (A_1)
4) "Same" means the allelic direction of association is same as that reported in Tibetans
5) Mean high-altitude ancestry across all 96 Tibetans (in standard deviation; only the SNPs with a direct estimate of their local ancestry from HAPMIX were shown)

Supplementary Table 2.10: The number of SNPs with nominal p < 0.05 among 26 *EPAS1* SNPS in 1,000 permutations

# SNPs	# permutations	# permutations (cumulative)
0	769	769
1	100	869
2	32	901
3	17	918
4	11	929
5	13	942
6	6	948
7	15	963
8	2	965
9	5	970
10	4	974
11	4	978
12	4	982
13	3	985
14	5	990
15	1	991
16	2	993
17	1	994
18	3	997
19	**1**	**998**
20	**0**	**998**
21	**0**	**998**
22	**1**	**999**
23	**1**	**1,000**

Supplementary Table 2.11: Genes in the two Reactome pathway gene sets for the HIF pathway

Gene Set	Genes			
Cellular response to hypoxia (n = 25)	*ARNT*	*CA9*	*CREBBP*	*CUL2*
	EGLN1	*EGLN2*	*EGLN3*	*EP300*
	EPAS1	*EPO*	*HIF1A*	*HIF1AN*
	HIF3A	*RBX1*	*RPS27A*	*TCEB1*
	TCEB2	*UBA52*	*UBB*	*UBC*
	UBE2D1	*UBE2D2*	*UBE2D3*	*VEGFA*
	VHL			
Oxygen-dependent Proline Hydroxylation of Hypoxia-inducible Factor Alpha (n = 18)	*CUL2*	*EGLN1*	*EGLN2*	*EGLN3*
	EPAS1	*HIF1A*	*HIF3A*	*RBX1*
	RPS27A	*TCEB1*	*TCEB2*	*UBA52*
	UBB	*UBC*	*UBE2D1*	*UBE2D2*
	UBE2D3	*VHL*		

Supplementary Table 2.12: The mean proportion of top 0.5, 1.0 and 5.0% high-altitude ancestry SNPs within 10 kb of the genes in the two Reactome pathway gene sets in the merged sample of all three Tibetan samples

	Top proportion		
Gene Set	0.005	0.010	0.050
Cellular response to hypoxia	0.224*	0.279**	0.299*
Cellular response to hypoxia (excluding *EGLN1* and *EPAS1*)	0.059	0.124*	0.146
Oxygen-dependent Proline Hydroxylation of Hypoxia-inducible Factor Alpha	0.356*	0.382**	0.408**
Oxygen-dependent Proline Hydroxylation of Hypoxia-inducible Factor Alpha (excluding *EGLN1* and *EPAS1*)	0.106	0.139*	0.180†

†, * and ** denote support from ≥ 90, 95 and 99% of bootstrap replicates, respectively.

Supplementary Table 2.13: Association test results of 64 SNPs in the *HYOU1*/*HMBS* gene region in the Sherpa with hemoglobin concentration (g/dL)

SNP	Position[1)]	A_1	A_2[2)]	Freq (A_2)	beta[3)]	p-value	Deviation in Local ancestry (SD)[4)]
rs602716	118801235	G	A	0.181	-0.396	0.126	3.793
rs587263	118816156	C	T	0.500	0.100	0.596	3.813
rs663003	118823148	C	T	0.312	0.201	0.321	3.841
rs1790191	118825636	A	G	0.309	0.191	0.345	3.841
rs12797009	118826192	C	A	0.246	0.097	0.650	3.839
rs11607835	118835258	T	C	0.275	0.281	0.200	3.855
rs4442562	118845813	G	A	0.239	0.392	0.087	3.865
rs4938603	118846482	A	G	0.399	0.209	0.258	3.865
rs4938604	118846574	T	C	0.449	0.182	0.294	3.863
rs4456264	118851138	C	T	0.384	0.226	0.226	3.865
rs3889526	118856491	C	A	0.449	0.182	0.294	3.864
rs3737504	118869460	A	G	0.094	-0.111	0.763	3.866
rs11822103	118869539	G	A	0.094	-0.111	0.763	3.866
rs10111	118886117	C	T	0.094	-0.111	0.763	3.865
rs7131534	118886656	G	A	0.094	-0.111	0.763	3.865
rs11602764	118888619	A	G	0.159	-0.123	0.695	3.860
rs11217125	118889185	T	C	0.094	-0.111	0.763	3.861
rs569	118894287	A	G	0.094	-0.111	0.763	3.866
rs11006	118895202	A	G	0.094	-0.111	0.763	3.852
rs4936459	118899218	C	T	0.072	0.330	0.406	3.886
rs1003081	118913993	C	T	0.072	0.071	0.856	3.983
rs568922	118919206	C	T	0.058	0.060	0.890	3.982
rs1804690	118922200	G	A	0.254	0.134	0.575	3.979
rs538478	118925341	A	G	0.094	0.132	0.712	3.981
rs582688	118932320	T	C	0.058	0.060	0.890	3.980
rs636283	118932445	C	T	0.058	0.060	0.890	3.969
rs540261	118947634	T	C	0.101	0.269	0.437	3.977
rs589925	118955679	T	C	0.094	0.132	0.712	3.961
rs17075	118959331	G	A	0.123	0.178	0.584	3.929
rs1784304	118962816	A	C	0.478	0.111	0.568	3.896
rs640603	**118964330**	**G**	**A**	**0.449**	**-0.508**	**0.007**	**3.864**
rs2509049	**118966521**	**T**	**C**	**0.435**	**0.441**	**0.016**	**3.857**
rs7759	**118967291**	**A**	**G**	**0.449**	**-0.508**	**0.007**	**3.857**
rs643788	**118967758**	**C**	**T**	**0.435**	**0.441**	**0.016**	**3.856**
rs10790282	**118973133**	**G**	**A**	**0.268**	**0.556**	**0.004**	**3.861**

Supplementary Table 2.13 – Continued

SNP	Position[1)]	A_1	A_2[2)]	Freq (A_2)	beta[3)]	p-value	Deviation in Local ancestry (SD)[4)]
rs10790283	118975927	G	A	0.210	-0.168	0.487	3.862
rs639373	**118976719**	**G**	**A**	**0.449**	**-0.508**	**0.007**	**3.843**
rs510435	**118980039**	**C**	**T**	**0.449**	**-0.508**	**0.007**	**3.839**
rs2239896	118983434	C	T	0.094	0.314	0.333	3.843
rs682795	**118983681**	**C**	**T**	**0.362**	**0.517**	**0.006**	**3.843**
rs583713	**118992177**	**T**	**C**	**0.370**	**0.526**	**0.005**	**3.846**
rs7234	**119005485**	**T**	**C**	**0.370**	**0.526**	**0.005**	**3.839**
rs11217164	**119008538**	**A**	**G**	**0.246**	**0.537**	**0.006**	**3.844**
rs671872	**119016176**	**G**	**A**	**0.246**	**0.537**	**0.006**	**3.839**
rs3809046	119018766	G	T	0.103	0.293	0.356	3.844
rs4301800	**119019520**	**A**	**C**	**0.246**	**0.537**	**0.006**	**3.850**
rs4148170	119027545	G	A	0.101	0.300	0.343	3.832
rs3802885	119032689	A	C	0.123	0.062	0.837	3.830
rs644252	119043500	G	T	0.232	0.388	0.068	3.824
rs561830	**119043745**	**C**	**T**	**0.428**	**-0.444**	**0.020**	**3.823**
rs4245191	119052826	A	C	0.232	0.388	0.068	3.831
rs10790286	**119054488**	**T**	**C**	**0.341**	**0.461**	**0.019**	**3.824**
rs1815811	119059404	A	G	0.246	0.353	0.086	3.791
rs11217183	**119078886**	**A**	**G**	**0.420**	**-0.519**	**0.007**	**3.815**
rs6589722	119085783	A	C	0.268	0.255	0.218	3.793
rs1893032	119088883	G	A	0.319	0.325	0.121	3.777
rs11217191	119094647	G	A	0.202	0.311	0.225	3.769
rs7937454	119120804	T	C	0.232	0.388	0.068	3.771
rs2298650	119155618	G	T	0.312	0.353	0.090	3.791
rs1047417	**119172536**	**A**	**G**	**0.419**	**-0.508**	**0.008**	**3.788**
rs2509671	119175075	A	C	0.370	0.260	0.193	3.785
rs2511844	**119176398**	**T**	**C**	**0.181**	**0.448**	**0.040**	**3.767**
rs11217234	**119177938**	**A**	**G**	**0.471**	**-0.414**	**0.020**	**3.752**
rs7914	119180316	G	A	0.309	0.181	0.390	3.725

1) Physical position of SNPs in hg19 (chromosome 11)
2) Minor allele in 69 Sherpa samples
3) Per allele effect size of the minor allele (A_2) in comparison to the major allele (A_1)
4) Mean high-altitude ancestry across all 96 Tibetans (in standard deviation; only the SNPs with a direct estimate of their local ancestry from HAPMIX were shown)

Supplementary Table 2.14: The number of SNPs with nominal p < 0.05 among 64 *HYOU1*/*HMBS* region SNPS in 1,000 permutations

# SNPs	# permutations	# permutations (cumulative)
0	279	279
1	204	483
2	100	583
3	41	624
4	68	692
5	57	749
6	48	797
7	41	838
8	38	876
9	22	898
10	20	918
11	10	928
12	14	942
13	17	959
14	10	969
15	8	977
16	6	983
17	4	987
18	1	988
19	**2**	**990**
20	**1**	**991**
21	**1**	**992**
22	**5**	**997**
23	**1**	**998**
24	**1**	**999**
25	**1**	**1,000**

Supplementary Table 2.15: Estimated selection coefficient of *EGLN1* and *EPAS1* SNPs in the HA-proxy.

Chromosome	Position (hg19)	SNP	Selection coefficient	Allele frequency	
				HA-proxy	HapMap3 East Asians[1)]
1	**231341569**	**rs6679627**	**0.0012**	**0.7381**	**0.2431**
1	231723425	rs16854592	0.0004	0.4762	0.3098
1	231727094	rs2275279	0.0004	0.4762	0.3078
1	231730693	rs12058117	0.0004	0.4762	0.3157
1	231828041	rs7416743	0.0006	0.3571	0.1667
2	**46845576**	**rs982414**	**0.0023**	**0.6190**	**0.0216**
2	46848446	rs6735530	0.0018	0.6667	0.0627
2	46874040	rs7599097	0.0019	0.4048	0.0196

1) The mean allele frequency of HapMap3 East Asians (CHD, CHB and JPT)

Supplementary Table 2.16: The Sherpa and Tibetan genotype data sets

Data sets	Sample size	Genotyping platform	Tissue of DNA extraction	Altitude (m)	Location
Sherpa	69	Illumina HumanOmni1-Quad	Saliva	3,800	Solukhumbu, Nepal
Lhasa Tibetan	30	Illumina Human 1M-Duo v3	Lymphoblastoid cell line	3,700	Tibetan Autonomous Region, China
Yunnan Tibetan	35	Illumina Human610-Quad	Saliva	3,200-3,500	Yunnan province, China
Qinghai Tibetan	31	Affymetrix Genome-wide Human SNP Array 6.0	Blood	4,350	Qinghai province, China

Supplementary Table 2.17: Eleven populations in the HapMap3 data set

Abbreviation	Description	Sample Size
YRI	Yoruba in Ibadan, Nigeria	115
MKK	Maasai in Kinyawa, Kenya	143
LWK	Luhya in Webuye, Kenya	90
ASW	African ancestry in Southwest USA	54
CEU	Utah residents with Northern and Western European ancestry from the CEPH collection	115
TSI	Toscani in Italia	88
GIH	Gujarati Indians in Houston, Texas	88
MXL	Mexican ancestry in Los Angeles, California	52
CHB	Han Chinese in Beijing, China	84
CHD	Chinese in Metropolis Denver, Colorado	85
JPT	Japanese in Tokyo, Japan	86

CHAPTER 3: LONG-TERM GENETIC STABILITY AND A HIGH ALTITUDE EAST ASIAN ORIGIN FOR THE PEOPLES OF THE HIGH VALLEYS OF THE HIMALAYAN ARC[2]

3.1: Abstract

The high-altitude transverse valleys [>3,000 m above sea level (masl)] of the Himalayan arc from Arunachal Pradesh to Ladahk were among the last habitable places permanently colonized by prehistoric humans due to the challenges of resource scarcity, cold stress, and hypoxia. The modern populations of these valleys, who share cultural and linguistic affinities with peoples found today on the Tibetan plateau, are commonly assumed to be the descendants of the earliest inhabitants of the Himalayan arc. However, this assumption has been challenged by archaeological and osteological evidence suggesting that these valleys may have been originally populated from areas other than the Tibetan plateau, including those at low elevation. To investigate the peopling and early population history of this dynamic high-altitude contact zone, we sequenced the genomes (0.04×-7.25×, mean 2.16×) and mitochondrial genomes (20.8×-1,311.0×, mean 482.1×) of eight individuals dating to three periods with distinct material culture in the Annapurna Conservation Area (ACA) of Nepal, spanning 3,150-1,250 y before present (yBP). We demonstrate that the region is characterized by long-term stability of the population genetic make-up despite marked changes in material culture. The ancient genomes, uniparental haplotypes, and high-altitude adaptive alleles suggest a high-altitude East Asian origin for prehistoric Himalayan populations.

[2] Citation for chapter: Jeong C et al. 2016. "Long-term genetic stability and a high altitude East Asian origin for the peoples of the high valleys of the Himalayan arc." Proc Natl Acad Sci USA 113: 7485-7490.

3.2: Introduction

The world's high plateaus and great mountain ranges were among the last places colonized by humans in prehistory (Gamble 1994; Aldenderfer 2006; Belwood 2014). The challenges of rough terrain, cold stress, hypoxia, and the relative scarcity of resources in these high places significantly slowed the pace at which permanent occupation took place. The Himalayan mountain range and the Tibetan plateau are among the highest places on earth. The Himalayas include 9 of the 10 tallest mountains in the world, and, at an average elevation of 5,000 m above sea level (masl), the Tibetan Plateau is ~25% higher than the Peruvian altiplano, the next highest plateau in the world (Houston and Hartley 2003; Aldenderfer 2006; Harris 2006). Genome-wide studies of the geographic structure of modern populations clearly point to the Himalayas as a barrier to gene flow between East Asians and Western Eurasians (Wang et al. 2012). However, there is also extensive evidence of cultural and linguistic diversity across the Himalayan arc, which hints at a long history of cross-regional contact (Gayden et al. 2009; Gayden et al. 2013).

Currently available archaeological and osteological data and genetic data from modern-day populations have been used to support contrasting hypotheses of South Asian (Stacul 1968; Hüttel 1997; Singh 1999), Central Asian (Alt et al. 2003), lowland Southeast Asian (Peng, Palanichamy, et al. 2011), and high-altitude East Asian (Gayden et al. 2009; Gayden et al. 2013) origins for the earliest Himalayan inhabitants, and there is likewise little agreement regarding subsequent regional population history (Alt et al. 2003; Aldenderfer 2013). Successful permanent habitation of high-altitude environments requires numerous physiological adaptations, and recent genetic studies have identified robust signals of positive natural selection underlying

adaptations to hypoxia in Tibetans (Beall et al. 2010; Simonson et al. 2010; Yi et al. 2010) and in the Sherpa (Jeong et al. 2014), an ethnic group that migrated from the eastern Tibetan plateau to Nepal 400–600 y ago (ya) (Oppitz 1974). Tibetans and Sherpa are the only two present-day high-altitude East Asian ethnic groups that have been studied to date, using genome-wide markers. Although the Tibetan plateau has been the subject of intense study regarding its population history and high-altitude adaptation (Beall et al. 2010; Aldenderfer 2011; Qi et al. 2013; Xiang et al. 2013), far less is known about the much later colonization of the surrounding high transverse valleys along the Himalayan arc. Elucidating this history is important because these valleys have long served as natural corridors and trade routes connecting the Tibetan plateau to the Indian subcontinent. Moreover, the role of adaptation to high-altitude hypoxia in the initial colonization of these valleys and in the subsequent gene flow through them is entirely unexplored.

The Annapurna Conservation Area (ACA) of Upper Mustang, Nepal (**Figure 3.1**) is a major high-elevation corridor (2,800-4,500 masl) that includes the earliest known archaeological sites containing preserved human remains in a Himalayan transverse valley (Aldenderfer and Eng 2016). Foodstuffs within ACA prehistoric funerary contexts include domesticates of both West Asian (e.g., barley, buckwheat, lentils, peas, sheep, and goats) and East Asian (e.g., rice) origin (Knörzer 2000). In addition to locally made utilitarian wares, prestige objects include copper ornaments and vessels, carnelian beads, marine shell pendants, and faience suggestive of a strong South Asian connection, as well as bamboo baskets, mats, and cups and wooden furniture and design motifs suggesting contact with Central Asia and Xinjiang (Alt et al. 2003; Massa 2013). Later periods after ca. 1,750 y before present (yBP) also include Chinese silk and glass beads from Sassania (modern-day Iran) and central and far southern India, as well as gold and silver masks that resemble those found in western Tibet, Ladahk, and Kyrgyzstan

(Aldenderfer 2013). Finally, local mortuary practices initially resemble those observed in northern Xinjiang, but after ca. 1,500 yBP include defleshing, a practice that may have multiple origins but is primarily associated with Western Asian cultures (Aldenderfer and Eng 2016). Therefore, there is evidence that early populations in the Himalayan transverse valleys were exposed to influences from a remarkably wide geographic extent, from Iran to eastern China.

Given the complexity in material culture, currently available archaeological data cannot determine whether population replacement, cultural diffusion, or both are responsible for these diverse influences. Furthermore, interpretation of linguistic and genetic data from present-day populations is complicated by multiple historically documented Tibetan migrations after ca. 1,300 yBP linked to the rise and fall of the Tibetan Empire, extensive warfare, and the establishment of modern nation states (LaPolla 2001; Childs 2012). For these reasons, the analysis of ancient human genomes provides a unique and direct means for resolving competing hypotheses regarding the population history of the high Himalayas.

To investigate the peopling and early population history of the ACA, we obtained genome-wide sequences and high-coverage mitochondrial sequences from eight individuals dating to three periods with distinct material culture: Chokhopani (3,150–2,400 yBP), Mebrak (2,400–1,850 yBP), and Samdzong (1,750–1,250 yBP) (**Table 3.1**). Following initial population affinity analyses, we then further sequenced the genomes of five individuals to >2× coverage to obtain higher-resolution genome data and increase the coverage of two genes associated with high-altitude adaptation, ELGN1 (egl-9 family hypoxia-inducible factor 1) and *EPAS1* (endothelial PAS domain protein 1). Our results are consistent with long-term genetic stability in the region; additionally, genome sequences, uniparental haplotypes, and high-altitude adaptive alleles support a high-altitude East Asian origin for these prehistoric Himalayan populations.

Figure 3.1: Map of the ACA and sampling locations. The ACA (dark gray), located in the Upper Mustang of north-central Nepal and bordering Tibet (Inset), is situated between the Annapurna and Dhaulagiri Massifs of the main Himalayan mountain range. The ACA includes 14 mountains in excess of 6,000 masl, and it contains a single major drainage, the Kali Gandaki River, which originates on the Tibetan plateau. Map is adapted from (Banskota and Sharma 1995).

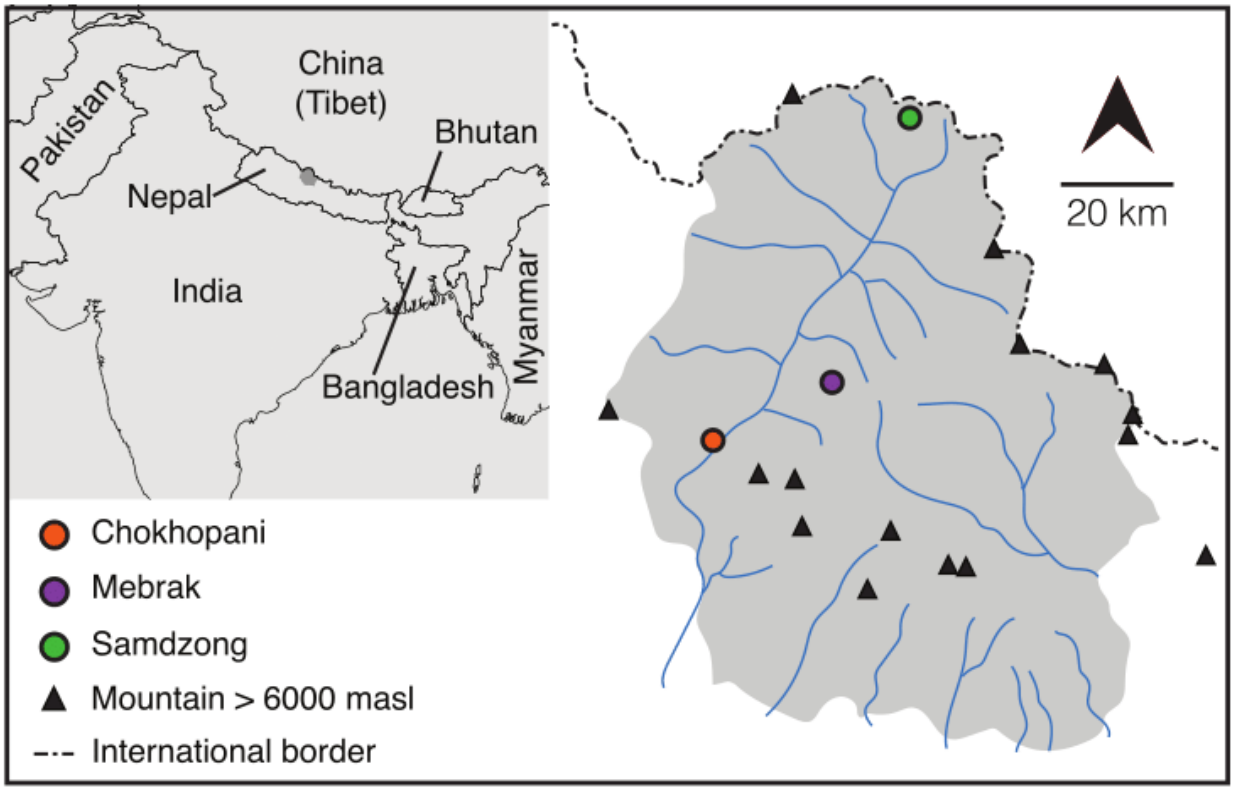


Table 3.1: ACA dental samples investigated in this study

Period / Site	Dates	No. samples	Sample ID
Chokhopani	3,150-2,400 yBP	1	C1
Mebrak	2,400-1,850 yBP	3	M63, M240, M344
Samdzong	1,750-1,250 yBP	4	S10, S35, S40, S41

3.3: Results

Ancient DNA Extraction and Sequencing Quality

Eight prehistoric ACA dental samples (C1, M63, M240, M344, S10, S35, S40, and S41) were sequenced in the first phase of this study and found to contain relatively high proportions of human DNA, ranging from 2.6% to 58.3% (**Supplementary Text 3.1, Supplementary Table 3.1**). Five of these samples were selected for deeper sequencing. This included three Samdzong period samples (S10, S35, and S41) containing ≥40% human reads and the oldest sample in the study, C1, dating to the Chokhopani period, containing 31.0% human reads, each of which was sequenced to >2× mean coverage. A Mebrak period sample, M63, with 18.9% human reads was also sequenced to 1× mean coverage. In total, mean sequence coverage at a genome-wide level for all eight samples ranged between 0.044× and 7.253× and between 20.8× and 1,311.0× for the mitochondrial genome (**Supplementary Text 3.1, Supplementary Table 3.1**). Genetic sex was confidently assigned for all eight individuals, of which seven were male (**Supplementary Text 3.1, Supplementary Figure 3.1**). Given the comparatively low proportions of human DNA reported in previous ancient DNA (aDNA) studies, the preservation of the ACA samples is very good, which is consistent with the arid and cold burial environment and relatively low thermal age of the sites (Ziesemer et al. 2015).

Assessment of Contamination from Modern Humans

After initial alignment, we assessed whether the human reads we recovered were likely to be endogenous (i.e., not resulting from modern contamination) by examining chemical damage patterns typical of aDNA (Briggs et al. 2007; Ginolhac et al. 2011; Jónsson et al. 2013) and

estimating the proportion of contaminant reads from mtDNA sequences (Fu et al. 2013). We observed typical ancient DNA damage patterns in all of the ACA samples, suggesting that the vast majority of DNA is of ancient origin. First, human DNA sequences were short in length, with median lengths of 55–87 bp (**Supplementary Text 3.1, Supplementary Table 3.1 and Supplementary Figure 3.2**). Second, 8.8–19.0% of sequences exhibited terminal 5' C > T miscoding lesions (**Supplementary Text 3.1, Supplementary Figure 3.3**), a characteristic pattern of aDNA damage. Finally, purines (A and G) were enriched at 5' -1 positions (**Supplementary Text 3.1, Supplementary Figure 3.4**), indicating depurination-driven strand breaks, another characteristic pattern of aDNA damage.

These features qualitatively support a high proportion of endogenous DNA in the ACA samples. However, the dataset can still contain a small number of contaminant human reads. Therefore, we estimated the proportion of contaminant mitochondrial reads, using a Bayesian method implemented in the program contamMix (Fu et al. 2013). The estimated proportion of endogenous reads in the ACA samples is >98% for all samples except M344 (94.4%), suggesting minimal contamination from other humans (**Supplementary Text 3.1, Supplementary Table 3.1**).

Genome-Wide SNP Profiling of Ancient DNA Samples

To understand the genetic relationship between the ACA aDNA samples and populations around the world, we compared sequences from our first-phase sequencing data to genetic data of 26 contemporary populations from the 1,000 genomes (1KG) project and high-coverage (≥30×) Illumina-sequenced whole genomes of 17 modern humans, including 4 Sherpa and 2 Tibetans from Nepal. Overlapping each aDNA sample dataset with the above population genetic

data panel, we retrieved 0.47–6.36 million autosomal SNPs for our first-phase analyses (**Supplementary Text 3.1, Supplementary Table 3.1**). All eight ACA individuals across the three time periods were found to be most closely related to East Asians (**Supplementary Text 3.1, Supplementary Figures 3.5-3.8**), a finding consistently supported by the results of several approaches, including principal components analysis (PCA), model-based unsupervised genetic clustering, and the outgroup f_3 statistic. The latter is a measure of genetic affinity that measures the branch length from an outgroup to the split point of a pair of populations (Raghavan et al. 2014).

To refine our inferences of genetic affinity, we further sequenced five ACA individuals to 1.0-7.3× coverage and compared the resulting genotypes to array genotyping data from Tibetans (Wang et al. 2011), Sherpa (Jeong et al. 2014), and populations from the Human Genome Diversity Panel (Li et al. 2008), as well as whole-genome sequences of two Nepali Tibetans. Multiple lines of evidence consistently indicate high-altitude East Asians (i.e., the Sherpa and Tibetans) as the closest contemporary populations to the ACA individuals, regardless of time period. First, ACA individuals cluster together with Tibetans in PCA (**Figure 3.2**). Second, model-based unsupervised clustering infers that a large proportion of ancestry in the ACA individuals is shared with the Sherpa and Tibetans (**Figure 3.3**). Third, all ACA individuals have the largest outgroup f_3 statistic with the Sherpa and Tibetans, followed by other Tibeto-Burman speaking groups such as Naxi, Yi, and Tujia (**Figure 3.4A and Supplementary Text 3.1, Supplementary Figures 3.8 and 3.9**). Finally, formal comparison of population affinity in the form of the D test shows that all of the ACA individuals are more closely related to Tibetans from Lhasa (Z = 2.7-8.0 SD), Tibetans from Nepal (Z = 0.8-4.6 SD), and the Sherpa (Z = 2.5-6.8 SD) than to any other population (**Figure 3.4B and Supplementary Text 3.1, Supplementary**

Figure 3.10). Additionally, outgroup f_3 (**Supplementary Text 3.1, Supplementary Figure 3.12A**) and D tests (**Supplementary Text 3.1, Supplementary Figure 3.12B**) support, albeit slightly less consistently, a greater genetic affinity of contemporary high-altitude populations with the ACA samples than with the Yi or Naxi (S41 is an exception, possibly due to a minor west-Eurasian component) (**Figure 3.3, Supplementary Text 3.1 and Supplementary Figure 3.12**). Taken as a whole, our results strongly suggest that the ACA individuals are closely related to contemporary high-altitude East Asian populations.

Figure 3.2: PCA of East Asian populations and ancient ACA individuals. All five ACA samples cluster with Tibetans. PC1 and PC2 were calculated using all contemporary East Asian samples. Ancient ACA samples were projected onto the PC plane, using the "*lsqproject: YES*" option.

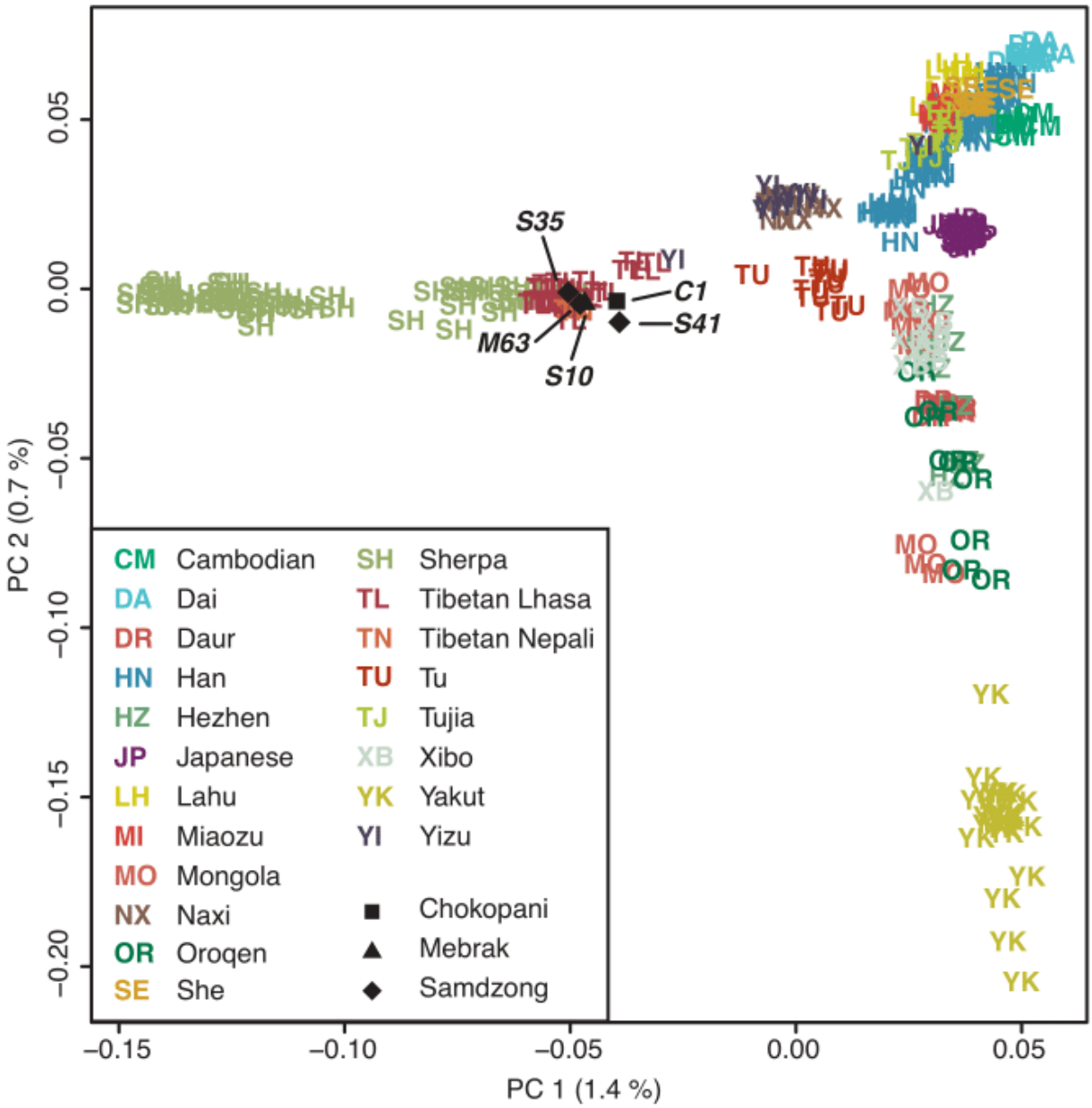


Figure 3.3: Unsupervised genetic clustering with two to nine ancestral populations (*K* = 2–9). All five ACA samples exhibit ancestry profiles most similar to Tibetans. The Sherpa, Tibetans, and ACA samples share a distinct high altitude ancestry (red) in the highest proportions, followed by other Tibeto-Burman speaking groups such as Naxi and Yi. *K* = 2 is shown at the bottom of the plot; *K* = 9 is shown at the top.

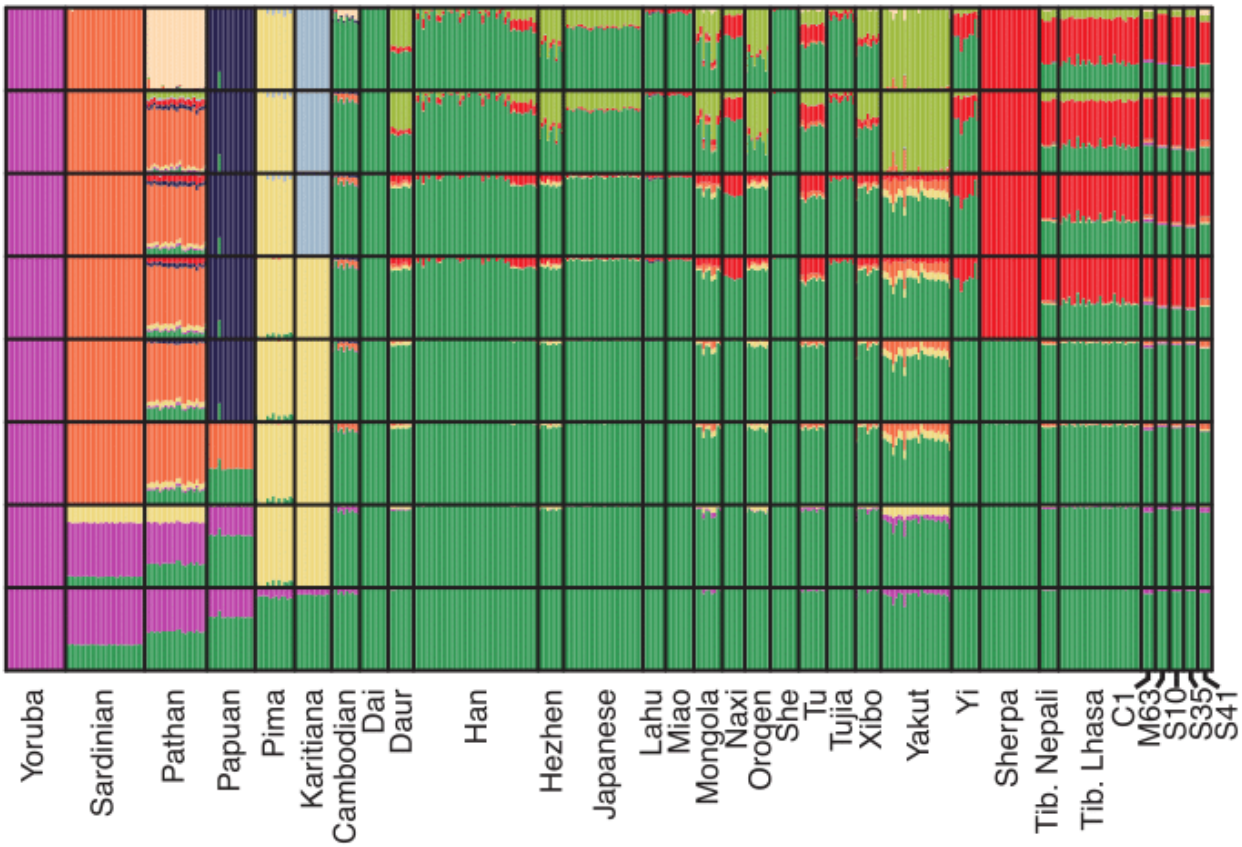
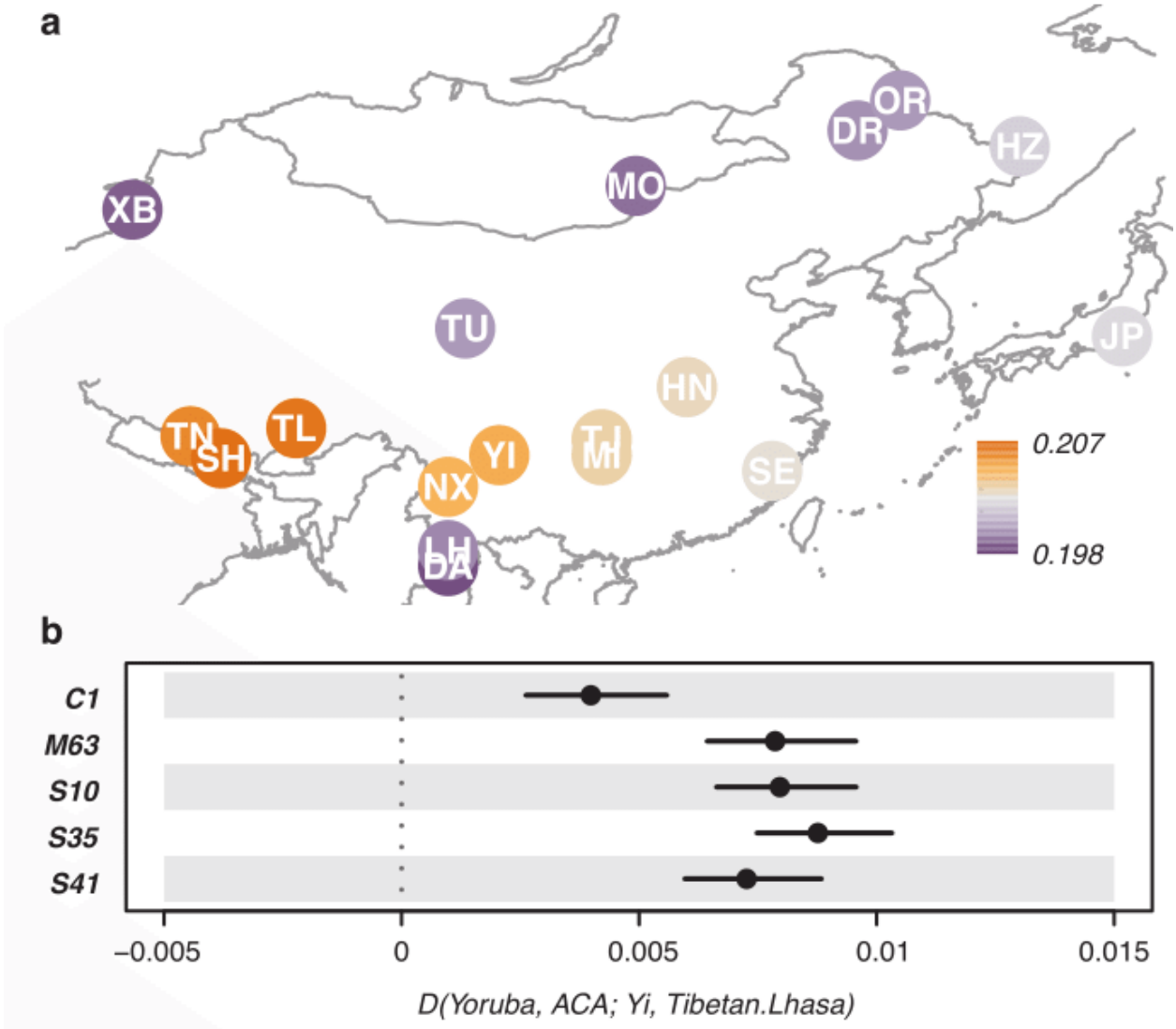


Figure 3.4: Genetic affinity of ACA individuals and East Asian populations, using genome-wide SNP data. (A) Genetic affinity with ancient sample C1 ismeasured by f_3 (Yoruba; ACA, X). For all ACA samples (**Supplementary Figure 3.9**), either Sherpa or Tibetans were the closest modern population (a larger f_3 value indicates a closer relationship), followed by other Tibeto-Burman speaking groups, such as Naxi and Yi. (B) All ACA samples are significantly more closely related to contemporary high altitude East Asians than they are to lowland Tibeto-Burman speaking groups, as shown by positive values of Patterson's *D* (Yoruba, ACA; Yi, Tibetan.Lhasa). Equivalent results are observed if the test is performed with alternative proxy populations for high-altitude East Asians (e.g., Tibetan.Nepali or Sherpa) and lowland Tibeto-Burman speakers (e.g., Naxi or Tujia) (**Supplementary Figure 3.10**).



High-Altitude Functional Alleles

Encouraged by the genetic profiles of the ACA individuals, we investigated whether the five more deeply sequenced ACA individuals share high-altitude adaptive genetic variants (Beall et al. 2010; Simonson et al. 2010; Yi et al. 2010) with Tibetan populations (**Supplementary Text 3.1, Supplementary Table 3.2**). More specifically, we determined whether they have derived alleles at 20 noncoding SNPs that tag the selected haplotype at the *EPAS1* gene (Huerta-Sánchez et al. 2014) or at two nonsynonymous SNPs (rs12097901 and rs186996510) with signatures of adaptive allele frequency divergence at the *EGLN1* gene (Xiang et al. 2013; Lorenzo et al. 2014). Currently, there is broad agreement for selection on the derived *EGLN1* alleles beginning ca. 8,000 ya (Xiang et al. 2013; Lorenzo et al. 2014), but dating the onset of selection for the derived *EPAS1* haplotype has proved more controversial. The derived *EPAS1* haplotype was recently shown to have originated in the Denisova genome and its presence in the human genome represents a recent archaic introgression (Huerta-Sánchez et al. 2014). Consequently, the subsequent selection of this haplotype in humans is difficult to model using genetic data from living populations, and dates ranging from 2,750 ya to 18,250 ya have been proposed (Yi et al. 2010; Peng, Yang, et al. 2011; Hackinger et al. 2016).

Interestingly, all reads from our ACA individuals match the derived allele for the nonsynonymous *EGLN1* SNP rs186996510 (**Supplementary Text 3.1, Supplementary Table 3.2**), including the oldest Chokhopani sample (C1). This derived allele, c.12G > C (p.Asp4Glu), is reported in high frequency in Tibetans (0.64–0.85) (Xiang et al. 2013; Lorenzo et al. 2014), but is rare in low-altitude East Asians (0.03 in 1KG phase 3 East Asians) and virtually absent outside East Asia. Functional studies have implicated this allele as playing a role in oxygen homeostasis under hypoxic conditions (Lorenzo et al. 2014; Petousi et al. 2014). In contrast,

reads supporting derived alleles at the *EPAS1* SNPs were found in two of the three later Samdzong individuals (S35 and S41), but not in the earlier Chokopani (C1) or Mebrak (M63) individuals. This observation of shared adaptive alleles between ancient ACA individuals and contemporary Tibetans is consistent with our genome sequence results suggesting that the ACA inhabitants are affiliated with contemporary high-altitude East Asians. In addition, the contrasting pattern of alleles for the two genes leads us to speculate that the *EGLN1* and *EPAS1* adaptive haplotypes rose to high frequency at different time points in these ancient high-altitude populations, although more samples must be sequenced to accurately estimate allele frequency change across time.

Mitochondrial and Y Chromosome Haplogroup Identification

Using high-coverage, consensus full mtDNA genome sequences (**Supplementary Text 3.1, Supplementary Table 3.1**), we next inferred haplogroup assignment for each ACA individual. All eight individuals are assigned to haplogroups reported to be present in contemporary Nepalis and/or Tibetans (**Supplementary Text 3.1, Supplementary Table 3.3**) (Fornarino et al. 2009; Qi et al. 2013) and rare or absent in present-day Indian and Pakistani populations (Metspalu et al. 2004). The oldest sample in our study, C1, belongs to haplogroup D4, a major maternal lineage among Tibetans. Interestingly, Tibetan D4 has a deep divergence time from other East Asian populations (26-27 kya), further supporting genetic affinity between the ACA individuals and contemporary high-altitude East Asians (Qi et al. 2013). Four male individuals with >2× coverage were determined to belong to Y chromosome haplogroups O-M117 and D (**Supplementary Text 3.1, Supplementary Table 3.4**), which are among the most

frequent haplogroups in contemporary Tibetans (Qi et al. 2013) based on haplogroup-tagging SNPs from the cleantree program (Ralf et al. 2015).

3.4: Discussion

The role of geography in migration and population structure has been a central topic in population genetics studies of our species and others (Novembre et al. 2008; Petkova et al. 2016). At a genetic level, the Himalayan arc delineates a sharp genetic barrier between South Asian and East Asian populations, a striking anomaly against a general isolation-by-distance pattern of human population structure across much of Eurasia (Wang et al. 2012). The asymmetric topography of the Himalayan massif, bordered by a high-elevation plateau to the north and lowland plains to the south, is reflected in the current regional genetic structure of human populations, evidenced by autosomal and Y chromosome STR (short tandem repeats) frequencies in modern Nepalese populations (Gayden et al. 2009; Gayden et al. 2013). Our results suggest that the Himalayas have long served as a remarkably resistant barrier to northward but not southward gene flow and that this genetic boundary has been stable for at least the last three millennia. The eight ancient ACA individuals analyzed in this study exhibit a strong and consistent genetic affiliation to contemporary East Asian, and especially to high-altitude East Asian (Sherpa and Tibetan), populations. Therefore, previous proposals of a South Asian (Stacul 1968; Hüttel 1997; Singh 1999), Central Asian (Alt et al. 2003), or low-elevation Southeast Asian (Peng, Palanichamy, et al. 2011) origin of the first inhabitants of this region, as well as speculation regarding subsequent prehistoric population replacement or large-scale

admixture with lowland populations (Tiwari 1985; Alt et al. 2003; Aldenderfer 2013), are not supported.

It is interesting that the high-altitude barrier to migration seems to be more permeable from the northern, as opposed to the southern, side of the Himalayan arc. One can speculate that this disparity may be due to the topographical differences between the northern and southern sides of the arc: the altitudinal gradient is much more gradual in the north than in the south. Thus, ascending populations on the north side may have been able to stay at intermediate altitudes for extended periods of time, allowing for acclimatization and the accumulation of genetic and subsistence adaptations, whereas potential migrants from the south side had no access to such a buffer zone because of the limited availability of sufficient habitable land at intermediate altitudes. This scenario is supported by the archaeological record of the Tibetan plateau. Archaeological data from the northeastern Tibetan plateau indicate an initial occupation ca. 15,000 ya (Zhao et al. 2009; Aldenderfer 2011), long before the colonization of the high-traverse valleys in the Himalayan arc. Archeological data also support later influences from the East Asian side of the plateau associated with the appearance of agriculture after 5,500 yBP, evidenced by the adoption of Neolithic domesticates, first from East Asia (millets and pigs) and later from West Asia (via Central Asia: barley, sheep, and goats). It has been proposed that these changes enabled populations on the plateau to move to higher and more marginal lands after ca. 4,000 yBP (Guedes et al. 2014), where they may have subsequently served as a source population for the Himalayan transverse valleys. It is beyond the scope of our current study, however, to address whether the spread of agriculture onto the plateau was accompanied by population migration.

Genetic adaptation to high altitude also likely facilitated this asymmetric colonization. Accumulation of beneficial mutations is a feature expected for a population gradually adapting to a new environment. Evolution of such beneficial mutations across time provides crucial information for understanding the strength and cause of natural selection. Contemporary high-altitude East Asians on the Tibetan plateau have at least two such genes, *EPAS1* and *EGLN1*, that exhibit strong signatures of positive natural selection as well as functional properties consistent with an adaptive role in high-altitude environments (Beall et al. 2010; Simonson et al. 2010; Sun et al. 2010; Lorenzo et al. 2014; Petousi et al. 2014). Importantly, we found that the oldest Chokhopani sample (C1) and three later Samdzong individuals (S10, S35, and S41) are most likely homozygous for a derived nonsynonymous allele of the *EGLN1* SNP (rs186996510), suggesting that this allele was already common in the founding population. In contrast, derived alleles from the *EPAS1* SNPs were observed only in Samdzong individuals, implying an asynchronous evolution of the two genes. However, sequencing of additional ancient samples through time is necessary to reconstruct the adaptive evolution of these and other beneficial mutations in the ACA. Given the unusually high quality of aDNA from the ACA, population-level ancient genome sequencing is likely an achievable goal once additional early archaeological specimens are available.

It is tempting to compare this case to archaeogenetic studies in Europe, which suggest that large-scale cultural transitions are frequently associated with massive population movements (Skoglund et al. 2012; Skoglund et al. 2014; Allentoft et al. 2015; Haak et al. 2015). In the Himalayas, we observe two discrete cultural transitions (associated with the Mebrak and Samdzong periods) without evidence of changes in the genetic makeup of the population. One sample from Samdzong (S41) may be an exception in that it is the only one showing some

amount of non-East Asian ancestry; however, this proportion is estimated to be small (**Figure 3.3**). Therefore, the predominance of East Asian ancestry in the ACA samples supports our hypothesis that certain topographies, specifically very high altitudes, require a unique set of adaptations, genetic or cultural, that differ from those sufficient for low-altitude migration and colonization. However, because current archaeological data are largely limited to funerary contexts, we caution that the archaeological changes we observe in the ACA may not represent full-scale cultural transitions.

In this study, we conducted to our knowledge the first successful ancient DNA investigation of prehistoric Himalayan populations and retrieved high proportions of endogenous aDNA from eight high-altitude ACA individuals dating to three distinct cultural periods spanning 3,150-1,250 yBP. Our population genetic analysis strongly supports the genetic affiliation of prehistoric Himalayan populations with contemporary East Asians and at a subcontinental level suggests a closer affinity with present-day high-altitude East Asians, such as Tibetans and Sherpa, than with low-altitude East Asians. Moreover, this affinity is consistent through time, suggesting that temporal changes in material culture and mortuary behavior largely reflect acculturation or cultural diffusion rather than largescale gene flow or population replacement from outside East Asia. Finally, we provide to our knowledge the first empirical evidence for differing evolutionary dynamics of selection on the *EGLN1* and *EPAS1* genes in prehistoric high-altitude populations. Considering the pivotal role of the Himalayan high transverse valleys in connecting far-flung Eurasian populations, as well as the environmental challenges they impose on their inhabitants, our study has deep implications for the understanding of human migration history and adaptation to local environments and for future genetic archaeology studies.

3.5: Experimental Procedures

Study Design and Samples

The ACA of Upper Mustang, Nepal is located in northern central Nepal and covers an area of ~7,630 km^2 (**Figure 3.1 and Supplementary Text 3.1, section 1**). Prior archaeological research in the region identified three distinct periods of occupation: Chokhopani (3,150-2,400 yBP) (Tiwari 1985; Simons and Schön 1998), Mebrak (2,400-1,850 yBP) (Simons and Schön 1998; Alt et al. 2003), and Samdzong (1,750-1,250 yBP) (Aldenderfer 2013; Aldenderfer and Eng 2016), each defined by a type site of the same name. Dental samples from 12 individuals were selected for DNA screening (**Supplementary Text 3.1, Supplementary Table 3.5**), of which 8 yielded sufficient data for continental-level ancestry analysis (**Table 3.1**). Of these, 5 were more deeply sequenced to investigate questions regarding regional ancestry and high-altitude adaptation. Use of ancient and preexisting, deidentified modern human genetic data was determined to be exempt from human subjects review (University of Chicago IRB12-1785).

Ancient DNA Extraction, Library Construction, and Sequencing

DNA extraction was performed in a dedicated ancient DNA facility in accordance with established contamination control precautions and workflows, as previously described (Ziesemer et al. 2015) (**Supplementary Text 3.1, section 2**). Following decalcification and digestion, two DNA extraction methods were compared: (i) phenol-chloroform separation followed by purification and concentration using a MinElute PCR Purification kit (Qiagen) (Ziesemer et al. 2015) and (ii) salting out followed by purification and concentration using a QIAamp DNA Mini Kit (Tito et al. 2011). For 3 of the 12 individuals, DNA extraction was performed using both

methods. Purified DNA was quantified using a Qubit High Sensitivity dsDNA assay (Life Technologies). DNA extracts were built into indexed Illumina libraries, using a double-stranded library protocol, with minor modifications (**Supplementary Text 3.1, section 3**). The resulting libraries were purified, quantified, and pooled for sequencing on the Illumina HiSeq platforms, using paired-end 100-bp, 125-bp, or 150-bp chemistry (**Supplementary Text 3.1, Supplementary Table 3.6**).

Sequence Data Filtering and Quality Control

Adapter sequences were removed and each read pair was merged into a single sequence, using a publicly available python script (Kircher 2012) (https://bioinf.eva.mpg.de/fastqProcessing/MergeReadsFastQ_cc.py). Merged reads were mapped to the human reference genome hg19, using BWAbacktrack 0.7.9a (Li and Durbin 2009). Uniquely mapped reads ≥35 bp were kept, and PCR duplicates were removed, keeping the one with the highest mapping quality score. Step-by-step filtering and quality parameters and statistics are provided in **Supplementary Text 3.1, section 4** and **Supplementary Table 3.1**.

Comparison of DNA Extraction Methods

Before proceeding further, the performance of the two DNA extraction methods was compared (**Supplementary Text 3.1, section 5**). The phenol-chloroform/MinElute method substantially outperformed the salting out/QIAamp method in both total DNA yield and human DNA content (**Supplementary Text 3.1, Supplementary Table 3.5 and Supplementary Figure 3.11**). Consequently, all subsequent genetic analyses were restricted to the eight samples

(C1, M240, M344, M63, S10, S35, S40, and S41) extracted using the phenol-chloroform/MinElute method.

Assessment of Genetic Sex, Sample Contamination, and DNA Damage

Genetic sex was estimated using previously described methods for shotgun sequence data (Skoglund et al. 2013) (**Supplementary Text 3.1, section 6**). Contamination was assessed by estimating the proportion of endogenous reads among human mitochondrial DNA sequences, using the Bayesian program contamMix (Fu et al. 2013) (**Supplementary Text 3.1, section 7**). For each sample, the estimated endogenous content and 95% confidence interval are provided in **Supplementary Table 3.1**. DNA fragment lengths and damage patterns typical of ancient DNA were assessed from uniquely mapped, nonduplicate reads, using the mapDamage program (Ginolhac et al. 2011; Jónsson et al. 2013) (**Supplementary Text 3.1, section 8**).

Data Filtering and Compilation for Population Genetic Analysis

For population genetic analysis, we retrieved ACA genetic information from sequence reads (**Supplementary Text 3.1, section 9**). High-quality base calls (≥Q30) from reads with high mapping quality scores (≥30) were collected for each genomic position, using the mpileup command of SAMtools v1.2 (Li et al. 2009), after masking 5 bp at both ends of reads to reduce the effect of cytosine deamination. For the analysis of the first-phase data (<1× coverage), one read at each position was then randomly sampled to generate haploid genotypes. ACA aDNA data were then overlapped with available genetic variation data of 26 worldwide populations from the 1KG project phase 3 haplotype set (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/), which includes 2,504 unrelated

individuals, and high-coverage (≥ 30×) published (Meyer et al. 2012; Jeong et al. 2014; Prüfer et al. 2014) and unpublished (https://www.simonsfoundation.org/life-sciences/simons-genome-diversity-project-dataset/) whole-genome sequences of 17 modern humans from 10 populations, including Tibetans and Sherpa. Finally, high-coverage genotypes from the chimpanzee genome assembly Pan_troglodytes-2.1.4 (panTro4), Altai Neandertal (Prüfer et al. 2014), and Denisovan (Meyer et al. 2012) were compiled to assess ancestral alleles and archaic hominin ancestry. Variants within human repeat regions or CpG islands (Wu et al. 2010), variable sites with multiple alternative alleles, sites with strand ambiguity (A/T or G/C SNPs), sites prone to cytosine deamination, and sites not present in the 1KG dataset were removed for PCA, clustering, and outgroup f_3 analyses. Additionally, sites with missing genotypes among the 17 modern humans or nonhuman samples were also excluded. This process led to variable numbers of SNPs for each aDNA sample, ranging from 0.47 million to 6.36 million SNPs. For the analysis of the second-phase data (1.0-7.3× coverage), we sampled a read for ~650,000 positions in the Human Genome Diversity Panel (HGDP) (Li et al. 2008). Additional array genotyping data for the Sherpa (Jeong et al. 2014) and Tibetans (Wang et al. 2011) were intersected with the HGDP dataset, as well as genotypes of two contemporary Nepali Tibetan individuals. For genetic clustering, we used genotype likelihoods calculated from the GATK v2.7-4 UnifiedGenotyper module (McKenna et al. 2010) (**Supplementary Text 3.1, section 9**).

Whole-Genome Ancestry Affiliation Analysis

PCA was performed using the smartpca program in the EIGENSOFT 6.0 package (Patterson et al. 2006). For analysis of the first-phase data, PCA was run for each ancient sample separately, using 0.15–2.03 million SNPs with minor allele frequency (MAF) ≥ 0.001. Results

for PC1 and PC2 were then merged by a Procrustes transformation, using the "procGPA" function in the R package "shapes" (Dryden 2014) (**Supplementary Text 3.1, section 10**). For analysis of the second-phase data, the "*lsqproject: YES*" option was used to project ancient samples onto the PC plane, calculated with 357,000 SNPs with MAF ≥ 0.01.We performed model-based genetic clustering analysis as implemented in the sNMF (Frichot et al. 2014) and NGSadmix (Skotte et al. 2013) programs. For analysis of the first-phase data, one allele from each modern sample was randomly sampled at each variable site to match the haploid nature of the ACA aDNA samples. SNPs with at least three copies of the minor allele were retained, and linkage disequilibrium (LD)-based SNP pruning was performed using PLINK v1.0.7 (Purcell et al. 2007) ($r^2 > 0.2$). The resulting SNPs were downsampled to match the number of SNPs available in each ACA sample (40,000–100,000), and analysis was performed in 50 replicates with random seeds for two to eight clusters (K) (**Supplementary Text 3.1, section 10**). For analysis of the second-phase data, 105,944 LD-pruned SNPs with MAF ≥ 0.01 were used, and 50 replicates were performed for K values of 2–9. Genetic affinity was estimated using the outgroup f_3 statistic (Raghavan et al. 2014) with 1KG YRI or HGDP Yoruba as an outgroup and using the D statistic, using the SNP set for PCA. The *qp3Pop* and *qpDstat* programs in the ADMIXtools v2 package (Patterson et al. 2012) were used to calculate f_3 and *D* statistics and associated SEs (**Supplementary Text 3.1, section 10**). In addition to comparison with ACA samples, f_3 and *D* statistics were also calculated comparing high-altitude East Asian and lowland Tibeto-Burman speaking populations (**Supplementary Text 3.1, Supplementary Figure 3.12**).

High-Altitude Adaptation Allele Analysis

For sites with a read depth ≥1, allelic variants were determined in the *EGLN1* gene and in 20 tagging SNPs in the *EPAS1* gene (**Supplementary Text 3.1, section 11**).

Uniparental Haplogroup Analysis

Consensus mtDNA sequences for all eight individuals were called from sequence reads, using the UnifiedGenotype module of the GATK v2.7-4 followed by haplogroup assignment using HaploGrep (Kloss-Brandstätter et al. 2011) (**Supplementary Text 3.1, section 12**). For five male individuals analyzed in the second phase of the study, the Y haplogroup was manually assigned based on reads piled up for 519 informative biallelic SNPs from a database associated with cleantree software (**Supplementary Text 3.1, section 13**).

3.6: Appendix: Supplementary Information

Supplementary Text 3.1: Supplementary Materials and Methods for Chapter 3

1. Study design and samples

The Annapurna Conservation Area of Upper Mustang, Nepal (ACA) is located in northern central Nepal and covers an area of approximately 7,630 km^2. It includes 14 mountains in excess of 6000 m, the tallest of which is Annapurna I (8091 m), the 10th highest mountain in the world. It has a single major drainage, the Kali Gandaki River, which has its origins on the Tibetan plateau (Banskota and Sharma 1995). The Kali Gandaki valley lies within a rain shadow, which has created an arid landscape that has promoted the long-term conservation of

archaeological and human remains. Annual precipitation and temperature vary within the ACA according to elevation. In Jomsom, Nepal, a town located at 2,729 masl in the ACA, annual precipitation is 307 mm and average annual temperature is 10.9°C (www.climate-data.org). With each 1,000 m rise in elevation, average annual temperature drops by 6°C, and annual precipitation in the trans-Himalayan region of the Upper Mustang is less than 200 mm (Banskota and Sharma 1995).

Prior archaeological research in the region identified three distinct periods of occupation: Chokhopani (3150-2400 BP) (Tiwari 1985; Simons and Schön 1998), Mebrak (2400-1850 BP) (Simons and Schön 1998; Alt et al. 2003), and Samdzong (1750-1250 BP) (Aldenderfer 2013; Aldenderfer and Eng 2016), each defined by a type site of the same name. At each site, the dead were housed in cliff-face, rock-cut, community mortuary shaft tombs in which adults, children, and both sexes are found. The three archaeological phases are defined primarily by tomb architecture and contents (Aldenderfer 2013).

The Chokhopani tombs contain collective burials, and artifacts include personal adornments such as carnelian, shell, and faience beads, musk deer teeth (possibly part of a necklace), schist bodkins, bronze bangles, mortuary ceramics, utilitarian wooden and stone objects, and copper jewelry and sheets (Simons et al. 1998; Aldenderfer 2013). The latter copper artifacts bear resemblance to poorly dated finds from the Copper Hoards of the upper Ganges river basin of the Indian subcontinent (Aldenderfer 2013). The Chokhopani tomb chronology has been determined by radiocarbon dating (Simons et al. 1998).

The Mebrak tombs contain collective burials, with individuals laid out in a flexed position on their sides on wooden platforms. For the first time, mummified heads of sheep and goats are included with the burials. Other tomb contents include a complete but disarticulated

horse skeleton, glass and carnelian beads, bamboo mats, bronze jewelry, a wooden bow, and wooden bowls and baskets (Alt et al. 2003). The burial goods additionally contain a wide variety of textiles (cotton, linen, wool, and other plant fibers) made using different weaving styles (including velveteen) and dyed in a variety of colors using alizarin, purpurin, indigo, lac-dye, ellagic aacid, and flavonol (Alt et al. 2003). The dead were laid out on intricately carved wooden platforms, some of which were painted with images of blue sheep (*Pseudois nayaur*), red deer (*Cervus elaphus*), and non-local markhor (*Capra falconeri*). A total of 28 samples from the Mebrak tombs have been radiocarbon dated and calibrated using dendrochronological data (Alt et al. 2003).

The Samdzong tombs contain both collective and individual burials, and for the first time the bodies show extensive evidence of cut marks from defleshing (and in some cases dismemberment) prior to being laid out on wooden platforms (Aldenderfer 2013; Aldenderfer and Eng 2016). The tombs contain an even greater number animal remains than in the Mebrak period, including sheep, goats, horses, and bovids. Other artifacts include wooden and bamboo artifacts, metal artifacts (vessels, knives, daggers, plates, arrowheads, and horse tack), three mortuary masks made of gold and silver, glass beads, ceramics, some textiles, including locally made materials and Chinese silks, and a wooden coffin painted with an image of figure riding a horse (Aldenderfer 2013; Gleba et al. 2016). The Samdzong chronology has been established based on a total of nine radiocarbon-dated samples collected from the tombs.

Initially, fifteen dental samples from twelve individuals of both sexes were selected for DNA screening and analysis (**Supplementary Table 3.5**). These samples were obtained during field excavations from 1990-2012 by the authors and others (Simons et al. 1998; Alt et al. 2003).

2. Ancient DNA extraction

DNA extraction was performed in a dedicated ancient DNA facility at the University of Oklahoma Laboratories of Molecular Anthropology and Microbiome Research (LMAMR) in accordance with established contamination control precautions and workflows, as previously described (Ziesemer et al. 2015). A non-template extraction control (negative control) was processed alongside experimental samples during all analytical steps. Prior to DNA extraction, all dental calculus was removed and the tooth surface was wiped clean with a 2% NaOCl solution, followed by molecular biology grade water. Areas of visible chemical or microbial damage were removed by mechanical abrasion using a Dremel rotary tool with a diamond bit. One tooth root was removed from each tooth, UV irradiated for 60 s on each side (CL-1000 UVP Crosslinker, 1.0 J/cm^2), and then crushed to a coarse powder. The tooth powder was agitated in a 0.5M EDTA solution for 15 minutes to remove remaining loosely bound contaminants and microbial DNA, and then decanted. The tooth powder was then resuspended in 1 mL of 0.5 M EDTA solution and incubated overnight at room temperature. 100 µl proteinase K (>600 mAU/ml; Qiagen) was then added and incubated at 37°C for 8 hours, followed by continued digestion under agitation at room temperature until decalcification was complete. The digestion buffer solution was refreshed after 48 hours and the two supernatants were combined for subsequent analyses. Following digestion, the samples were divided into two groups and two alternative ancient DNA extraction techniques were compared: 1) phenol-chloroform separation followed by purification and concentration using a MinElute PCR Purification kit (Qiagen) (Ziesemer et al. 2015), and 2) salting out followed by purification and concentration using a QIAamp DNA Mini Kit (Tito et al. 2011). Purified DNA was quantified using a Qubit High Sensitivity dsDNA assay (Life Technologies) (**Supplementary Table 3.5**).

3. DNA library construction and sequencing

For samples C1, M240, M344, S35, S40, and S41 extracted by technique 1, 28 µl of sample extract was constructed into double-stranded, single-indexed Illumina libraries using a blunt-end protocol at the LMAMR using previously described methods (Ziesemer et al. 2015) (**Supplementary Table 3.5**). Library construction was completed by PCR amplification using Platinum Taq HiFi polymerase (Life Technologies) and indexed primers. Each 50 µl reaction contained the following: 0.2 µl Platinum Taq HiFi Polymerase (5 U/µl), 5 µl 10X HiFi PCR Buffer, 1.5 µl 50mM MgCl2, 1 µl 2.5mg/ml BSA, 3 µl 2mM dNTPs, 1.5 µl 10 µM IS4 primer, and 1.5 µl of 10 µM indexed P7 primer, 24 µl of template, and 12.3 µl water. PCR amplification was performed in a post-PCR laboratory under the following conditions: initial denaturation at 95°C for 2 min, followed by 17 cycles of denaturation at 95°C for 15 s, annealing at 60°C for 30 s, and elongation at 72°C for 30 s, followed by a final elongation step at 72°C for 7 min. Resulting libraries were purified using a MinElute PCR Purification kit (Qiagen) and quantified using a Bioanalyzer 2100 High Sensitivity DNA assay (Agilent). In the first phase of the study, each library was pooled at equimolar concentration and sequenced at the University of Chicago Genomics Core using an Illumina HiSeq 2500 in Rapid Run mode with v2 paired-end 100 bp chemistry. In the second phase of the study, C1, S35 and S41 samples were sequenced at the same facility using an Illumina HiSeq 4000 with paired-end 100 bp chemistry.

For samples M63 and S10 extracted by technique 1 and samples C4, M344, M458, M294, S10, S39, and S41 extracted by technique 2, 20 µl of sample extract was constructed into double-stranded, single-indexed Illumina libraries using a blunt-end protocol at the Uppsala Ancient DNA Laboratory using previously described methods (Günther et al. 2015) (**Supplementary**

Table 3.5). Library construction was completed by PCR amplification using AmpliTaq Gold polymerase (Life Technologies) and indexed primers. Each library was amplified in quadruplicate 25 µl reactions containing the following: 0.1 U/µl AmpliTaq Gold, 1X AmpliTaq Gold Buffer, 2.5 mM MgCl2, 250 µM of each dNTP, 0.2 µM of IS4 primer, 0.2 µM of indexed P7 primer, 3 µl of template, and water. PCR amplification was performed in a post-PCR laboratory under the following conditions: initial denaturation at 94°C for 12 min, followed by 12-18 cycles of denaturation at 94°C for 30 s, annealing at 60°C for 30 s, and elongation at 72°C for 45 s, followed by a final elongation step at 72°C for 10 min. The replicate PCR reactions were pooled and purified using AMPure XP beads (Agencourt) and quantified using High Sensitivity DNA1000 screen tapes and reagents on a 2200 TapeStation (Agilent Technologies). In the first phase of the study, each library was pooled at equimolar concentration with eight or nine other libraries and sequenced at the SciLife Sequencing Centre in Uppsala using an Illumina HiSeq 2500 with v2 paired-end 125 bp chemistry. In the second phase of the study, M63 and S10 samples were sequenced at the same facility using an Illumina HiSeq X-Ten with paired-end 150 bp chemistry.

Base calling and base quality scoring for all libraries were performed using default Illumina software RTA v1.17.21.3 and CASAVA v1.8.2. Demultiplexing was performed using bcl2fastq conversion software v1.8.3, with no mismatch in sample indexes allowed. Metagenomic DNA sequences have been deposited in the NCBI Short Read Archive (SRA) under the project accession SRP065070, and sample accessions SRR2751055-SRR2751058, SRR2751060- SRR2751063, SRR2751066-SRR2751067, SRR2751070, SRR2751142, SRR2751148, SRR2751152, SRR3222643, SRR3222649, SRR3222655, SRR3222659,

SRR3222661, SRR3222664, SRR3222686, SRR3222749, SRR3222758, SRR3222765, and SRR3222772.

4. Sequencing data filtering and quality control

We first removed adapter sequences flanking inserted molecules and merged each pair of reads into a single sequence using a publicly available python script (Kircher 2012) (https://bioinf.eva.mpg.de/fastqProcessing/MergeReadsFastQ_cc.py). Only those paired reads with ≥ 11 overlapping bp were retained for alignment. Merged reads were mapped to the human reference genome hg19 (GRCh37, GenBank accession GCA_000001405.1) using BWA-backtrack 0.7.9a (Li and Durbin 2009). Seeding was disabled (-l 9999) and more mismatches were allowed with two non-default arguments: a larger maximum edit distance (-n 0.01) and maximum gap openings (-o 2). These arguments were used to consider damage patterns of ancient DNA samples, as previously suggested (Meyer et al. 2012). We kept uniquely mapped reads with length ≥ 35 bp and removed PCR duplicates, keeping one with the highest mapping quality score, using in-house scripts. The resulting data set is referred to as analysis-ready reads in the text. Step-by-step filtering and quality statistics are provided in **Supplementary Table 3.1**. Genomic sequence coverage information is provided in **Supplementary Table 3.7**.

5. Comparison of DNA extraction methods

Before proceeding further, we compared the performance of the two DNA extraction methods. The phenol-chloroform/MinElute method significantly outperformed the salting out/QIAamp method in both total DNA yield (27.5x; two-tailed t test, $p \leq 0.01$) and human DNA content (13.1x; two-tailed t test, $p \leq 0.01$) (**Supplementary Table 3.5 and Supplementary**

Figure 3.9). This is likely due to the fact that salting out involves a precipitation step that is known to disfavor the recovery of short DNA fragments (Gaillard and Strauss 1990), such as those typical of ancient DNA. Because salting out/QIAamp samples did not yield sufficient human DNA for ancestry analysis, all subsequent genetic analyses were restricted to the eight samples (C1, M240, M344, M63, S10, S35, S40, S41) extracted using the phenol-chloroform/MinElute method (**Table 3.1**).

6. Genetic sex typing

Genetic sex was estimated by comparing the ratio of DNA sequences aligning to the X and Y chromosomes using previously described methods for shotgun sequence data (Skoglund et al. 2013). All eight individuals were confidently assigned to sex, and seven out of eight individuals were male (**Supplementary Figure 3.1**). Genetic sex assignment for two individuals (M63 and S10) was then independently confirmed using a TaqMan duplex qPCR assay targeting a 106/112 bp region of the amelogenin gene following previously described methods (Krüttli et al. 2014). The assay was performed in triplicate and yielded genetic sex assignments consistent with the shotgun data analysis approach.

7. Assessment of DNA contamination

We estimated the level of contamination from exogenous human sources in our sequence data based on a method implemented in the contamMix (Fu et al. 2013) program (**Supplementary Table 3.1**). In brief, this program uses the majority-based mtDNA consensus sequence as a representative of the endogenous sequence, assuming that sequence coverage is high enough to accurately call a consensus sequence given presence of some contaminant reads

and sequencing errors. Then, it models observed sequence reads as a sample from a mixture of mitogenomes, including the endogenous one and potential contaminants, as represented by 311 contemporary human mitogenomes. The probability of reads drawn from the endogenous mitogenome, and its 95% confidence interval, is estimated through a Markov chain Monte Carlo algorithm.

8. Assessment of DNA damage

Endogenous DNA molecules from ancient samples typically show specific patterns of chemical damage, which have been used for supporting authenticity of aDNA sequence data in previous studies (Briggs et al. 2007). We assessed size distribution (**Supplementary Figure 3.2**) and damage patterns (**Supplementary Figures 3.3 and 3.4**) of inserted molecules using uniquely mapped, non-duplicate reads of each of eight samples, using mapDamage program (Ginolhac et al. 2011; Jónsson et al. 2013).

9. Data filtering and compilation for population genetic analysis

For population genetic analysis, two different sets of data were compiled, one for the first phase of data analysis and the other for the second phase. In both cases, we first retrieved genetic information of ACA aDNA samples from sequence reads. We collected high-quality base calls (base quality score ≥ 30) from reads with mapping quality score ≥ 30 for each genomic position, after masking 5 bp at both ends of reads to reduce effect of cytosine deamination. For the first phase of the study, which focused on low coverage (< 1x) data from eight individuals, we randomly picked up one read for each position and used them as haploid genotypes. For the second phase of the study, which focused on samples with 1.0-7.3x coverage, we sampled one

read per position for genotype-based analyses or calculated genotype likelihoods using the UnifiedGenotyper module of the Genome Analysis Toolkit (GATK) v2.7-4 (McKenna et al. 2010).

A worldwide population panel for the first phase of data analysis was compiled in the following way. First, the majority of data was from the 1000 genomes (1KG) project phase 3 haplotype set (https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.tgz), consisting of 2,504 unrelated individuals from 26 populations (http://www.1000genomes.org/about#ProjectSamples): Chinese Dai in Xishuangbanna, China, CDX; Han Chinese in Beijing, China, CHB; Japanese in Tokyo, Japan, JPT; Kinh in Ho Chi Mihn City, Vietnam, KHV; Southern Han Chinese, China, CHS; Bengali in Bengladesh, BEB; Gujarati Indian in Houston, TX, GIH; Indian Telugu in the UK, ITU; Punjabi in Lahore, Pakistan, PJL; Sri Lankan Tamil in the UK, STU; African Ancestry in Southwest US, ASW; African Caribbean in Barbados, ACB; Esan in Nigeria, ESN; Gambian in Western Division, the Gambia, GWD; Luhya in Webuye, Kenya, LWK; Mende in Sierra Leone, MSL; Yoruba in Ibadan, Nigeria, YRI; British in England and Scotland, GBR; Finnish in Finland, FIN; Iberian populations in Spain, IBS; Toscani in Italia, TSI; Utah residents with Northern and Western European Ancestry, CEU; Colombian in Medellin, Columbia, CLM; Mexican Ancestry in Los Angeles, California, MXL; Peruvian in Lima, Peru, PEL; and Puerto Rican in Puerto Rico, PUR. Second, we called genotypes of 13 modern humans (1-2 individuals from Yoruba, Sardinian, Karitiana, Han, Dai, Sherpa, Papuan and Australian aborigine) from published high-coverage (≥ 30x coverage) short read data (Meyer et al. 2012; Jeong et al. 2014; Prüfer et al. 2014). In addition, to increase local geographic coverage of the dataset, high-coverage genomes of two additional Sherpa and two Nepali Tibetans were added (https://www.simonsfoundation.org/life-sciences/simons-genome-diversity-project-dataset/).

Short reads of the above 17 genomes were processed following the "best practice workflows" from GATK v2.8-1 (McKenna et al. 2010; DePristo et al. 2011; Auwera et al. 2013). Specifically, reads were aligned to the human reference (GRCh37) using BWA-backtrack with "-q 15" option, duplicate reads were removed with Picard tool v1.98 (http://broadinstitute.github.io/picard/), locally realigned around indels and base quality score recalibrated using GATK. Only the properly paired non-duplicate reads with phred-scaled mapping quality score ≥ 30 were kept for genotype calling. For each sample, we called genotypes across all sites using the GATK UnifiedGenotyper module, based on bases with phred-scaled quality score ≥ 30, and kept sites with phred-scaled quality score ≥ 50. Finally, genotypes from the chimpanzee genome assembly Pan_troglodytes-2.1.4 (panTro4), Altai Neandertal (Prüfer et al. 2014) and Denisovan (Meyer et al. 2012) were compiled to assess ancestral alleles and archaic hominin ancestry. For the archaic hominin genomes, we downloaded genotype calls in VCF format. For chimpanzee reference sequence, we first converted human reference genome coordinates to chimpanzee ones using LiftOver tool (http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64/liftOver) and retrieved corresponding chimpanzee reference sequences (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/liftOver/hg19ToPanTro4.over.chain.gz).

After compiling the above data sets, we filtered out variants within human repeat regions or CpG islands (Wu et al. 2010). We removed variable sites with multiple alternative alleles, sites with strand ambiguity (A/T or G/C SNPs), sites prone to cytosine deamination (C>T or G>A SNPs) or sites not present in the 1KG data set to minimize impact of high error rate of genotype call from low-coverage data. Sites were also excluded if they have missing data among 17 modern humans or non-human samples (chimpanzee reference, Altai Neandertal and

Denisovan). This process led to variable number of SNPs for each aDNA sample, ranging from 0.47 to 6.36 million SNPs (**Supplementary Table 3.1**).

A population panel for the second phase analysis focused on fine-scale comparisons of population relationships among East Asians. For this purpose, we intersected array-based genotyping data from three studies: 938 individuals from 52 world-wide populations in the Human Genome Diversity Panel (HGDP) (Li et al. 2008), 30 Tibetans from near Lhasa, Tibet Autonomous Region in China (Wang et al. 2011), and 21 "high-altitude proxy" Sherpa samples from Khumbu, Nepal (Jeong et al. 2014). We also added genotypes of two contemporary Nepali Tibetan individuals used in the first phase. Autosomal SNPs with strand ambiguity (A/T or G/C) were excluded, as were any SNPs in which data were missing for more than one individual in any population. After further removing SNPs with no sequence information for any of five ACA samples, a total of 364,842 SNPs were retained for the downstream analysis.

10. Whole genome ancestry affiliation analysis

Principal component analysis (PCA) was performed using the smartpca program in the EIGENSOFT 6.0 package (Patterson et al. 2006). For the first phase of data analysis, PCs were calculated using all modern human samples including the ACA samples (**Supplementary Figure 3.5**). We kept SNPs with minor allele frequency (maf) ≥ 0.001, ranging 0.15-2.03 million SNPs. After performing PCA for each ACA sample separately, we merged results by applying a Procrustes transformation to PC1 and PC2, using "procGPA" function in an R package "shapes" (Dryden 2014). For the second phase of data analysis, PCs were calculated using the following contemporary East Asian populations: Cambodian, Dai, Daur, Han, Hezhen, Japanese, Lahu, Miao, Mongola, Naxi, Oroqen, She, Tu, Tujia, Xibo, Yakut, Yi from the HGDP, and Tibetan and

the Sherpa. After calculating PCs using 357K SNPs with maf ≥ 0.01, ancient ACA samples were projected onto PC plane using "*lsqproject: YES*" option (**Figure 3.2**).

We next conducted a model-based genetic clustering analysis as implemented in the sNMF (Frichot et al. 2014) and NGSadmix (Skotte et al. 2013) programs (**Figure 3.3, Supplementary Figures 3.6 and 3.7**). For the first phase of data analysis, we chose the sNMF program for properly modeling the haploid nature of our data set of eight ACA individuals. Thirty individuals for each of four 1KG populations, YRI (Yoruba in Ibadan, Nigeria), TSI (Toscani in Italy), CDX (Chinese Dai in Xishuangbanna, China) and STU (Sri Lankan Tamil from the UK), and all 17 high coverage sequenced contemporary humans from the high-coverage data (1 Yoruba, 1 Sardinian, 1 Karitiana, 2 Papuan, 2 Australian aborigine, 2 Han, 2 Dai, 4 Sherpa and 2 Tibetans) were chosen for a population panel. We randomly chose one of two alleles for genotypes of all contemporary samples to match haploid nature of the ACA aDNA samples. After this step, SNPs with at least three copies of minor allele were retained. SNPs were pruned based on linkage disequilibrium (LD) using PLINK v1.0.7 (Purcell et al. 2007), by randomly removing a SNP in each pair with $r^2 > 0.2$, and further thinned, leaving 40K to 100K SNPs for the analysis, depending on the ACA sample. We ran 50 replicates with random seeds for the number of clusters (K) ranging from 2 to 6 and chose one run with the smallest cross entropy value for each K value. For the second phase of data analysis, we analyzed all five samples of our more deeply sequenced samples together using genotype likelihoods as implemented in the NGSadmix program. HGDP Yoruba, Sardinian, Pathan, Papuan, Pima, Karitiana, Cambodian, Dai, Daur, Han, Hezhen, Japanese, Lahu, Miao, Mongola, Naxi, Oroqen, She, Tu, Tujia, Xibo, Yakut, Yi and Tibetans and the Sherpa were included in the analysis,

together with the five ACA samples. For each of the K values from 2 to 9, a replicate with the highest likelihood was chosen from 50 replicates with random seeds.

We estimated genetic affinity of the ACA samples with contemporary human populations in our data set using outgroup-f_3 statistic (Raghavan et al. 2014) using either 1KG YRI or HGDP Yoruba as an outgroup, using the same SNP set we used for PCA (**Figure 3.4, Supplementary Figures 3.8 and 3.9**). A large value of outgroup-f_3 statistic suggests a higher genetic affinity between two groups. We used *qp3Pop* program in the ADMIXtools v2 package (Patterson et al. 2012) to calculate f_3 statistics and associated standard errors. Outgroup-f_3 statistics were also used to compare the genetic relationships between contemporary East Asian populations (**Supplementary Figure 3.12**).

The D statistic was used as a statistical comparison of the affinity of the ACA samples to Tibetans and the Sherpa, as well as to other Tibeto-Burman populations (i.e., HGDP Yi, Naxi, and Tujia). Specifically, D (Yoruba, ACA; Tibeto-Burman, Tibetan/Sherpa) was calculated using qpDstat program in the ADMIXtools v2 package (**Figure 3.4 and Supplementary Figure 3.10**). The D statistic was also used to compare the genetic relationships of contemporary East Asian populations (**Supplementary Figure 3.12**).

11. High altitude adaptation allele analysis

We tested if ACA aDNA samples share adaptive haplotypes in the *EGLN1* (egl-9 family hypoxiainducible factor 1) and *EPAS1* (endothelial PAS domain protein 1) genes (**Supplementary Table 3.2**), which are common in contemporary Tibetans with strong signatures of recent positive selection (Beall et al. 2010; Simonson et al. 2010; Yi et al. 2010). Specifically, we determined if the ACA aDNA samples have derived alleles in two

nonsynonymous SNPs (rs12097901 and rs186996510) in the *EGLN1* gene (Xiang et al. 2013; Lorenzo et al. 2014) or in 20 SNPs tagging the *EPAS1* haplotype in Tibetans (Huerta-Sánchez et al. 2014).

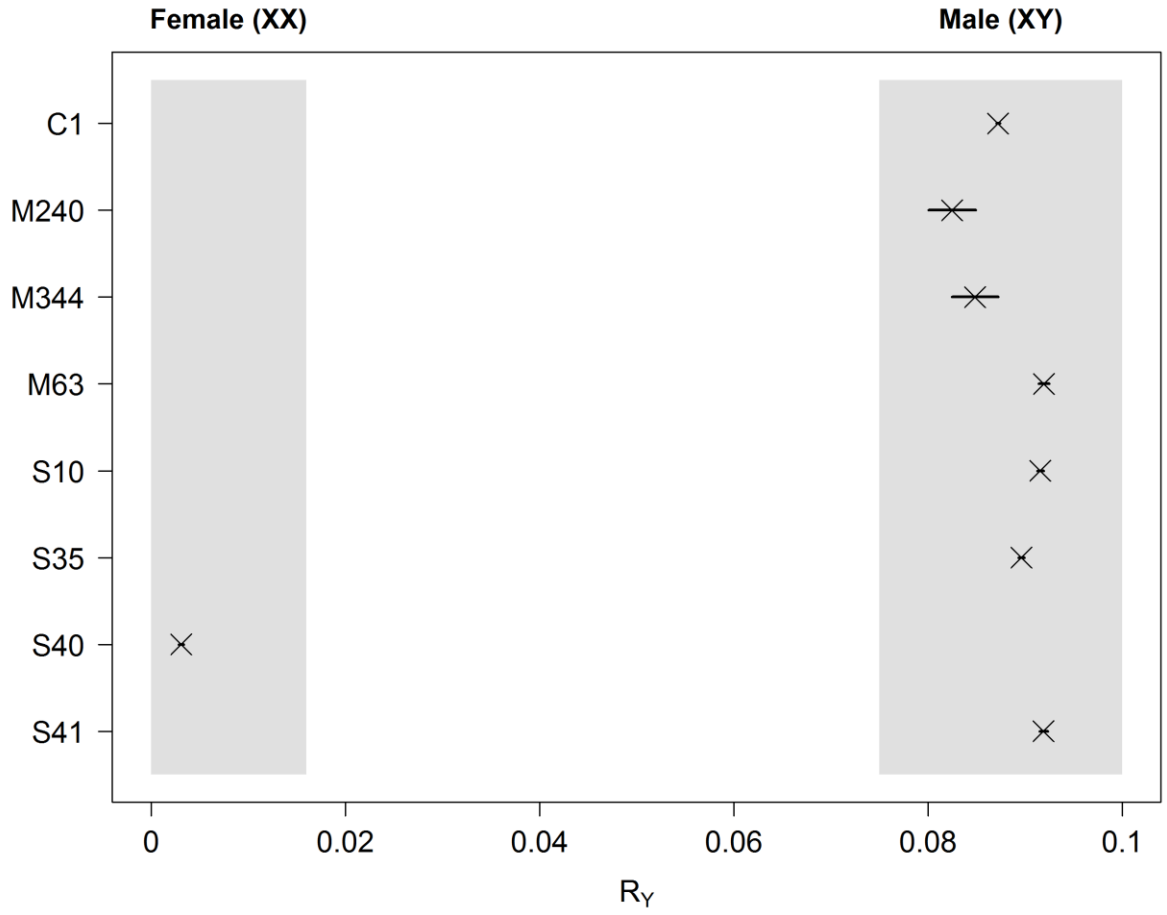
12. Mitochondrial haplogroup analysis

Consensus mtDNA sequences were called from sequence reads for all eight samples (21-1311x coverage; **Supplementary Table 3.1**) using SAMtools (Li et al. 2009). Then, we queried haplogroup information of the consensus mtDNA sequences (**Supplementary Table 3.3**) using HaploGrep (Kloss-Brandstätter et al. 2011), a web-based tool based on PhyloTree build 16, a database of > 20,000 human mtDNA sequences (Van Oven and Kayser 2009).

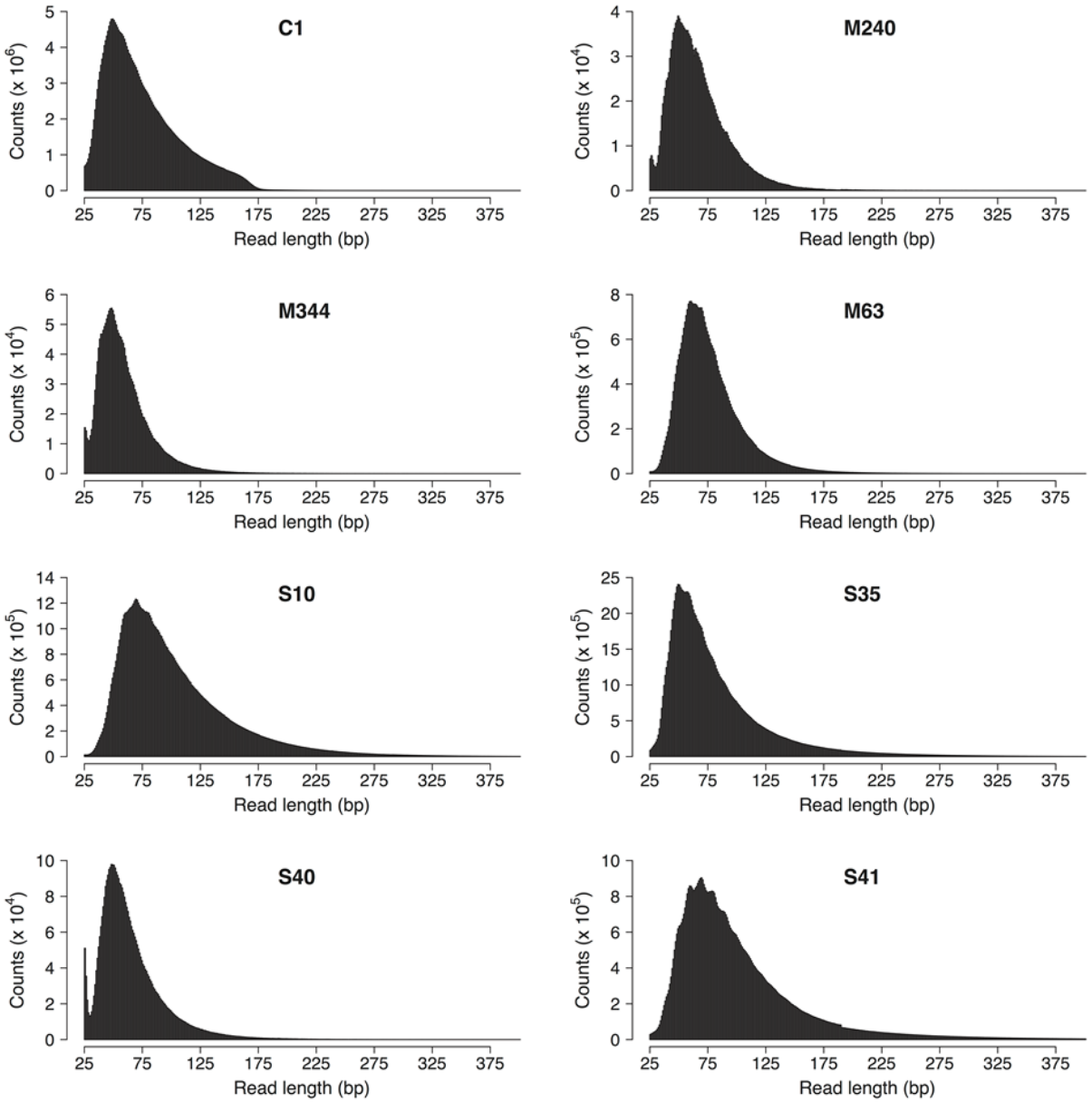
13. Y chromosome haplogroup analysis

Reads mapped to 519 informative bi-allelic Y-chromosome SNPs were piled up using the GATK v2.7-4 UnifiedGenotyper module. Y haplogroup (**Supplementary Table 3.4**) was manually assigned based on the informative SNPs and their associated information, retrieved from the cleantree program (Ralf et al. 2015).

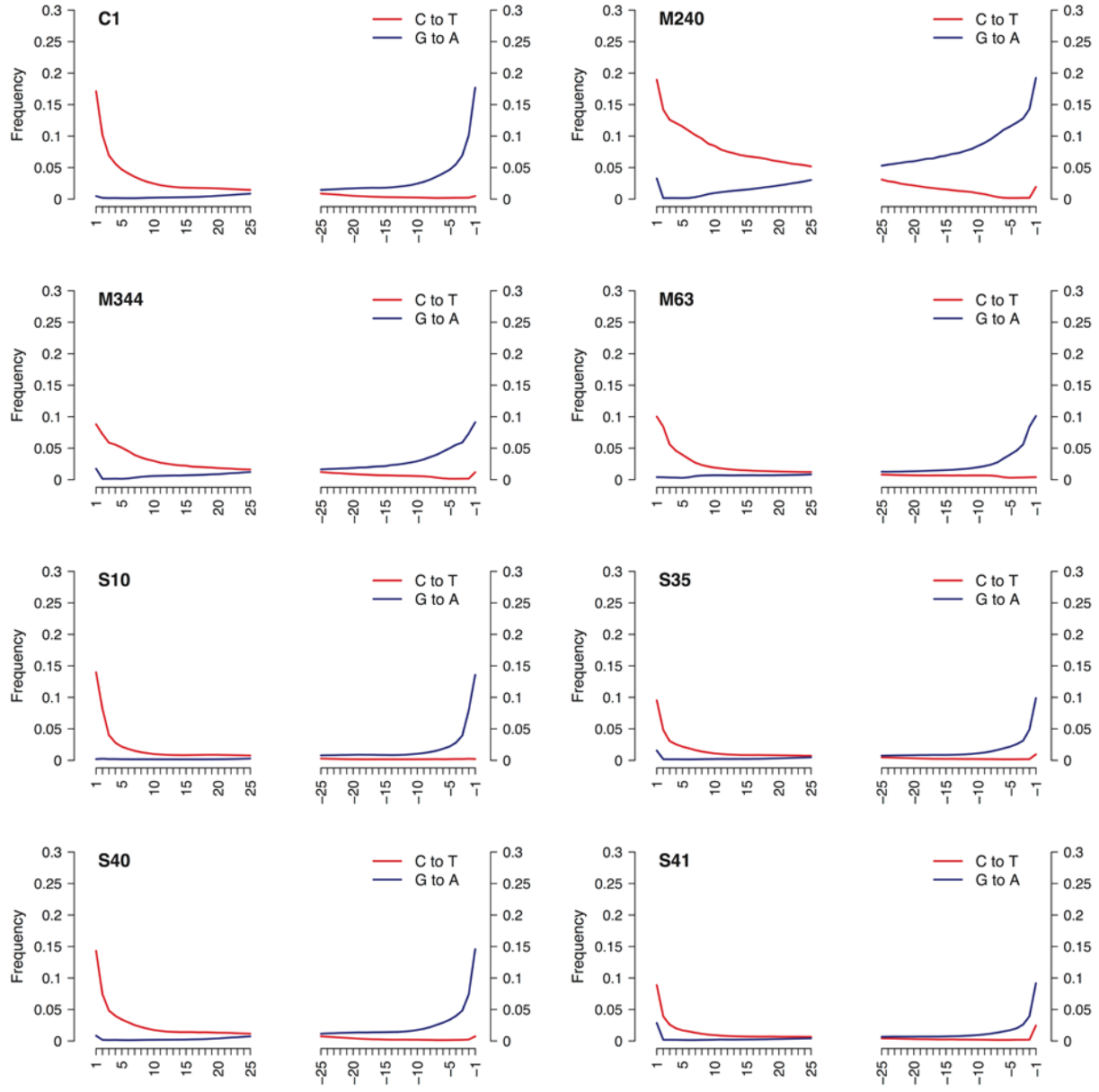
Supplementary Figure 3.1: Genetic sex assignment for the ACA individuals. Seven of the eight individuals, including one child (M63), are identified as male.



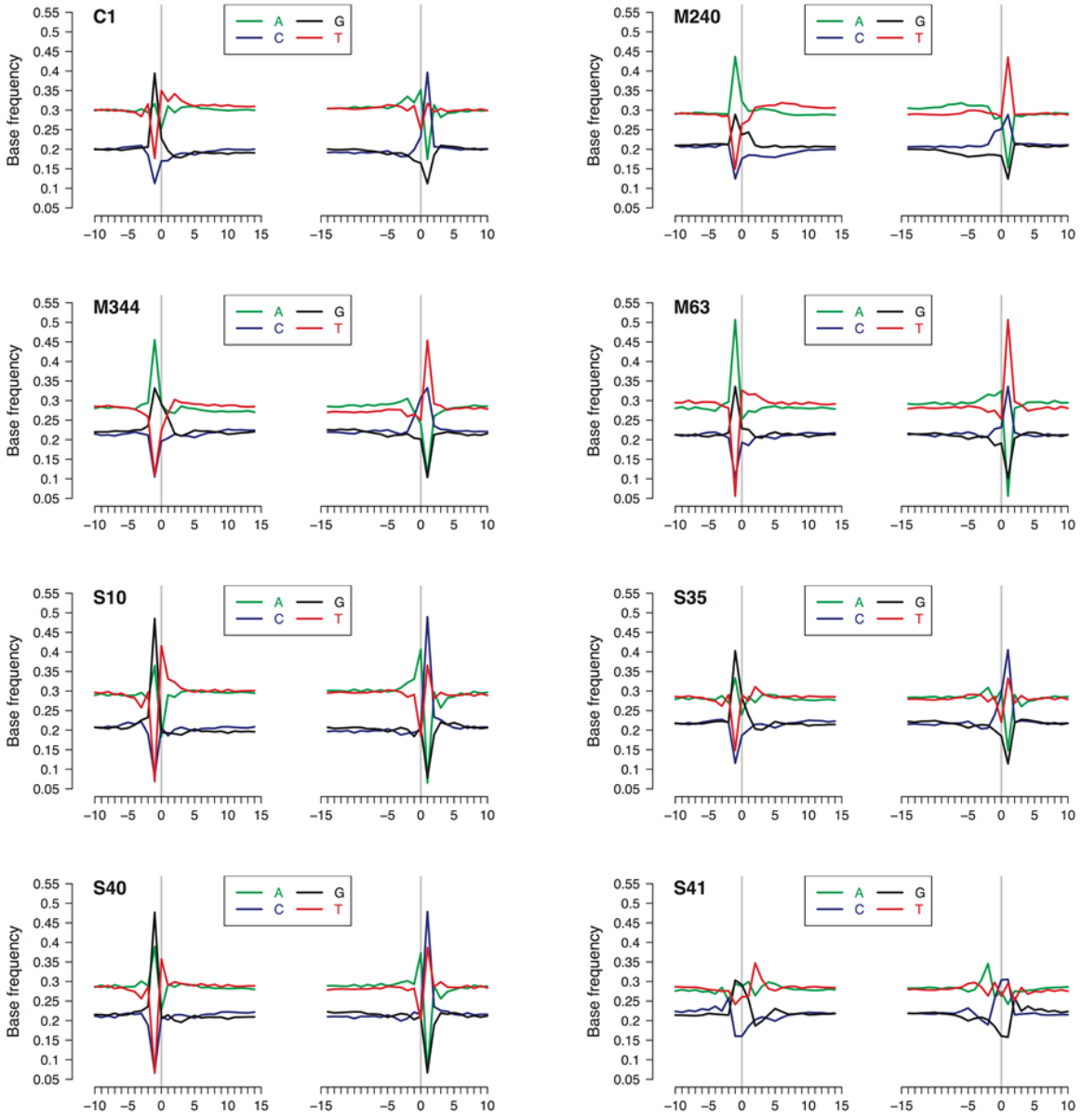
Supplementary Figure 3.2: Histogram of aDNA fragment lengths (bp) in the eight ACA samples. Data includes all non-duplicate reads ≥ 25 bp and uniquely mapped to the human reference genome (hg19) with a mapping quality score ≥ 30. For subsequent data analysis, reads < 35 bp were discarded.



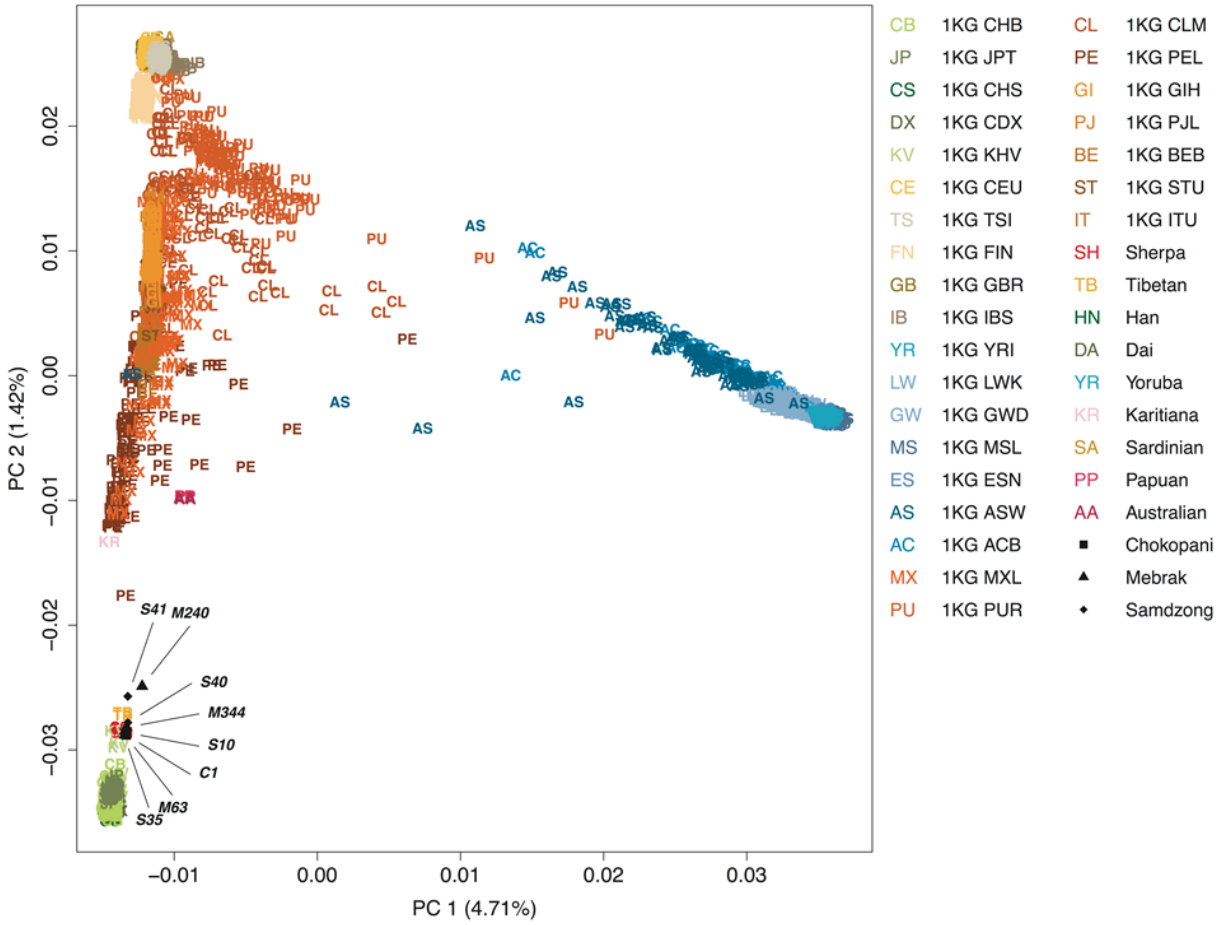
Supplementary Figure 3.3: Proportion of C>T and G>A substitutions in human DNA across DNA fragments in the ACA samples. C>T substitutions increase at the 5'-end and G>A substitutions increase at the 3'-end, as expected for authentic ancient DNA.



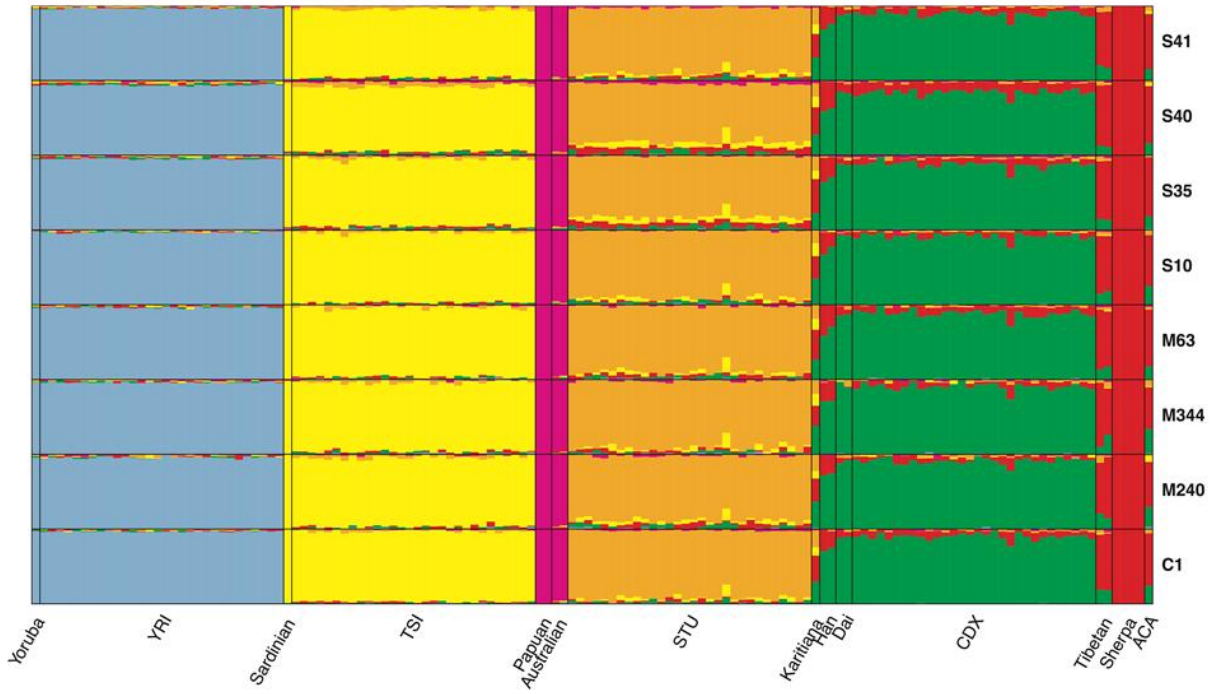
Supplementary Figure 3.4: Base frequencies flanking human DNA reads from ACA samples. An excess of purines (A or G) at -1 position of 5'-end and a corresponding excess of pyrimidine (C or T) at +1 position of 3'-end of human DNA in the ACA samples is consistent with depurination and hydrolysis breakage patterns characteristic of ancient DNA.



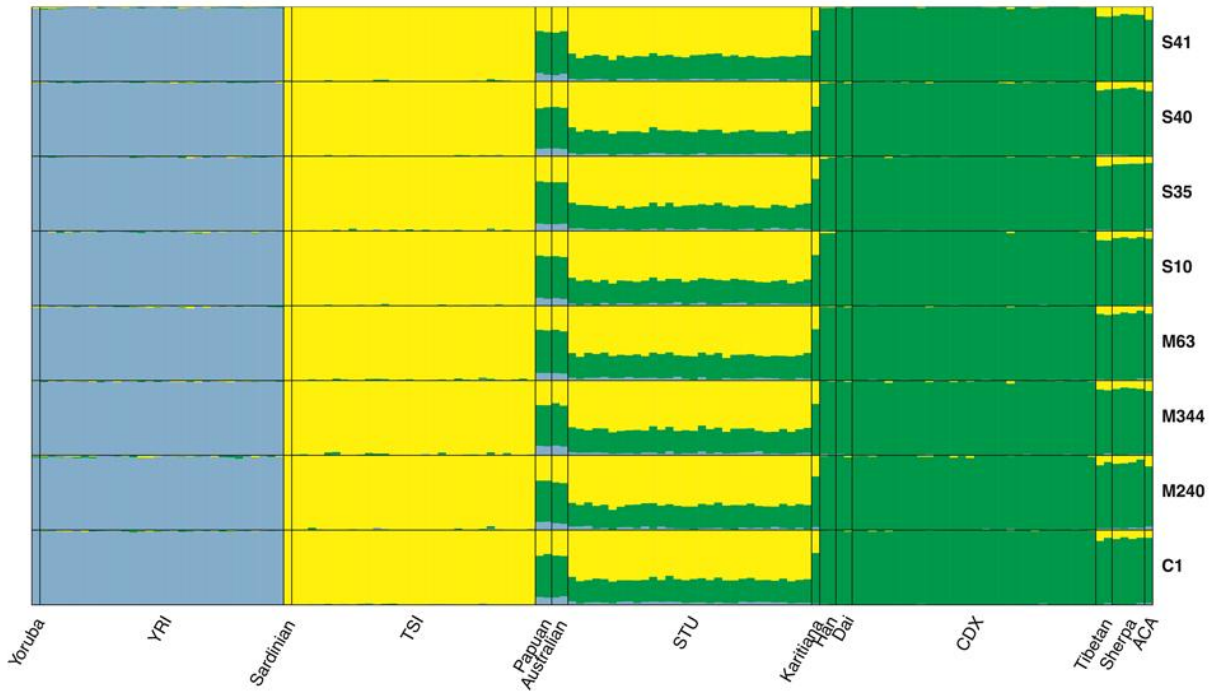
Supplementary Figure 3.5: PCA of global populations and ancient ACA samples using first phase sequencing data. All eight ACA samples cluster with East Asians. PC1 and PC2 were calculated using all contemporary samples and ACA samples. PCA was performed for each aDNA sample separately and the results were summarized using the Procrustes analysis as implemented in R package "shapes".



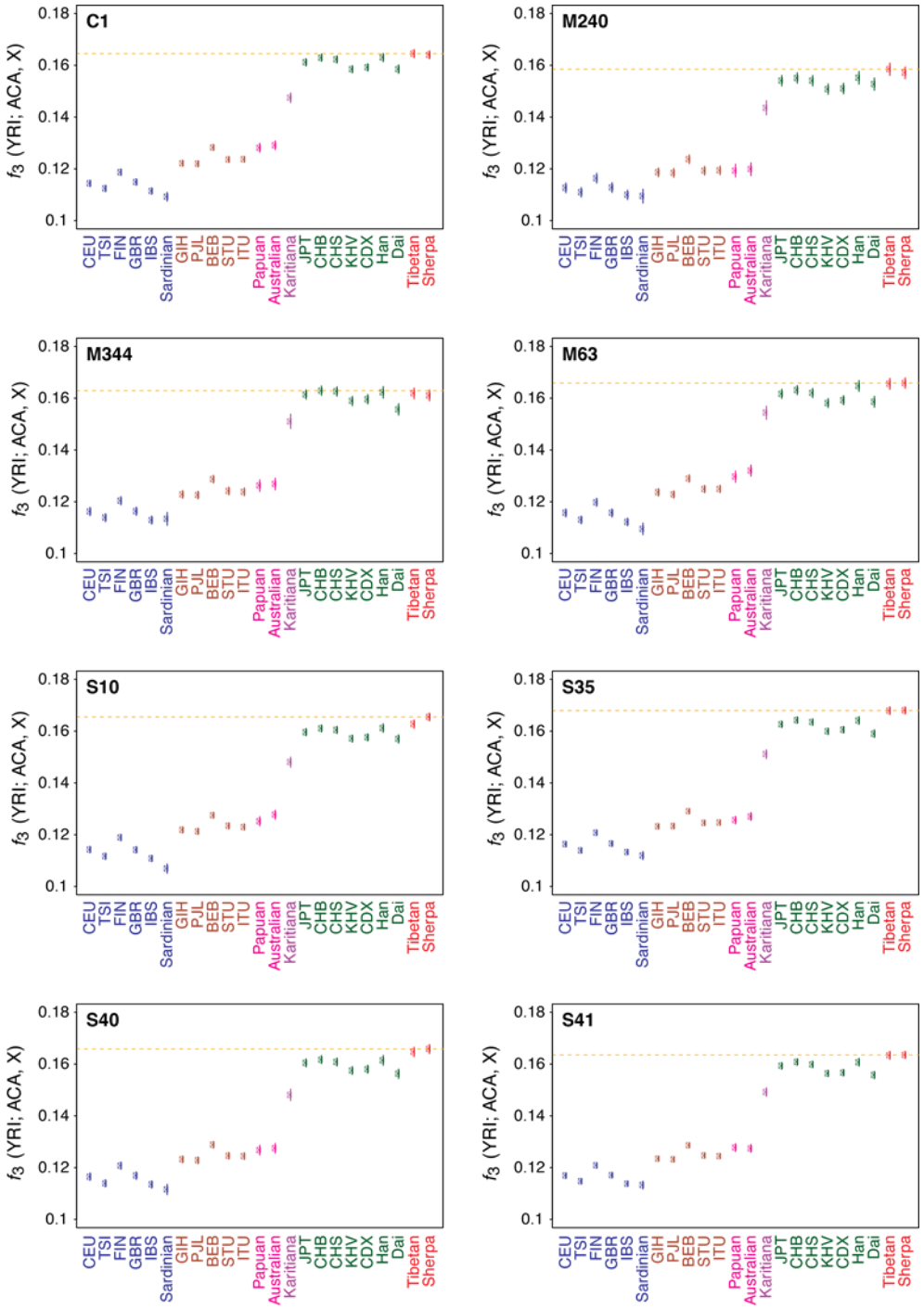
Supplementary Figure 3.6: Unsupervised genetic clustering with six ancestral populations (K=6). Six ancestries were allocated, respectively, to Africans (YRI [Yoruba in Ibadan, Nigeria]; blue), Europeans (TSI [Toscani in Italy]; yellow), Australasians (Papuan and Australian aborigine; magenta), South Asians (STU [Sri Lankan Tamil in the UK]; orange), lowland East Asians (CDX [Chinese Dai in Xishuangbanna]; green), and high altitude East Asians (Sherpa and Tibetans; red). All eight ACA samples share most of their genetic ancestry with high altitude East Asians.



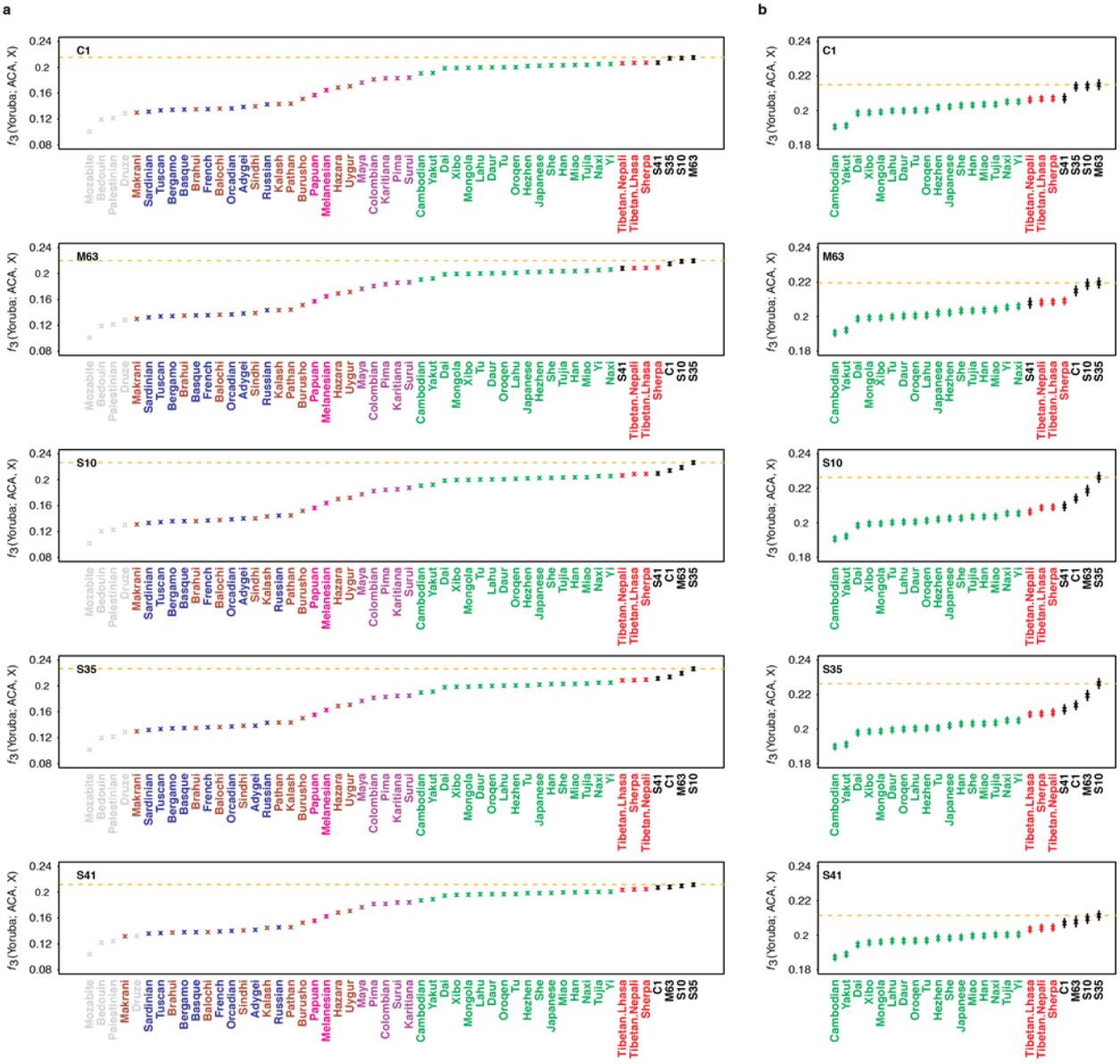
Supplementary Figure 3.7: Unsupervised genetic clustering with three ancestral populations (K=3). Three ancestries were allocated to Africans (YRI [Yoruba in Ibadan, Nigeria]; blue), Europeans (TSI [Toscani in Italy]; yellow) and East Asians (CDX [Chinese Dai in Xishuangbanna]; green). 17 highcoverage genomes of Yoruba, Sardinians, Papuans, Australian aborigine, Karitiana, Han, Dai, Tibetans, and Sherpa are also included for comparison. All eight ACA samples share most of their genetic ancestry with East Asians.



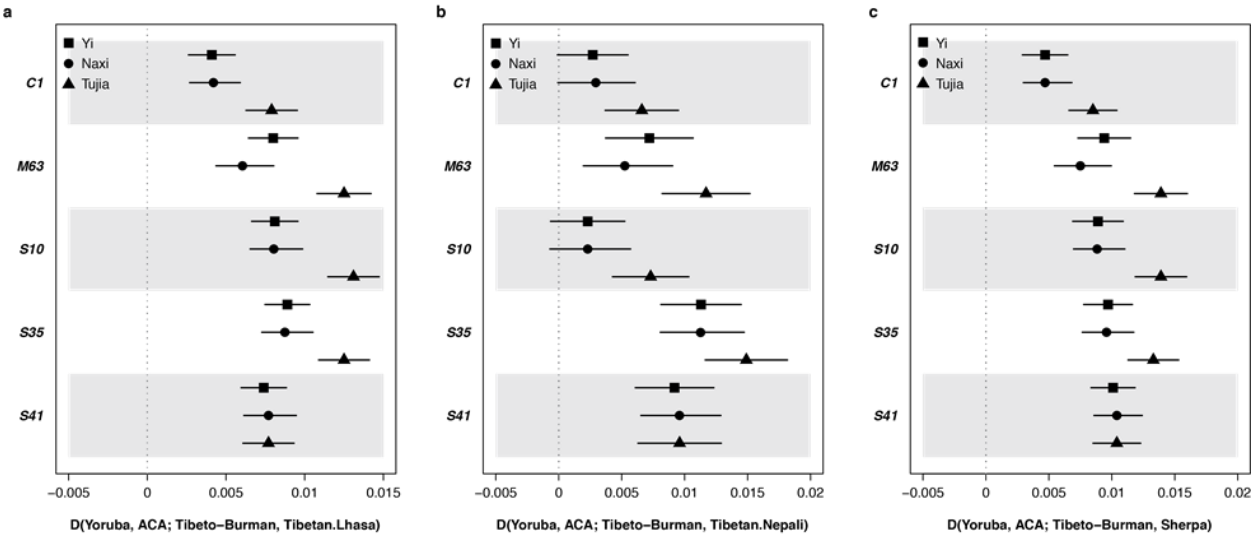
Supplementary Figure 3.8: Genetic affinity (f_3) of ACA samples and global populations using genome wide SNP data obtained from first phase sequencing. Genetic affinity with ancient samples is measured by f_3 (Yoruba; ACA, X). Vertical bars represent ± 1 standard error (SE) from 5 cM block jackknifing. For all ACA samples either Sherpa or Tibetans were the closest modern population (a larger f_3 value indicates a closer relationship).



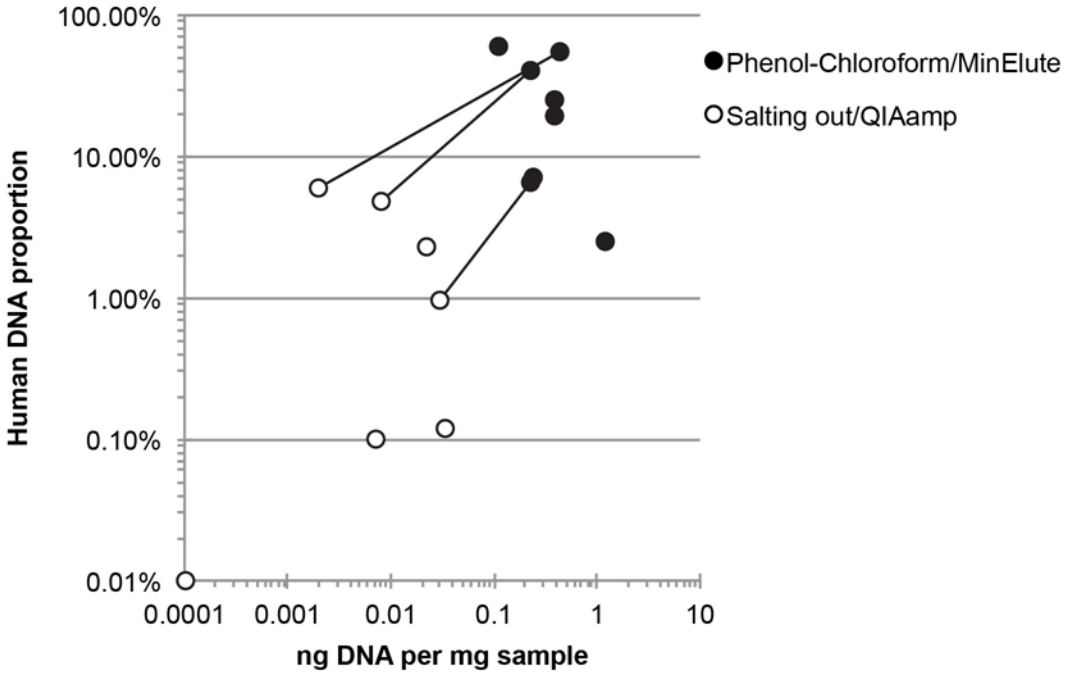
Supplementary Figure 3.9: Genetic affinity (f_3) of ACA samples and global populations using genome wide SNP data obtained from second phase sequencing. (a) Results for all populations and (b) East Asians in detail are provided. Genetic affinity with ancient samples is measured by f_3 (HGDP Yoruba; ACA, X). Vertical bars represent ± 1 standard error (SE) from 5 cM block jackknifing. For all ACA samples either Sherpa or Tibetans were the closest modern population (a larger f_3 value indicates a closer relationship).



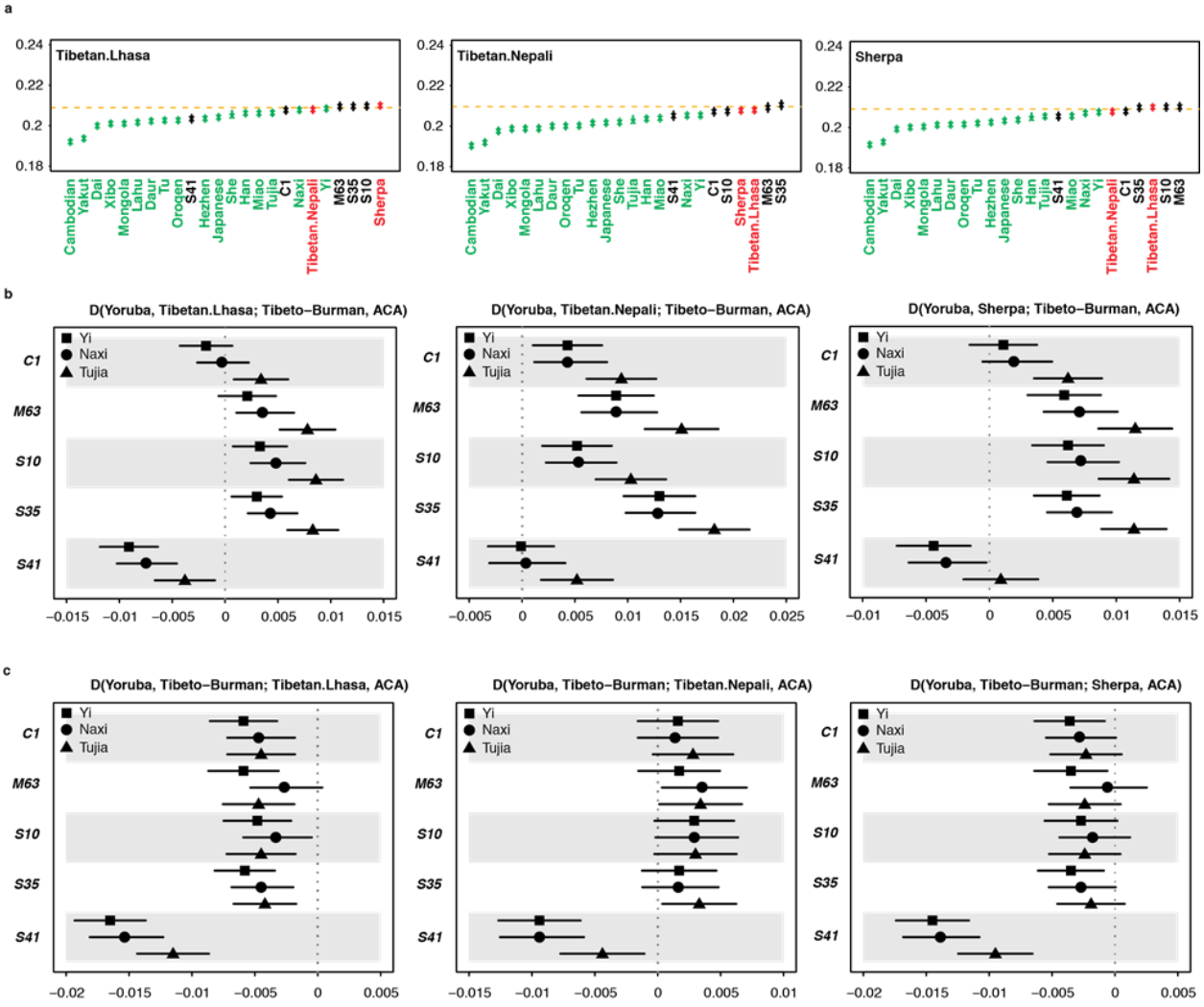
Supplementary Figure 3.10: Genetic affinity (D) of ACA samples to high altitude East Asian and lowland Tibeto-Burman speakers using genome wide SNP data obtained from second phase sequencing. Genetic affinity with ACA samples is measured by D(Yoruba, ACA; Tibeto-Burman, high altitude East Asian), where the Tibeto-Burman populations are Ni, Yaxi, and Tujia, and the high altitude East Asian populations are (a) Tibetans from near Lhasa (Tibetan.Lhasa), (b) Tibetans from Nepal (Tibetan.Nepali), and (c) Sherpa. Positive values indicate greater genetic affinity with high altitude East Asians; negative values indicate greater affinity with lowland Tibeto-Burman speakers. Horizontal bars represent ± 1 SE from 5 cM block jackknifing. Regardless of proxy population, the ACA individuals consistently show a significantly greater genetic affinity with high altitude East Asian populations.



Supplementary Figure 3.11: Comparison of ancient DNA extraction methods with respect to total DNA yield and human DNA content. On average, DNA extraction using the phenol-chloroform/MinElute method yields 27.5-fold more DNA and a 13.1-fold higher proportion of human DNA than the salting out/QIAamp method. Paired DNA extracts from the same sample are connected by a line.



Supplementary Figure 3.12: Genetic affinity (f_3 and D) of high altitude East Asian and lowland Tibeto-Burman populations to other East Asian populations. (a) Genetic affinity is measured by f_3(HGDP Yoruba; high altitude East Asian, X). In general, high altitude East Asian populations show high genetic affinity with each other and with the ancient ACA samples. The reduced genetic affinity with S41 may be due to a small proportion of west Eurasian ancestry in this individual, as evidenced by the NGSAdmix results in **Figure 3.3**. (b) Genetic affinity is measured by D(HGDP Yoruba, high altitude East Asian; Tibeto-Burman, ACA). In general, high altitude East Asian populations show greater genetic affinity with ancient ACA individuals than with Tibeto-Burman speakers, but this difference is not consistent or significant with respect to C1 and S41. (c) Genetic affinity is measured by D(HGDP Yoruba, Tibeto-Burman; high altitude East Asian, ACA). Tibeto-Burman speakers show small negative D values with Tibetans from near Lhasa and Sherpa, suggesting their greater genetic affinity to them than to ACA samples. However, all values are close to zero, especially those with Sherpa. Small positive D values with Nepali Tibetans may be due to a low proportion of South Asian ancestry in these Nepali Tibetan samples. Horizontal bars in (b) and (c) represent ± 1 SD.



Supplementary Table 3.1: Sequencing output and summary of data filtering and quality statistics for ACA samples.

	C1	M63	M240	M344
Total reads[a]	1,918,868,106	1,384,178,962	28,064,830	33,228,833
Merged reads[b]	1,867,605,894	1,284,061,577	25,268,226	32,566,882
	(97.3%)	(92.8%)	(90.0%)	(98.0%)
Uniquely mapped to hg19	631,145,688	263,725,693	2,519,316	2,577,894
	(32.9%)	(19.1%)	(9.0%)	(7.8%)
Uniquely mapped & length ≥ 35 bases	594,700,000	262,138,360	2,290,630	2,240,052
	(31.0%)	(18.9%)	(8.2%)	(6.7%)
Uniquely mapped & length ≥ 35 bases	277,205,013	41,104,521	1,991,595	2,176,982
& no PCR duplicate	(14.4%)	(3.0%)	(7.1%)	(6.6%)
Median fragment length (bp)[a]	66 (66)	71 (70)	61 (62)	53 (55)
Endogenous[d]	100.0%	99.2%	99.8%	94.4%
Nuclear DNA coverage[e]	7.253x	1.048x	0.045x	0.044x
mtDNA coverage[e]	1077.4x	313.1x	72.5x	20.8x
First Phase SNPs[f]	2,442,236	759,039	468,381	521,344
Second Phase SNPs[g]	362,780	222,340	-	-

	S10	S35	S40	S41
Total reads[a]	456,744,916	244,594,518	167,200,988	138,904,789
Merged reads[b]	441,915,476	229,752,968	162,592,938	123,471,628
	(96.8%)	(93.9%)	(97.2%)	(88.9%)
Uniquely mapped to hg19	184,876,748	146,037,858	5,180,436	75,837,156
	(40.5%)	(59.7%)	(3.1%)	(54.6%)
Uniquely mapped & length ≥ 35 bases	184,225,674	142,647,353	4,355,340	74,957,456
	(40.3%)	(58.3%)	(2.6%)	(54.0%)
Uniquely mapped & length ≥ 35 bases	97,275,054	115,663,044	4,219,420	57,044,922
& no PCR duplicate	(21.3%)	(47.3%)	(2.5%)	(41.1%)
Median fragment length (bp)[a]	90 (87)	69 (67)	57 (59)	86 (83)
Endogenous[d]	100.0%	98.4%	99.4%	99.4%
Nuclear DNA coverage[e]	3.264x	3.493x	0.090x	2.072x
mtDNA coverage[e]	249.5x	784.8x	27.8x	1311.0x
First Phase SNPs[f]	1,982,791	6,358,061	1,126,324	5,259,226
Second Phase SNPs[g]	346,210	348,165	-	311,680

Notes:

[a]Calculated from total first phase and second phase reads >25 bp mapping to hg19 after duplicate removal. Median fragment length of analysis-ready reads is given in parentheses.

[b]Adapter trimming, chimera removal, and read merging performed using MergeReadsFastQ_cc.py.

[c]Uniquely mapped to hg19, >35 bp, no PCR duplicate.

[d]Estimated proportion of endogenous human reads (Bayesian estimate from Fu et al 2013).

[e]Mean coverage.

[f]Total SNPs in the first phase analysis-ready genetic data set.

[g]Number of SNPs covered with reads, used for PCA, f_3-statistic, and *D* analyses. Genotype likelihoods for NGSadmix were calculated separately.

Supplementary Table 3.2: Read mapped to *EGLN1* and *EPAS1* SNPs

SNP	Position	Ref	Alt	C1	M63	S10	S35	S41
EGLN1 (chr1)								
rs12097901	231557255	C	**G**	**2G**	1C	**1G**	**3G**	
rs186996510	231557623	G	**C**	**3C**		**3C**	**7C**	**4C**
EPAS1 (chr2)								
rs115321619	46567916	G	**A**	1G		2G	4G, **2A**	**1A**
rs73926263	46568680	A	**G**	3A	1A	3A	2A, **1G**	**1G**
rs73926264	46569017	A	**G**	16A		3A	2A, **3G**	**2G**
rs73926265	46569770	G	**A**	5G	2G	2G	**2A**	**3A**
rs55981512	46570342	G	**A**	3G		1G	2G, **5A**	**2A**
rs149306391	46571017	C	**G**	4C	2C	1C	2C, **1G**	**1G**
.	46571435	G	**C**	7G		5G	1G, **4C**	**2C**
rs188801636	46577251	T	**C**	9T		5T	**1C**	1T
.	46579689	A	**G**	4A	2A	3A	3A, **3G**	**2G**
rs189807021	46583581	G	**A**	1G		1G	2G, **1A**	**2A**
.	46584859	A	**G**	1A	2A	1A	1A, **2G**	**1G**
rs150877473	46588019	C	**G**	7C		1C	1C, **1G**	1C, **2G**
rs142826801	46588331	G	**C**	4G		2G	4G, **1C**	**1C**
rs74898705	46589032	C	**T**	2C		3C	2C, **2T**	**2T**
rs141366568	46594122	A	**G**	10A	2A	2A	2A, **1G**	**4G**
rs116062164	46597756	A	**C**	7A	2A	1A	1A, **2C**	**2C**
rs141426873	46598025	C	**G**	6C		2C	**4G**	1C, **4G**
rs116611511	46600030	A	**G**	13A	1A	4A	1A, **1G**	**2G**
.	46600358	A	**G**	10A	3A	7A	5A, **2G**	**2G**
rs58160876	46600661	A	**C**	7A	5A	2A	1A, **1C**	**1C**

Notes:
Derived alleles are shown in bold type. Insufficient *EGLN1* and *EPAS1* sequence coverage was obtained for samples M240, M344, and S40; these samples have been omitted from the table.

Supplementary Table 3.3: Mitochondrial haplogroup assignments for ACA dental samples

Sample	Haplogroup	Mismatches	Additional mutations	Haplogroup Rank[a]	Haplogroup frequency in Tibetans[b]	Haplogroup frequency in Indians and Pakistanis[c]	Haplogroup frequency in Han Chinese[d]
C1	D4j1b	-	12	0.965	12.72% (D4)	0.00% (D4)	14.03% (D*)
M63	M9a1a1c1b1a	5899.XC	1	0.976	16.30% (M9a1a)	0.04% (M9a)	2.14% (M9a)
M240	M9a1a2	16362T	3	0.975	16.30% (M9a1a)	0.04% (M9a)	2.14% (M9a)
M344	Z3a1a	249A	8	0.940	1.92% (Z)	0.00% (Z)	3.22% (Z)
S10	M9a1a1c1b1a	5899.XC	3	0.959	16.30% (M9a1a)	0.04% (M9a)	2.14% (M9a)
S35	M9a1a	16362T	8	0.966	16.30% (M9a1a)	0.04% (M9a)	2.14% (M9a)
S40	F1c1a1a	249A	6	0.955	2.16% (F1c)	0.04% (F1c)	1.07% (F1c)
S41	F1d	146T 249A	10	0.865	6.81% (F1)	0.00% (F1)	0.00% (F1)

[a]Haplogroup rank calculated using the HaploGrep algorithm, as previously described (Kloss-Brandstätter et al. 2011).
[b]Estimated from 6,109 individuals from 41 Tibetan populations (Qi et al. 2013).
[c]Estimated from 2,440 individuals from Indian (n=2295) and Pakistani (n=145) populations, excluding Tibeto-Burman speakers (Metspalu et al. 2004).
[d]Estimated from 1,119 individuals in Han Chinese populations (Wen et al. 2004). D* does not include D5 or D5a.

Supplementary Table 3.4: Y chromosome haplogroup assignments for ACA samples

Sample[a]	Haplogroup[b]	Haplogroup frequency in Tibetans[c]
C1	O-M117	29.86%
M63[d]	nd	nd
S10	O-M117	29.86%
S35	O-M117	29.86%
S41	D	54.50%

[a]Y-chromosome haplotyping was only performed on data generated during the second phase of the study.
[b]Haplogroup was assigned based on informative SNPs from a database associated with the cleantree program.
[c]Estimated from 2,354 individuals from 41 Tibetan populations (Qi et al. 2013).
[d]Not determined. Insufficient data for haplogroup assignment.

Supplementary Table 3.5: ACA DNA extraction and NGS library information

Sample	Sex[a]	Dentine (mg)	Extraction Location	Extraction Method[b]	DNA yield (ng/mg)[c]	NGS Library Location	NGS Sequencing Location	Human DNA content[d]	Median DNA Fragment Length (bp)
Chokhopani									
C1	M	101	LMAMR	1	0.385	LMAMR	U. Chicago	31.0%	66
C4	-	1940	LMAMR	2	0.033	Uppsala	Uppsala	0.12%	84
Mebrak									
M63	M	97	LMAMR	1	0.384	Uppsala	Uppsala	18.9%	71
M240	M	106	LMAMR	1	0.235	LMAMR	U. Chicago	8.2%	61
M294	-	1800	LMAMR	2	0.000	Uppsala	Uppsala	*	*
M344	M	112	LMAMR	1	0.226	LMAMR	U. Chicago	6.7%	53
M344	-	1610	LMAMR	2	0.029	Uppsala	Uppsala	0.93%	65
M458	-	1980	LMAMR	2	0.007	Uppsala	Uppsala	0.10%	75
Samdzong									
S10	M	117	LMAMR	1	0.224	Uppsala	Uppsala	40.3%	90
S10	-	1570	LMAMR	2	0.008	Uppsala	Uppsala	4.71%	78
S35	M	101	LMAMR	1	0.107	LMAMR	U. Chicago	58.3%	69
S39	-	1420	LMAMR	2	0.022	Uppsala	Uppsala	2.31%	48
S40	F	99	LMAMR	1	1.192	LMAMR	U. Chicago	2.5%	57
S41	M	104	LMAMR	1	0.434	LMAMR	U. Chicago	54.0%	86
S41	-	1530	LMAMR	2	0.002	Uppsala	Uppsala	5.98%	77

Notes:

*Insufficient DNA for successful library construction.

[a]Genetic sex: M = male; F = female; - = not analyzed. Only samples extracted using Extracted Method 1 were analyzed.

[b]Extraction methods: 1 = phenol-chloroform/MinElute; 2 = salting out/QIAamp. Because of the dramatic differences in DNA extraction performance, only samples analyzed using Extraction Method 1 were further analyzed for ancestry information in this study.

[c]Determined by measurement of 1µl of DNA extract using a Qubit fluorometer high sensitivity DNA assay and normalized to ng of DNA recovered per mg of dentine.

[d]Reads uniquely mapped to hg19 with a length >35 bp.

Supplementary Table 3.6: Outline of DNA sequencing scheme for data generated in this study

Study	Quantity[a]	Platform	Sample[b]	Location
First Phase	2	HiSeq 2500 (rapid run mode)	C1, M240, M344, S35, S40, S41	U. Chicago
	1	HiSeq 2000	M63[c]	Uppsala
	1	HiSeq 2000	S10[c]	Uppsala
Second Phase	5	HiSeq 4000	C1	U. Chicago
	1	HiSeq 4000	S35, S41	U. Chicago
	1	HiSeq X-Ten	S10	Uppsala
	3	HiSeq X-Ten	M63	Uppsala

[a]Number of lanes sequenced.
[b]List of pooled samples sequenced in each experiment.
[c]M63 and S10 were each pooled and sequenced with other samples not included in this study (list not shown).

Supplementary Table 3.7: Genomic sequence coverage information

	Autosomal (2,881,033,286)			X (155,270,560)		
ACA ID	Sites	Total	Avg.	Sites	Total	Avg.
C1	2,618,065,082	20,895,219,390	7.253	138,902,915	582,701,156	3.752
M63	1,692,329,770	3,019,603,191	1.048	57,984,210	80,125,936	0.516
M240	125,330,641	129,470,638	0.045	3,358,010	3,418,825	0.022
M344	122,583,409	126,989,101	0.044	3,158,277	3,212,457	0.021
S10	2,518,355,450	9,404,687,422	3.264	117,260,089	259,091,314	1.669
S35	2,496,870,820	10,065,483,731	3.493	115,335,284	261,753,041	1.686
S40	246,021,289	260,648,390	0.090	12,954,861	13,677,628	0.088
S41	2,253,399,978	5,969,434,907	2.072	90,940,543	154,670,556	0.996

	Y (59,373,566)			mtDNA (16,571)		
ACA ID	Sites	Total	Avg.	Sites	Total	Avg.
C1	15,059,983	71,760,677	1.209	16,570	17,853,125	1077.4
M63	6,293,395	10,620,215	0.179	16,570	5,188,142	313.1
M240	383,924	419,325	0.007	16,551	1,202,384	72.560
M344	367,816	418,666	0.007	16,195	345,060	20.823
S10	12,884,603	32,067,300	0.540	16,569	4,134,151	249.5
S35	12,265,592	33,313,350	0.561	16,571	13,005,585	784.8
S40	92,788	151,166	0.003	16,190	460,958	27.817
S41	9,730,272	20,430,817	0.344	16,571	21,724,347	1311.0

Notes:
Sites: the number of sites covered at least once; Total: the number of all bases mapped to each chromosome; Average: average number of reads per site in each chromosome.

CHAPTER 4: DEEP HISTORY OF EAST ASIAN POPULATIONS REVEALED THROUGH GENETIC ANALYSIS OF THE AINU[3]

4.1: Abstract

Despite recent advances in population genomics, much remains to be elucidated with regard to East Asian population history. The Ainu, a hunter-gatherer population of northern Japan and Sakhalin island of Russia, are thought to be key to elucidating the prehistory of Japan and the peopling of East Asia. Here, we study the genetic relationship of the Ainu with other East Asian and Siberian populations outside the Japanese archipelago using genome-wide genotyping data. We find that the Ainu represent a deep branch of East Asian diversity more basal than all present-day East Asian farmers. However, we did not find a genetic connection between the Ainu and populations of the Tibetan plateau, rejecting their long-held hypothetical connection based on Y chromosome data. Unlike all other East Asian populations investigated, the Ainu have a closer genetic relationship with northeast Siberians than with central Siberians, suggesting ancient connections among populations around the sea of Okhotsk. We also detect a recent genetic contribution of the Ainu to nearby populations, but no evidence for reciprocal recent gene flow is observed. Whole genome sequencing of contemporary and ancient Ainu individuals will be helpful to understand the details of the deep history of East Asians.

[3] Citation for chapter: Jeong C, Nakagome S and Di Rienzo A. 2016. "Deep history of East Asian populations revealed through genetic analysis of the Ainu." Genetics 202: 261-272.

4.2: Introduction

The rapid development of genomic technologies has greatly enhanced our understanding of the history of modern human dispersal out of Africa and the peopling of different continents (Li et al. 2008; Green et al. 2010; Reich et al. 2010). However, the history of populations in East Asia, including Siberia, remains poorly understood even though they account for a large fraction of the human population (Stoneking and Delfin 2010; Oota and Stoneking 2011). There is still no clear consensus regarding basic questions such as when, where and how many times modern humans migrated into East Asia and Siberia. For example, several previous studies based on contemporary samples concluded that one migration from south to north could explain the genetic structure of East Asians (Li et al. 2008; The HUGO Pan-Asian SNP Consortium 2009). However, recent studies indicate that this scenario is too simplistic: a recent study detected western Eurasian ancestry in an individual in southern Siberia 24,000 years ago as well as a substantial contribution of this ancestry to the gene pool of Native Americans (Raghavan et al. 2014), and several studies detected a clear genetic differentiation between northeast Siberians and central-south Siberians, with the latter being more closely related to northeast Asians (Rasmussen et al. 2010; Fedorova et al. 2013). Indigenous high-altitude populations of the Tibetan plateau provide another example of East Asian populations that do not fit into the simple one migration hypothesis. Tibetans and Sherpa show a divergent history from lowland East Asian populations such as Han Chinese (Jeong et al. 2014), and their adaptive haplotype spanning the *EPAS1* (endothelial PAS domain-containing protein 1) gene shares its ancestry with that of an archaic hominin (Huerta-Sánchez et al. 2014).

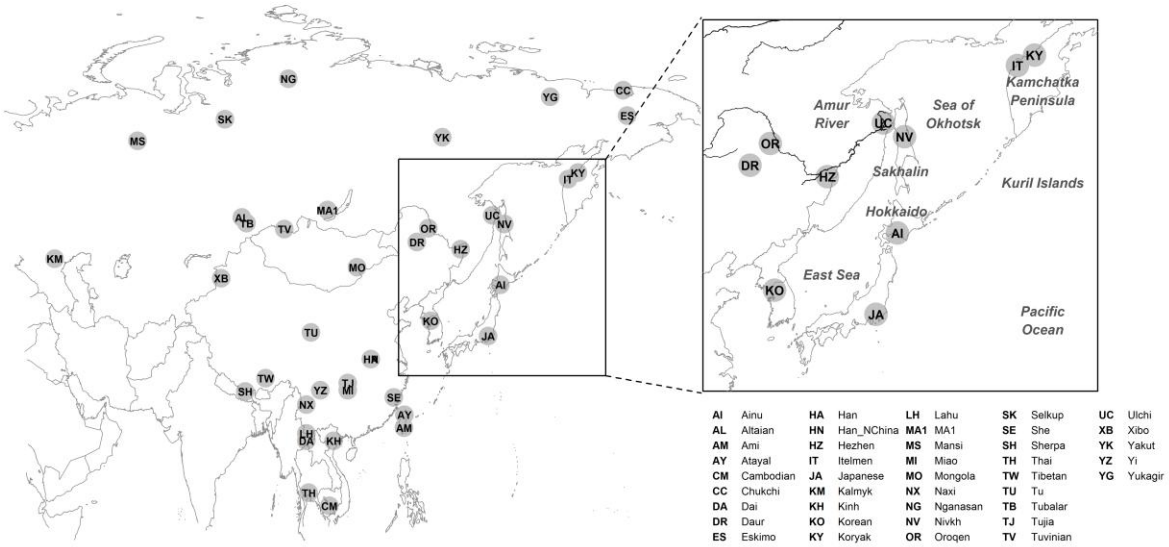
Considering robust evidence of human habitation in Arctic Siberia before the last glacial maximum (LGM) (Pitulko et al. 2004), it is possible that there were multiple expansions (from south to north) and contractions (from north to south) of human populations in mainland East Asia and Siberia over a long period of time, generating a complex pattern of genetic relationships among contemporary populations. Population isolates frequently provide critical information to understand the genetic structure of surrounding populations with more complex histories. For example, it has been proposed that Sardinians are key to understanding the changes in population structure in mainland Europe with the arrival of Neolithic farmers (Keller et al. 2012; Skoglund et al. 2012). The Onge people from the Andaman Islands have also been proposed as the best contemporary representatives of the "Ancestral South Indian" ancestry (Reich et al. 2009; Moorjani et al. 2013). Therefore, unraveling the genetic history of population isolates in East Asia and Siberia may provide new insights into the initial colonization of these regions.

The Ainu people are an indigenous population of Hokkaido, a northern island in the Japanese archipelago, and of the southern part of Sakhalin islands (**Figure 4.1**). They have been proposed by archaeologists, linguists and geneticists to be the direct descendants of prehistoric Japanese hunter-gatherers, associated with the Jomon pottery culture, dating back to 16,500 years before present (Hanihara 1991; Habu 2004). The dual structure model for the Japanese population (Hanihara 1991) envisions that the contemporary Japanese are a mixture of two distinct genetic sources, one from the indigenous Jomon hunter-gatherers and the other from East Asian rice farmers who first migrated into the archipelago approximately 2,300 years ago ("Yayoi culture"). Genetic data clearly support this model in two main regards. First, the genetic profile of the mainland Japanese reveals a strong signature of admixture, best modeled as a mixture of Ainu-related ancestry and continental East Asians (Jinam et al. 2012; Nakagome et al.

2015). Second, the Ainu are genetically closer to the Ryukyuan, who live in the southern-most islands of the Japanese archipelago, than to the mainland Japanese (Tajima et al. 2002; Matsukusa et al. 2010; Jinam et al. 2012; Koganebuchi et al. 2012). This suggests that inhabitants of the northern and southern-most parts of the archipelago were genetically most isolated from the incoming farmers, who first arrived in the central part of the archipelago. However, the origin of the Ainu people in the context of eastern Eurasian population history has not been thoroughly investigated using genome-scale variation data.

In this study, we investigated the genetic relationship of the Ainu with surrounding East Asian and Siberian populations, using genome-wide genotype data, and found that the Ainu gene pool is basal to all other East Asian populations. In addition, the Ainu show unusual patterns of excess genetic affinity with low-altitude East Asians and northeast Siberians, as well as signatures of genetic adaptations to their local environments and maritime hunter-gatherer life style.

Figure 4.1: Geographic location of East Asian and Siberian population samples used in this study. The zoom-in plot highlights the region around the Japanese archipelago and the sea of Okhotsk.



4.3: Materials and Methods

Genotype data

We obtained genome-wide genotyping data of world-wide populations from several previous publications. First, we obtained genotype data of 36 Ainu individuals from a previous study, typed on the Affymetrix Genome-wide Human SNP 6.0 array (Jinam et al. 2012). We estimated genetic relatedness for all pairs of Ainu individuals using PLINK v1.07 (Purcell et al. 2007), with 396,552 autosomal SNPs with minor allele frequency ≥ 0.05. We randomly removed one individual from each pair with coefficient of relationship (r) > 0.1875, which corresponds to relatedness between first cousins (r = 0.125) and half siblings (r = 0.25). This step removed 11 individuals, involved in 17 of 630 pairs, some of which correspond to first degree relatives (**Supplementary Figure 4.1**). Second, we used genotype data of 1,963 individuals from 183 world-wide populations, "Affymetrix Human Origins fully public dataset" (Lazaridis et al. 2014), typed on the Affymetrix Axiom® Genome-wide Human Origins 1 array (Patterson et al. 2012). We removed three individuals with genotype missing rate higher than 5% and kept SNPs with missing rate not exceeding our criteria in all 152 populations with ≥ 5 individuals. Specifically, we allowed one missing genotype in populations with < 20 individuals, and two missing genotypes in populations with ≥ 20 individuals. After this filtering, an overlap of 103,218 SNPs between the Ainu and Human Origins data sets was used for the majority of analyses ("WA" data set, for world-wide and Ainu data). Third, we overlapped the "WA" data set with genotype data for the Sherpa and Tibetans genotyped on Illumina arrays. Specifically, we took 21 Sherpa individuals described as the "High-altitude proxy" samples in a previous study (Jeong et al. 2014) and 30 Tibetans from near Lhasa, Tibet Autonomous Region in China (Wang et al. 2011). We

used the overlapping 45,513 SNPs ("WHA" data set, for world-wide, high-altitude and Ainu data) for most of the analyses along with the "WA" data set to investigate the genetic relationships of the Ainu and the high-altitude East Asians. Fourth, we overlapped the "WHA" data set with the genotype data of two Nivkh individuals (Fedorova et al. 2013) for additional genetic clustering analysis. Fifth, for genome scans of positive selection in the Ainu, we overlapped the Ainu genotype data with the 1000 genomes project phase 3 data set, downloaded from https://mathgen.stats.ox.ac.uk/impute/impute_v2.html#reference. This includes 540,304 SNPs ("1KG-Ainu" data set). Finally, we overlapped the "1KG-Ainu" data set with available high-coverage Illumina whole genome sequences of contemporary humans and archaic hominins. For this, we retrieved genotype calls of high-coverage Denisovan (Meyer et al. 2012) and Altai Neandertal (Prüfer et al. 2014), in VCF format. Chimpanzee alleles were extracted from the chimpanzee genome assembly Pan_troglodytes-2.1.4 (Pantro4), using LiftOver tool (http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64/liftOver) to convert coordinates between the human reference sequence (GRCh37) and Pantro4. We also obtained short read data of 13 individuals (1-2 individuals from Han, Dai, Sherpa, Yoruba, Karitiana, Sardinian, Papuan and Australian aborigine) from previous studies (Meyer et al. 2012; Jeong et al. 2014; Prüfer et al. 2014). Short reads were aligned to the human reference (GRCh37) using BWA backtrack 0.7.4-r385 (Li and Durbin 2009) with "-q 15" option, duplicate removed with Picard tool v1.98 (http://broadinstitute.github.io/picard/), locally realigned around indels and base quality score recalibrated using the Genome Analysis Toolkit (GATK) v2.8-1 (McKenna et al. 2010; DePristo et al. 2011) following the "best practice workflows" from GATK (Auwera et al. 2013). We kept only properly paired non-duplicate reads with phred-scaled mapping quality ≥ 30 using Samtools v1.2 (Li et al. 2009). For each sample, we called genotypes across all sites using the GATK

UnifiedGenotyper module, based on bases with phred-scaled quality score ≥ 30, and kept sites only if phred-scaled quality score ≥ 50. We combined genotype calls of all modern and archaic individuals using bcftools v1.1 (Li et al. 2009), removed sites if they have any missing genotype, show more than one alternative allele (including 1KG data and chimpanzee alleles), or locate within either human CpG islands (Wu et al. 2010) or human/chimpanzee repeat regions. Finally, we intersected these data with the Ainu genotype data and removed strand ambiguous (A/T or G/C) SNPs, leaving a total of 322,011 SNPs ("CND-1KG-Ainu" data set, CND for Chimpanzee-Neandertal-Denisova).

Assessment of genetic homogeneity within the Ainu individuals

To determine if 25 unrelated Ainu individuals represent a homogenous gene pool, we performed a model-based genetic clustering analysis using *ADMIXTURE* v1.22 (Alexander et al. 2009). For this analysis, we included 25 unrelated Ainu individuals and sets of 30 randomly chosen individuals per each 1KG East Asian population from the "1KG-Ainu" data set. We removed SNPs with minor allele frequency < 0.01 and randomly removed one from each pair of SNPs with $r^2 > 0.2$ ("--indep-pairwise 200 25 0.2" option in PLINK), leaving 84,462 SNPs for the analysis. We ran 50 replicates with random seeds for the number of clusters (K) from 2 to 6 and chose a run with the maximum log likelihood for each K. The optimal value K=2 was chosen based on its lowest five-fold cross validation error. We also performed a principal component analysis (PCA) of 1KG East Asians and the Ainu individuals, as implemented in the smartpca program in the EIGENSOFT package v4.1 (Patterson et al. 2006; Price et al. 2006).

We further estimated admixture time in the Ainu, using a decay of weighted admixture linkage disequilibrium (LD) as implemented in *ALDER* v1.03 (Loh et al. 2013). For this, we

performed a 2-reference *ALDER* analysis, using 1KG JPT (Japanese in Tokyo, Japan) and 12 Ainu with 100% Ainu ancestry in the *ADMIXTURE* analysis ("Ainu" in **Supplementary Figure 4.2A**) as references and the other 10 Ainu as the target ("Ainu2" in **Supplementary Figure 4.2A**). Three individuals who clustered together with mainland Japanese ("Ainu3" in **Supplementary Figure 4.2**) were excluded from the analysis. We also ran *ALDER* with all 22 Ainu individuals as the target population, with SNP loadings for the Ainu-Japanese cline in PCA (PC1 in **Supplementary Figure 4.2B**) as a weight function instead of specifying reference populations, to check if our split of Ainu individuals into two groups generates a bias in estimation. For both analyses, we applied bin size of 0.025 centiMorgan (cM).

Because both analyses above suggested a recent admixture (12.3 and 11.6 generations, respectively) and we are interested in investigating the ancient history of the Ainu (**Supplementary Figure 4.3**), we focused on the 12 individuals with no mainland Japanese ancestry ("Ainu" in **Supplementary Figure 4.2**).

Population clustering and TreeMix analyses

We conducted a genetic clustering analysis of world-wide populations, with a subset of the "WHA" data set including all East Asian and Siberian individuals, using *ADMIXTURE* v1.22. For each data set, we ran 50 replicates with random seeds for the numbers of clusters (K) from 2 to 9 and chose a run with the maximum log likelihood for each K. In all analyses, five to ten best runs for each K had log likelihoods ranging within a difference of 1, supporting a convergence of the best runs. We chose the optimal K for each data set by taking one with the lowest five-fold cross validation error.

Then, we built a consensus tree of 15 world-wide populations representing major ancestry components inferred from the *ADMIXTURE* analysis. For this, we first removed all populations showing a negative three-population (f_3) statistic (Reich et al. 2009; Patterson et al. 2012) to exclude populations that experienced recent admixture. Two additional populations were excluded because they showed significant evidence for admixture using *ALDER*: the Ulchi (using Korean and Itelmen as references; $Z = 4.65$ and $p = 3.3 \times 10^{-6}$) and the Eskimos (using Surui and the Chukchi as references; $Z = 4.19$ and $p = 2.8 \times 10^{-5}$). Second, for populations outside East Asia and Siberia, we arbitrarily chose Ju_hoan_North and Mandenka for Africans, Sardinian and Basque for Europeans, Papuan for Oceania, and Karitiana and Surui for Native Americans. These populations represent major branches of human continental diversity and have been used as such representatives in many population genetic studies (Li et al. 2008; Reich et al. 2011; Keller et al. 2012; Pickrell et al. 2012; Skoglund et al. 2015). Last, we removed the She from the analysis because of its unstable position in the population trees generated by *TreeMix* (Pickrell and Pritchard 2012). The resulting set of 15 populations covers well all ancestry components inferred from the *ADMIXTURE* analysis (**Figure 4.2**). Five hundred bootstrap replicates of the maximum-likelihood tree were generated for the final 15 populations by *TreeMix*, with 50 SNPs per block ("-k 50"). A majority consensus tree was inferred using the R package "ape" (Paradis et al. 2004). We also performed *TreeMix* analysis with 1 to 5 migration edges allowed ("-m 1" to "-m 5") to detect major patterns of extra population affinity not captured by the tree in **Figure 4.3**. One hundred bootstrap replicates were conducted for each m value.

Formal tests of admixture

We calculated three-population (f_3) and Patterson's D statistics (Green et al. 2010) for all combinations of 71 world-wide populations (**Supplementary Table 4.1**), using the *qp3Pop* and *qpDstat* programs in the *ADMIXTOOLS* v1.1 package (Patterson et al. 2012). All contemporary populations from the human genome diversity panel were included. In addition, all the other East Asian and Siberian populations were included if they had sample size ≥ 5. Four Yukagir individuals were removed because they show a large proportion of European-related ancestry, likely due to recent admixture.

We used the admixture LD decay based method implemented in *ALDER* to provide additional evidence for admixture as well as an estimate of time since admixture, assuming a single pulse of admixture. We ran *ALDER* with two reference populations chosen based on the three-population test results. We applied bin size of 0.025 centiMorgan (cM) and required a minimum genetic distance of 0.5 cM between bins.

Frequency of derived alleles with selection signals in East Asians

We interrogated three variants, *EDAR* V370A (rs3827760), *OCA2* H615R (rs1800414) and a non-coding SNP rs3811801 in the *ADH* gene cluster, which carry a selective sweep signal in East Asians. Because rs3827760 and rs3811801 were not included in the Ainu data, we imputed them using *IMPUTE2* (Howie et al. 2009) with 1KG phase 3 data set as a reference. Rs3827760 was imputed with high confidence (genotype posterior probability > 0.98) for all twelve individuals. Except for two individuals who were omitted from further analysis, all imputed genotypes at rs3811801 had posterior probability ≥ 0.89. Therefore, although imputation

in an isolated population may have reduced accuracy, genotypes were imputed with high confidence in our samples.

Genome scans of recent positive selection in the Ainu

We used autosomal SNPs from the "1KG-Ainu" data set to detect genomic regions showing signatures of recent positive selection in the Ainu. For this, we calculated the cross population extended haplotype homozygosity (XP-EHH) statistic (Sabeti et al. 2007) against 1KG phase 3 CHB (Han Chinese in Beijing, China) and the population branch statistic (PBS) (Yi et al. 2010) using CHB as a comparison group and 1KG phase 3 CEU (CEPH Utah residents with northern and western European ancestry) as an outgroup. We removed strand ambiguous (G/C and A/T) SNPs from the analysis. To perform the XP-EHH analysis, we first phased the Ainu genotype data using *SHAPEIT2* v2.r790 (Delaneau et al. 2013) with 1KG phase 3 data as a reference. The most likely haplotypes were chosen after running *SHAPEIT2* with default parameter values. We summarized signals for each of 500 kb windows sliding by 50 kb. First, we counted the number of total SNPs (n_{total}) and top 1% signal SNPs (n_{top}) for each window and each test. Second, we calculated a simple binomial probability $P(X \geq n_{top})$ assuming a binomial distribution with success probability 1%, i.e. $X \sim B(n_{total}, 0.01)$. Probability of 1 was assigned to windows with $n_{total} < 20$. Then, we prioritized windows with binomial $p \leq 0.01$ for both XP-EHH and PBS and merged adjacent windows. Finally, we narrowed down the peaks by removing 50 kb windows that do not harbor any top 1% SNP, resulting in 66 signal peaks for further analysis.

We used a simulation-based approach to test if our selection scans have enough power to distinguish loci under positive selection from neutrally evolving ones, given the small effective population size of the Ainu and our small sample size. For this purpose, we focused on the top 10

regions among the above 66, which have the highest PBS statistics. In our procedure, we first simulated neutral trajectories of derived alleles under the Wright-Fisher model with a constant size of N_e=2,219, which we estimated from LD decay. Specifically, we fit a non-linear regression model with the equation $E(r^2) = 1/(1+4N_ec)+1/n$, where c is a genetic distance in Morgans and n is the number of sampled chromosomes (Tenesa et al. 2007). We also repeated our simulations with a more conservative estimate N_e=1,000. Although it is hard to model accurately demographic history using array genotyping data, the range of N_e values we used is likely to span a range of relevant demographic scenarios. The time of divergence between Ainu and the other East Asians was assumed to be 800 or 1,000 generations ago and the frequency in Ainu at the time of the split was taken from the current frequency in the other East Asians. At the end of each simulation, we took the frequency ($f_{present}$) and sampled 24 alleles following a binomial distribution of probability $f_{present}$. We repeated this process 10,000 times and calculated an empirical probability of our data by taking the fraction of simulations where the counts of simulated derived alleles are greater than those observed in the Ainu sample.

4.4: Results

A subset of the Ainu samples are the result of recent admixture

We first investigated if the Ainu samples in our study represent a homogenous gene pool. Both a principal component analysis (PCA) and a genetic clustering analysis showed that the Ainu samples are genetically heterogeneous and form a few distinct clusters (**Supplementary Figure 4.2**). Based on these analyses, three individuals labeled as Ainu were indistinguishable

from the mainland Japanese samples (**Supplementary Figure 4.2**); therefore, they were removed from the analysis. The same analyses also identified 10 Ainu individuals with substantial non-Ainu ancestry (> 11%). Using *ALDER*, which is based on weighted admixture LD decay, we estimated the admixture time for these individuals to be 12.3 generations ago (**Supplementary Figure 4.3**), further supporting a recent mixture ($p = 1.1 \times 10^{-6}$). If we included the entire sample of unrelated Ainu, a similar analysis with SNP loading on PC1 as a weight vector, without specifying reference populations, we obtained a similarly recent estimate (11.6 generations ago). Furthermore, to explore more complex admixture scenarios, we also fit data to a two-pulse admixture model, resulting in a combination of a younger (5.2 ± 2.1 generations ago) and an older (40.7 ± 8.9 generations ago) admixture event (**Supplementary Figure 4.3**). When using a more stringent cutoff of 2.0 cM for the minimum genetic distance between markers, estimates were younger: 8.8 ± 2.1 for a single-pulse model, and 4.6 ± 2.0 and 30.4 ± 10.1 for a two-pulse model. Even the older estimates of 30-40 generations ago from the two-pulse model, which set an upper bound of a continued gene flow, are too recent to be consistent with the initial expansion of the Yayoi culture into the Japanese archipelago (Jinam et al. 2012; Nakagome et al. 2015) or with the hypothesized gene flow from the so called "Okhotsk culture", which spread throughout the Sakhalin and Hokkaido islands between 5th and 11th centuries (Befu and Chard 1964; Ohyi 1975). Because we are primarily interested in the ancient history of the Ainu gene pool and its relationship with world-wide populations, we decided to primarily use 12 individuals with 100% Ainu ancestry ("Ainu" in **Supplementary Figure 4.2**) as representative of the original Ainu gene pool in the following analyses. However, we also performed several analyses with all 22 Ainu individuals in parallel and obtained comparable results, as presented below.

The Ainu form an outgroup to all East Asian farmers including Tibetan populations

To infer the genetic relationship of the Ainu gene pool with world-wide populations, we compiled genotype data of the Ainu, the Sherpa, Tibetans and 183 populations around the world as described in the Materials and Methods section ("WHA" data set). With this data set, we first characterized the Ainu ancestry in the context of East Asian genetic diversity, by performing a model-based unsupervised genetic clustering as implemented in the program *ADMIXTURE*. With the optimal number of ancestral components (K = 8), the Ainu individuals are assigned to a distinct ancestry (**Figure 4.2**). In suboptimal runs with fewer ancestral components ($K \leq 6$), the Ainu, as most other East Asians, are modeled as a mixture of Siberian and East Asian ancestries, with limited contributions from other populations (**Supplementary Figure 4.4**). Interestingly, Ainu-related ancestry is present in the Japanese and the Ulchi people from the lower basin of Amur river (17.8% and 13.5% mean ancestry in **Figure 4.2**, respectively), as well as in two Nivkh individuals, an indigenous population from the Sakhalin island, in a similar analysis with additional samples (27.2% mean ancestry; **Supplementary Figure 4.5**). This suggests a potential gene flow from an Ainu-related gene pool into these surrounding populations.

Then, we further explored the genetic relationships among the ancestry components inferred from the *ADMIXTURE* analysis. To do this, we aimed to choose population samples that were good representatives of the global population structure (**Figure 4.2**). First, we removed populations with signatures of recent admixture not involving the Ainu ancestry, suggested either by negative values of the three-population (f_3) statistic or by significant decay of weighted admixture LD (see Materials and Methods). To simplify the analysis, we arbitrarily chose one or two populations to represent each of the major continental groups outside of East Asia, but kept all East Asian and Siberian populations with no signal of recent admixture. This process resulted

in a total of fifteen populations (**Figure 4.3**), which cover all ancestry components identified by *ADMIXTURE*.

Using a maximum likelihood-based algorithm implemented in *TreeMix*, we found that the Ainu can be modeled as an outgroup to all East Asian farmers (Ami, Atayal, Dai, Lahu and the Sherpa; **Figure 4.3**) in all 500 bootstrap replicates. The long terminal branch leading to the Ainu (**Figure 4.3**), as well as slow decay of LD (**Supplementary Figure 4.6**), suggests a strong genetic drift, expected for a small population isolate. The Ainu's position as an East Asian outgroup in the tree is unlikely due to gene flow from outside East Asia, as no significant results were found with Patterson's D statistic in the form of D(African, non-African outgroup; Ainu, East Asian farmer) (**Supplementary Table 4.2**). We further investigated this point by using an additional data set of 320K SNPs in a smaller set of populations including also the Neandertal and Denisova data ("CND-1KG-Ainu" data set). Consistent with the *TreeMix* analysis, the Ainu form a clade with East Asian populations relative to Europeans or South Asians (**Supplementary Table 4.3 and Supplementary Figure 4.7**). Additionally, at the resolution provided by this data set, the Ainu are not inferred to contain more archaic ancestry than other East Asian populations. The *TreeMix* result is unlikely to be an artifact of either genetic drift or variant ascertainment bias: this algorithm was shown to work well even if population-specific variants are common (extreme genetic drift) or if all variants are ascertained in a single population (Pickrell and Pritchard 2012). In addition, the allele frequency distribution of the SNPs included in the analysis was similar across the Ainu and the other East Asian, Siberian and Native American populations (**Supplementary Figure 4.8**) and it did not vary substantially between different intersected sets of SNPs (**Supplementary Figure 4.9**). The Sherpa formed an outgroup to the lowland East Asian farmers, consistent with our previous study showing a deep

split between high- and low-altitude East Asians (Jeong et al. 2014). Siberian populations (Nganasan and Itelmen) were modeled either as a sister group of all East Asians including the Ainu (76.8%; **Figure 4.3**) or as a sister group of Native Americans, Karitiana and Surui (23.2%; **Supplementary Figure 4.10**). When we allowed for migration edges in the *TreeMix* analysis, gene flow events between Europeans and Native Americans or Siberians were robustly inferred in all bootstrap replicates (**Supplementary Figure 4.11 and Supplementary Table 4.4**). This pattern is in agreement with the findings of previous studies on the genetic history of Native Americans and Siberian populations (Rasmussen et al. 2010; Fedorova et al. 2013; Raghavan et al. 2014).

Figure 4.2: *ADMIXTURE* analysis of East Asian and Siberian populations with K = 8. Ainu individuals are assigned to a distinct ancestry component (dark grey), which is present also in Japanese and Ulchi individuals. Siberian ancestry is divided into two major components, one for northeast Siberians (skyblue in Chukchi, Eskimo, Itelmen and Koryak) and the other for central Siberians (orange in Nganasan and other Siberian and northeast Asian populations). Four Yukagir individuals harboring a large proportion of European ancestry were labeled as "Yukagir2".

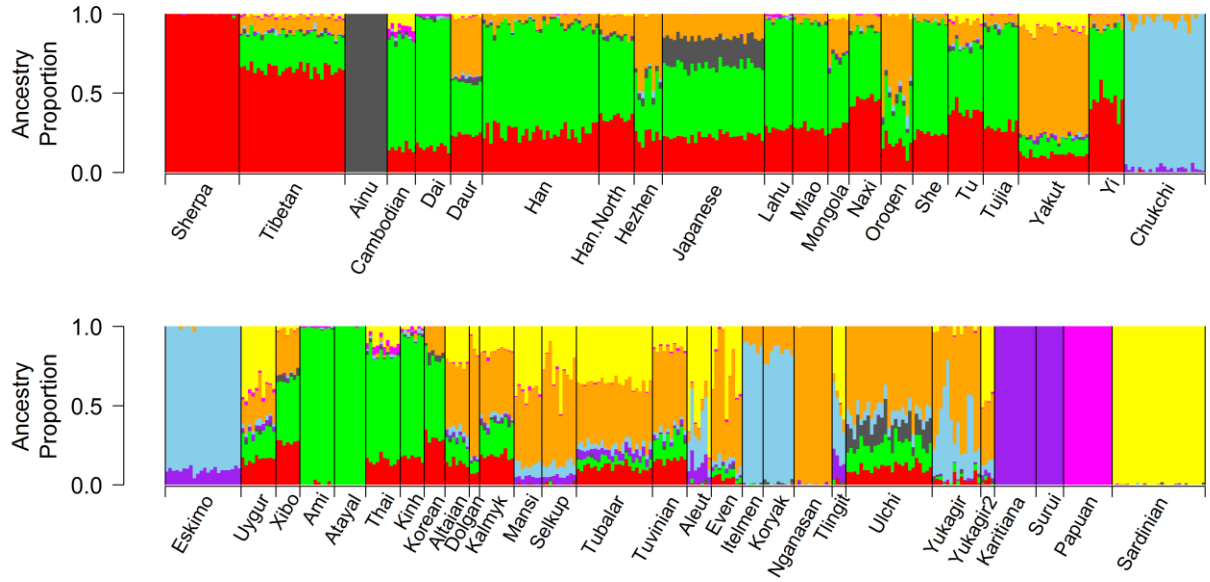
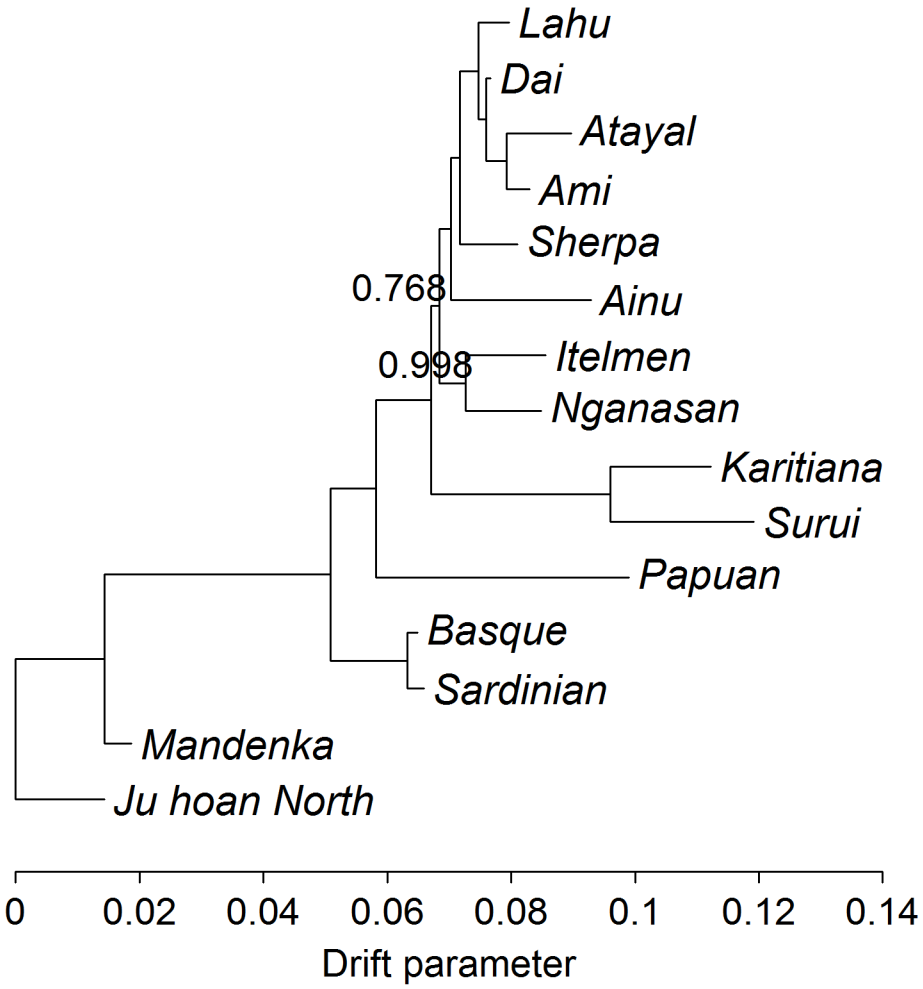


Figure 4.3: A consensus tree of 15 world-wide populations inferred from 500 bootstrap replicates of maximum likelihood trees using *TreeMix*. Numbers show the bootstrap support on the corresponding nodes. Nodes with no number were supported in 100% of the bootstrap replicates.



The Ainu share more ancestry with low-altitude than with high-altitude East Asians

Human populations often have a complicated genetic history, which cannot be fully captured by a simple tree-based model. The distribution of residual covariances from the maximum likelihood trees indeed suggests that the consensus tree cannot fully explain the data for many of the populations including the Ainu (**Supplementary Figure 4.12**). Therefore, as a next step, we investigated if the Ainu have extra affinity with other populations beyond what could be inferred in a bifurcating tree. First, we tested if East Asian farmer populations are symmetric to each other in terms of their relationship with the Ainu, as suggested by the Ainu position as an outgroup to these populations in the consensus population tree (**Figure 4.3 and Supplementary Figure 4.13A**). If the population relationships in the consensus tree hold, two East Asian farmer populations should be equally close to the Ainu. However, the D statistics in the form of D(Outgroup, Ainu; Sherpa, other East Asian) consistently showed significantly positive values (D > +2.9 SD; **Supplementary Table 4.5**), suggesting gene flow between the Ainu and lowland East Asian populations after their split from the high-altitude populations. Unsurprisingly, the inclusion of the 10 recently admixed Ainu individuals further strengthened this pattern (D = +3.0 to +8.6 SD for the same set of tests). Genetic affinity tests of East Asian populations with the Ainu, assessed by the outgroup f_3 statistic (Raghavan et al. 2014), also supported a closer relationship between the Ainu and lowland East Asians than between the Ainu and the Sherpa (**Supplementary Figure 4.14**). *TreeMix* analysis allowing migration edges also detected a similar relationship: an edge between the Ainu and lowland East Asians was inferred for 59-88% of bootstrap replicates, when 3 or more migration edges were allowed (**Supplementary Figure 4.11 and Supplementary Table 4.4**).

The Ainu share ancestry with northeast Siberians but not with central Siberians

Previous genetic studies of Siberian populations clearly demonstrated genetic differentiation between northeast Siberians and the rest of the Siberian populations (Volodko et al. 2008; Rasmussen et al. 2010; Fedorova et al. 2013; Raghavan et al. 2014). Our genetic clustering analysis recapitulates this observation, by separating the northeast Siberian ancestry (skyblue in **Figure 4.2**; concentrated in Eskimo, Chukchi, Itelmen and Koryak) from central Siberian ancestry (orange in **Figure 4.2**; most prevalent in the Nganasan and present in southern Siberians and northeast Asians). Even though the Itelmen and the Nganasan cluster together in our population tree (**Figure 4.3 and Supplementary Figures 4.10 and 4.13B**), most East Asian populations are genetically closer to the Nganasan than to the Itelmen, as shown by their negative D(African, East Asian; Nganasan, Itelmen) statistic (-2.4 to -12.0 SD; **Figure 4.4 and Supplementary Figure 4.15**). Interestingly, the Ainu were the only East Asian population showing a closer affinity to the Itelmen than to the Nganasan, although the observed negative D statistic was within statistical noise (+1.5 SD; **Figure 4.4**). This pattern was robust to the inclusion of the 10 recently admixed Ainu individuals, which – as expected – dampens the signal due to the presence of mainland Japanese ancestry (+1.0 SD; **Supplementary Figure 4.16**). Consistent with this result, a D test in the form D(African, Siberian, Ainu, East Asian) showed that the Itelmen are genetically closer to the Ainu than to East Asian farmer populations, but the Nganasan are symmetric in their relationship to the Ainu and the East Asian farmers (**Supplementary Table 4.6**). Northeast Asian or southern Siberian populations could not be directly compared in this way because of the shared ancestry between the Itelmen and the Nganasan. We obtained qualitatively similar results when we replaced the Itelmen with the Chukchi (**Supplementary Figure 4.17 and Supplementary Table 4.6**). The *TreeMix* analysis

also detected migration edges between the Itelmen and the Ainu when 4 or 5 migration edges were allowed, but only in 13-23% of bootstrap replicates (**Supplementary Table 4.4**).

An Ainu-related ancestry was introduced into nearby populations

Our genetic clustering analysis strongly suggests that the Ainu-related ancestry substantially contributed to the gene pools of nearby populations, such as the Japanese or the Ulchi (**Figure 4.2 and Supplementary Figures 4.4 and 4.5**). However, strong genetic drift in the Ainu may artificially generate such a signal. Therefore, we applied two formal tests of admixture, which use different aspects of genetic variation data, to test for Ainu-related admixture in these populations. First, three-population test statistics were significantly negative when the Ainu were used as a reference: -22.2 SD for the Japanese (using the Ainu and Han as references) and -3.9 SD for the Ulchi (using the Ainu and Nganasan as references). Second, admixture LD decay was clear in both populations ($p = 3.7 \times 10^{-26}$ and $p = 8.7 \times 10^{-9}$ for the Japanese and the Ulchi, respectively), with estimates of admixture time around 70 and 22 generations ago for the Japanese and the Ulchi, respectively (**Supplementary Figures 4.18 and 4.19**).

Figure 4.4: The genetic affinity of East Asian and Siberian populations to the Nganasan and the Itelmen, respectively, measured by Patterson's D(Yoruba, X; Nganasan, Itelmen). Horizontal bars around the value represent ± 3 standard deviations.

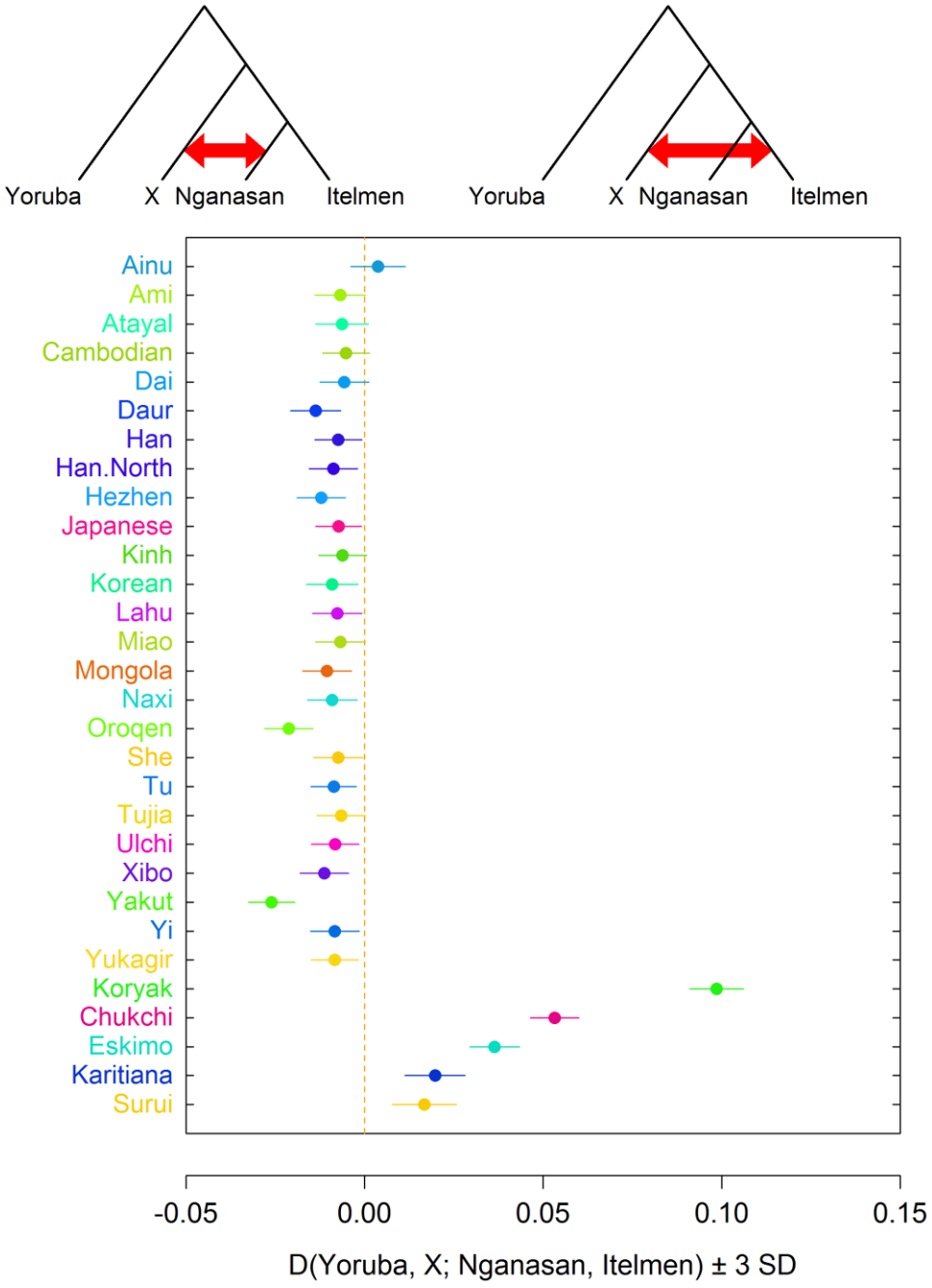
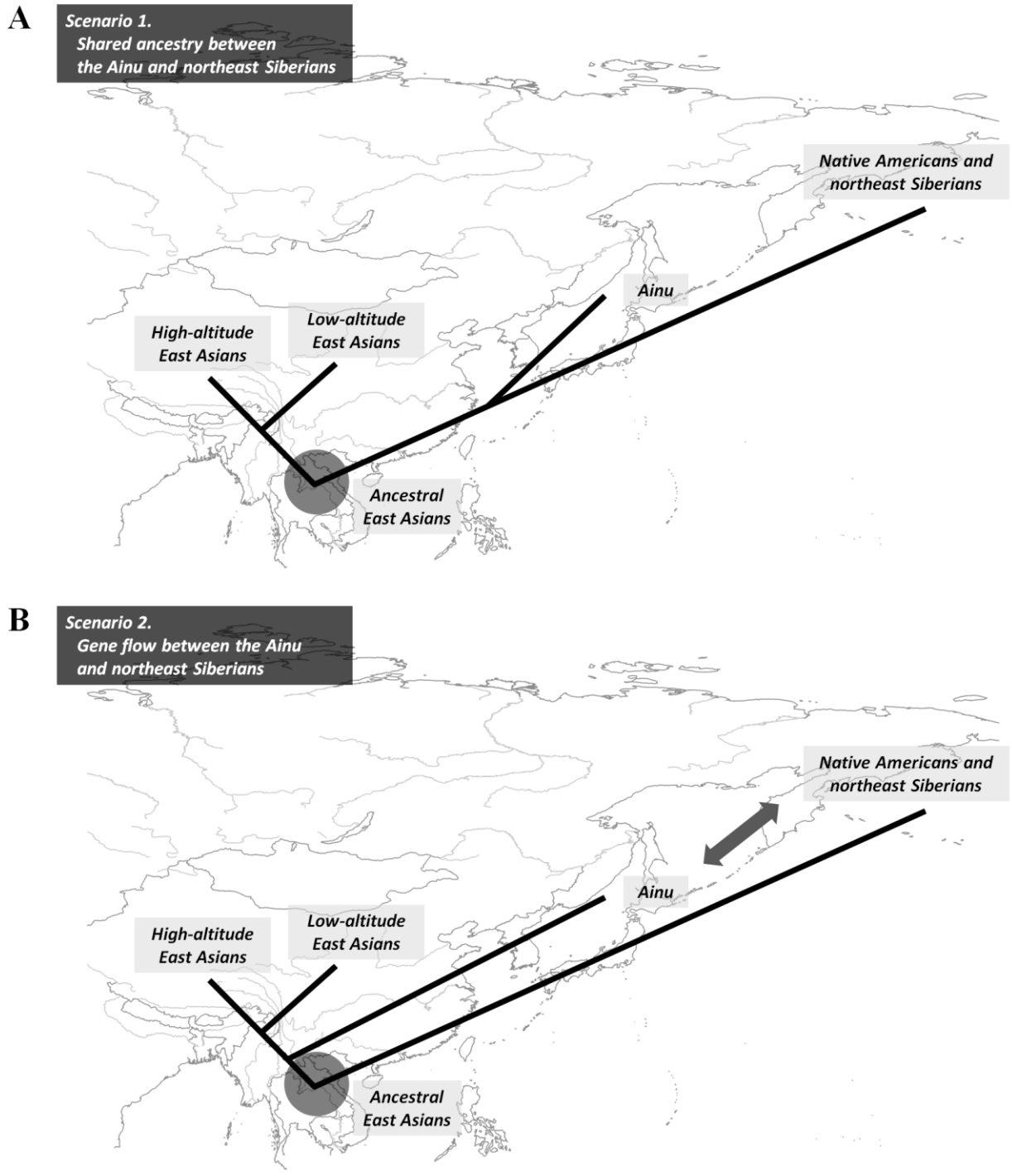


Figure 4.5: A summary of competing scenarios for the observed excess affinity of the Ainu with northeast Siberians. (A) "Scenario 1" proposes a shared ancestry between the Ainu and northeast Siberians. (B) "Scenario 2" proposes a later gene flow between the Ainu and northeast Siberians.



The Ainu genome harbors shared and unique signatures of adaptations

To investigate adaptive evolution occurring in the Ainu, we analyzed the allele frequencies of variants known to be swept to high frequency in East Asians and performed genomic scans of recent positive selection across the Ainu genome.

A nonsynonymous V370A (rs3827760) mutation in the *EDAR* (ectodysplasin A receptor) gene harbors a strong selective sweep signal shared among low- and high-altitude East Asians and Native Americans, which is not present in contemporary western Eurasian populations (Kamberov et al. 2013). In addition, the derived allele is associated with "East Asian phenotypes", such as shovel shaped incisors (Kimura et al. 2009). In sharp contrast to surrounding populations, this allele occurs at only 25% (6 out of 24) frequency in the Ainu (**Supplementary Table 4.7**). Consistent with this finding, the Ainu are also reported to have the sundadont dental pattern, even though the sinodont pattern, which is associated with shovel shaped incisors, is the dominant one in northeast Asia (Howells 1997). This suggests that the Ainu ancestors may not have shared the selective pressures for *EDAR* V370A with other ancestral East Asian and Native American populations.

In contrast to *EDAR* V370A, two variants occurring at high frequency in East Asia and virtually absent elsewhere, rs1800414 (H615R) in the *OCA2* gene (Hider et al. 2013) and rs3811801 in the *ADH* gene cluster (Li et al. 2011), have high frequency in the Ainu (**Supplementary Table 4.7**). The onset of positive selection on these two variants was estimated to have occurred less than 11,000 years ago (Peng et al. 2010; Li et al. 2011; Chen, Hey, et al. 2015). Therefore, the Ainu ancestors may have shared Holocene environmental factors favoring these variants with other East Asians, although gene flow between the Ainu and other East Asians, as we inferred from genome-wide SNP data, may also have contributed to their high

frequencies in the Ainu. Interestingly, they occur at low frequency around 10% in the Sherpa and Tibetans, raising the possibility that selective pressure on these variants were different in the high-altitude environments.

To find genomic loci involved in local adaptations in the Ainu, we performed a LD-based test of recent positive selection (XP-EHH), and an allele frequency-based test (PBS), in the Ainu against CHB from the 1KG phase 3 data set. We found a total of 66 genomic regions containing excess SNPs with extreme values (top 1%), defined by a binomial probability ≤ 0.01 of having the number of extreme SNPs equal to or greater than the observed one, of both XP-EHH and PBS statistics (**Supplementary Table 4.8**). One such region on chromosome 11 spans the Apolipoprotein (*APO*) gene cluster, including the *APOA5*, *APOA4*, *APOC3* and *APOA1* genes. This region contains a noncoding SNP rs964184 associated with levels of blood triglycerides and LDL cholesterol, HDL cholesterol, and risk of coronary artery disease and ischemic stroke (Global Lipids Genetics Consortium 2013; Dichgans et al. 2014). Interestingly, the risk allele has much higher frequency in the Ainu, reaching 75% in comparison to 22% in 1KG CHB, and is found on a haplotype with extended LD (**Supplementary Figure 4.20**).

In addition, multiple loci among the above 66 regions include SNPs showing extreme allele frequency differentiation between the Ainu and 1KG East Asians excluding JPT (**Supplementary Table 4.9**). Several genes in these loci have been reported to be associated with other metabolic traits, such as body mass index, glomerular filtration rate and serum metabolite levels (**Supplementary Table 4.9**). Because the Ainu sample size is small (n=12) and the Ainu have had historically small effective population size (**Supplementary Figure 4.6**), high allele frequency divergence may not necessarily be due to local adaptations. Therefore, we relied on simulations to test if the observed allele frequency difference is greater than expected by chance

for a small population and a sample size as small as ours. We found that for all of the 10 top PBS regions the difference in allele frequency is unlikely to be due to chance (empirical p-value ≤ 0.01), based on neutral simulations of population of constant effective population size (N_e) of 2,219, as estimated using an LD decay method (**Supplementary Table 4.9**). We obtained similar results (9 out of 10 top PBS regions with empirical p-value ≤ 0.05) (**Supplementary Table 4.9**), when we used a smaller, more conservative estimate of N_e=1,000.

4.5: Discussion

Our genome-wide analysis shows that the Ainu are one of the deepest branches of East Asian diversity, forming an outgroup to all present-day East Asian farmers, including high-altitude populations (**Figure 4.3**). The deep history of the Ainu is consistent with the archaeological record for the Jomon culture in Japan starting 16,500 years ago (Habu 2004) as well as their hunter-gatherer life style. Therefore, the ancestors of the Ainu are likely to have reached the Japanese archipelago in an early migration event distinct from the spread of farmer populations across East Asia.

Interestingly, we find evidence for extra genetic affinity between the Ainu and northeast Siberians (Itelmen and Chukchi), who share ancestry with Native Americans. This finding coupled with the ancient origin of the Ainu raises the possibility that the same migration event led to the settlement of Jomon hunter-gatherers and to the initial dispersal of Native American ancestors. If this is the case, this first northward migration took place before the LGM ("Scenario 1" in **Figure 4.5A**). This proposal is consistent with previous studies that suggested a connection

between Jomon or Ainu people and Native Americans based on morphological and genetic evidence (Tokunaga et al. 2001; Adachi et al. 2009; Owsley and Jantz 2014) (**Figure 4.4 and Supplementary Table 4.6**). Under this scenario, the split between the Ainu and Native American ancestors is likely to have occurred earlier than the gene flow of western Eurasian ancestry into the Native American ancestors (Raghavan et al. 2014), because the Ainu and other East Asians are symmetrically related to contemporary Europeans and to the ancient MA1 sample (**Supplementary Table 4.2**). However, a recent study reported no genetic affinity between the Ainu and the Kennewick man (Rasmussen et al. 2015). An alternative to this scenario is that there may have been more recent gene flow between the Ainu and northeast Siberian populations ("Scenario 2" in **Figure 4.5B**). Our *TreeMix* analysis with migration edges suggests a gene flow from the northeast Siberians to the Ainu, although both the pattern itself and its direction are not robust (**Supplementary Table 4.4**). As explained above, the spread of the "Okhotsk culture" is unlikely to account for this finding, although such a contact across the Okhotsk sea may have happened earlier than the Okhotsk culture. Whole genome sequence data of the Ainu, ancient Jomon samples and northeast Siberians will shed more light on the details of this history (Li and Durbin 2011; Schiffels and Durbin 2014).

Surprisingly, we also find extra genetic affinity between the Ainu and lowland farmer populations in comparison to the Sherpa (**Supplementary Figure 4.14 and Supplementary Table 4.5**), indicating gene flow between these two groups of populations. A long-standing hypothesis posits that Ainu and Tibetans share a part of their ancestry that is not present in other East Asian populations based on patterns of Y chromosome variation. The Y chromosome haplogroup D-M174 shows a striking pattern of geographic distribution: it is highly prevalent in Tibetans and Japanese (especially in the Ainu) and virtually absent everywhere else in Eurasia

(Hammer et al. 2006; Shi et al. 2008; Chiaroni et al. 2009; Stoneking and Delfin 2010). A possible explanation for the Y chromosome data is that the Tibetan and the Japanese branches of this haplogroup have deep coalescence times, i.e. older than 30,000 years before present (Shi et al. 2008). Therefore, even if the presence of the D-M174 haplogroup in the Ainu and Tibetans is due to shared ancestry, the shared history of these populations was short and left only a weak genome-wide signature of shared variation in their gene pools.

Even though we find strong evidence of gene flow between the Ainu and lowland East Asian farmers, it is hard to establish whether migrations were mainly unidirectional and, if so, which direction was predominant. One possibility is that Ainu-related populations, probably hunter-gatherers, once occupied mainland East Asia preceding the expansion of farmers and that they contributed to the gene pool of the latter. The observation of Ainu-related ancestry in the Ulchi from the lower Amur river basin (**Figure 4.2 and Supplementary Figures 4.4 and 4.5**) is consistent with the presence of such an Ainu-related population in mainland northeast Asia. This model of gene flow is not expected to generate a signature of admixture in the Ainu. Consistent with this scenario, we fail to detect an admixture signal in the Ainu beyond the 10 recently admixed individuals (**Supplementary Figures 4.2 and 4.3**): three-population statistics are strongly positive for all combinations of reference populations listed in Table S1 (> +25 SD). Extended LD in the Ainu and the lack of a reference population representing the Jomon ancestry made it difficult to test for admixture in the Ainu using *ALDER*: indeed, decay constant and amplitude parameter estimates from one-reference *ALDER* analysis did not change regardless of our choice of reference population (decay constant = 9.0 to 9.8 generations ago, amplitude = 1.6 to 2.3×10^{-4} across all 26 1KG populations as a reference). This pattern was reported to be a likely false positive in the *ALDER* analysis, which occurs when the target population experienced

strong genetic drift (Loh et al. 2013). Therefore, we probably did not have enough power to accurately infer the timing and direction of ancient gene flow events, such as those we found between the Ainu and lowland East Asians or northeast Siberians. Genetic analysis of ancient Jomon and Ainu samples over a range of time periods will be critical to distinguish among these hypotheses.

While we confirm the evidence first reported by Jinam *et al* (2012) for an admixture event between Yayoi farmers and Ainu ancestors, we do not find evidence supporting a claim for a gene flow into the Ainu from an unknown population. This claim was based on a group of five Ainu individuals clustering away from the rest in PCA plots (Jinam et al. 2012). We think this is an artifact of including close relatives in PCA, a well known phenomenon. Indeed, a recent re-analysis by Jinam *et al* (2015) also found that exclusion of close relatives from the analysis removed this clustering pattern (Jinam et al. 2015). We discuss this issue in more detail in Supporting Information (**Supplementary Text 4.1**).

We took a cautious approach in performing and interpreting genome scans of recent positive selection in the Ainu, because of their small effective population size and our small sample size. For example, we decided not to use the integrated haplotype score approach (Voight et al. 2006) because it requires substantial numbers of haplotypes harboring both ancestral and derived alleles at the focal variant. Considering this limitation, it is particularly encouraging that we find several loci harboring extreme allele frequency differentiation in the Ainu (greater than expected under neutrality based on simulations) in comparison to 1KG CHB data (**Supplementary Table 4.9**). In particular, lipid metabolism may have been a key process for local adaptation in the Ainu, as suggested by the selection signature around the *APO* gene cluster and their historical heavy dependence on a marine subsistence. Archaeological evidence for

heavy reliance of the prehistoric Jomon culture, particularly in the northeastern part of Japan, on marine mammals and fish (Chisholm and Koike 1999; Yoneda et al. 2002) may provide a plausible link between our findings and local adaptations of the Ainu and Jomon people.

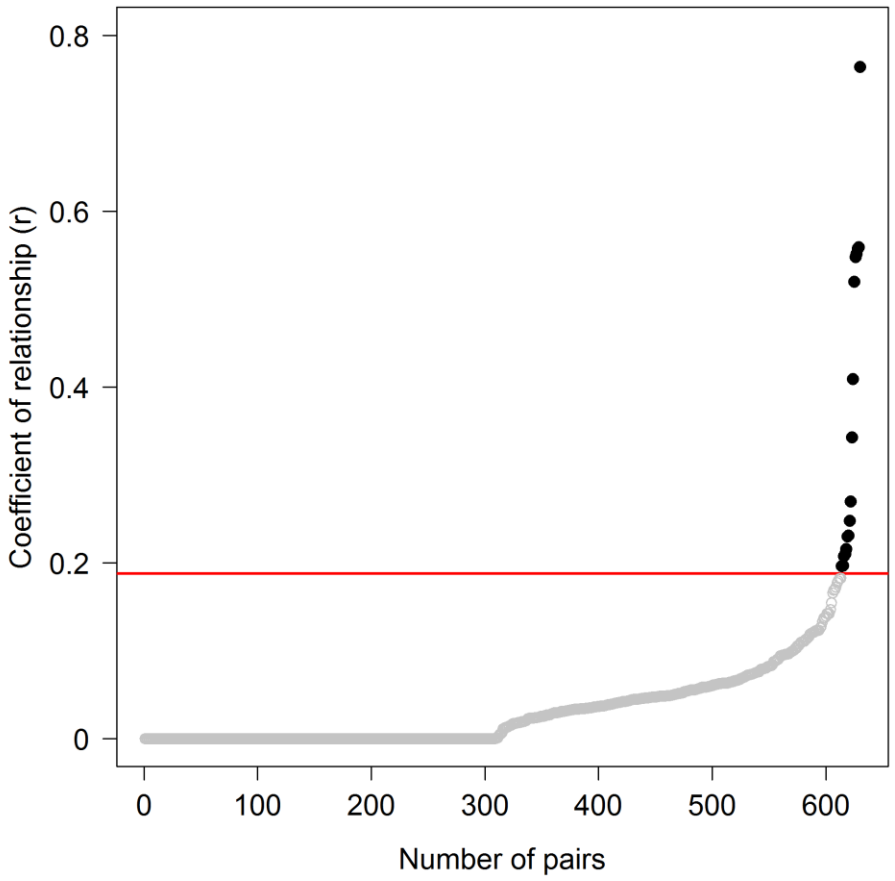
4.6: Appendix: Supplementary Materials

Supplementary Text 4.1: Inclusion of close relatives in PCA generates artificial clusters in the Ainu

In a previous study of the same Ainu genotype data set, it was claimed that the Ainu received gene flow from mainland Japanese and from an unknown population (Jinam et al. 2012). These gene flow events were suggested based on their principal component analysis (PCA) using HapMap East Asians, mainland Japanese, Ryukyuans and all 36 Ainu individuals. More specifically, a scattered distribution of the Ainu individuals across their first PC, which separates the Ainu from other East Asians, was interpreted as evidence for admixture with mainland Japanese. Likewise, a cluster of five Ainu individuals apart from the others along their second PC was interpreted as evidence for another admixture event with the unknown second source population. Analyzing the same Ainu data, we clearly detected a recent gene flow from mainland Japanese into the Ainu (**Supplementary Figures 4.2 and 4.3**), confirming Jinam *et al*'s first finding. However, we did not find supporting evidence for the second admixture event from an unknown source in our PCA results, in which we used 25 unrelated Ainu individuals (**Supplementary Figure 4.2B**). Our additional analysis strongly suggests that this inconsistency

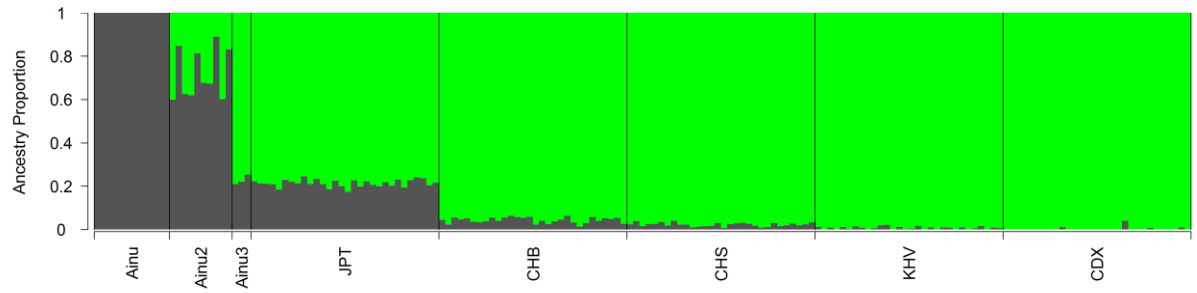
is easily explained by the fact that we omitted 11 closely related individuals from the analysis **(Supplementary Figure 4.1**). It is well known that relatedness may affect PCA results whereby related individuals tend to cluster closely in PCA plots. Here, we provide PCA plots including related individuals (**Supplementary Figures 4.21A and B**), which generate clusters similar to those in Jinam *et al* but not present in PCA plots without related individuals (**Supplementary Figures 4.21C and D**). Specifically, **Supplementary Figure 4.21B** recapitulates Figure 1a of Jinam *et al*, which reported a cluster of five Ainu individuals along their second PC (upper left side in the plot). We found that they are highly related to each other. So, we think this difference in the data is a major contributor to the apparent discrepancy. For running PCA, we used 504 1KG phase 3 East Asians and the Ainu individuals, either all 36 (**Supplementary Figures 4.21A and B**) or 25 unrelated ones (**Supplementary Figures 4.21C and D**). For **Supplementary Figures 4.21C and D**, 11 related Ainu individuals were projected onto PC planes.

Supplementary Figure 4.1: Cumulative distribution of coefficient of relationship (r) between pairs of Ainu individuals. 11 individuals from 17 pairs with r > 0.1875 (black filled dots above the red line for r = 0.1875) were removed from the downstream population genetic analysis.

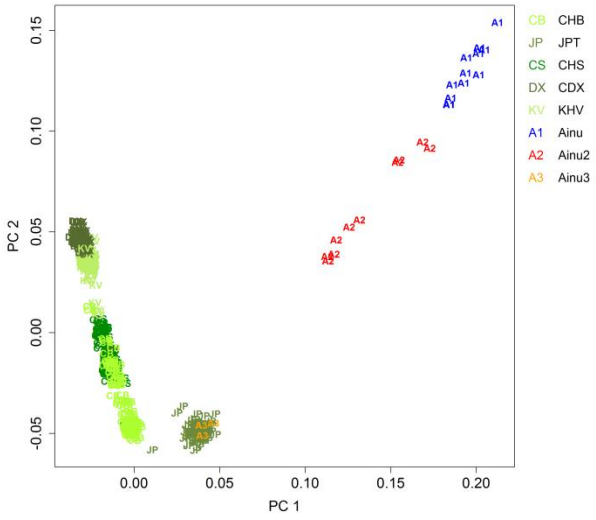


Supplementary Figure 4.2: Identification of Ainu individuals with recent mainland Japanese ancestors. (A) *ADMIXTURE* analysis with K=2 and (B) PCA of 1KG East Asian and Ainu individuals identified 12 Ainu individuals with the highest Ainu ancestry (dark grey in A and PC 1 in B). Another 10 Ainu individuals ("Ainu2") formed two discrete clusters between the "Ainu" cluster and mainland Japanese (JPT). Three Ainu individuals clustered closely with mainland Japanese ("Ainu3"). (C) *ADMIXTURE* with K=2 and PCA results closely match, detecting the same four clusters, including those with 100% Ainu ancestry.

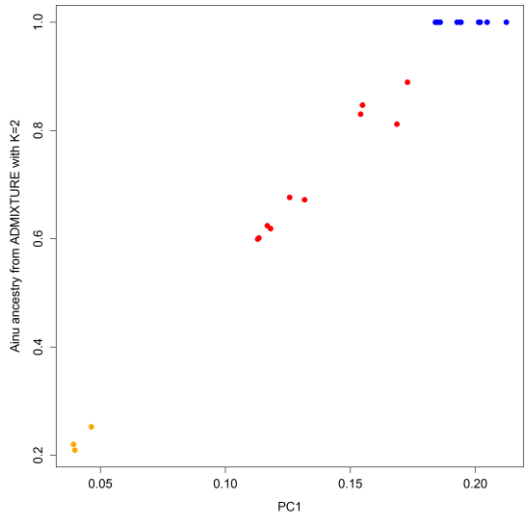
A



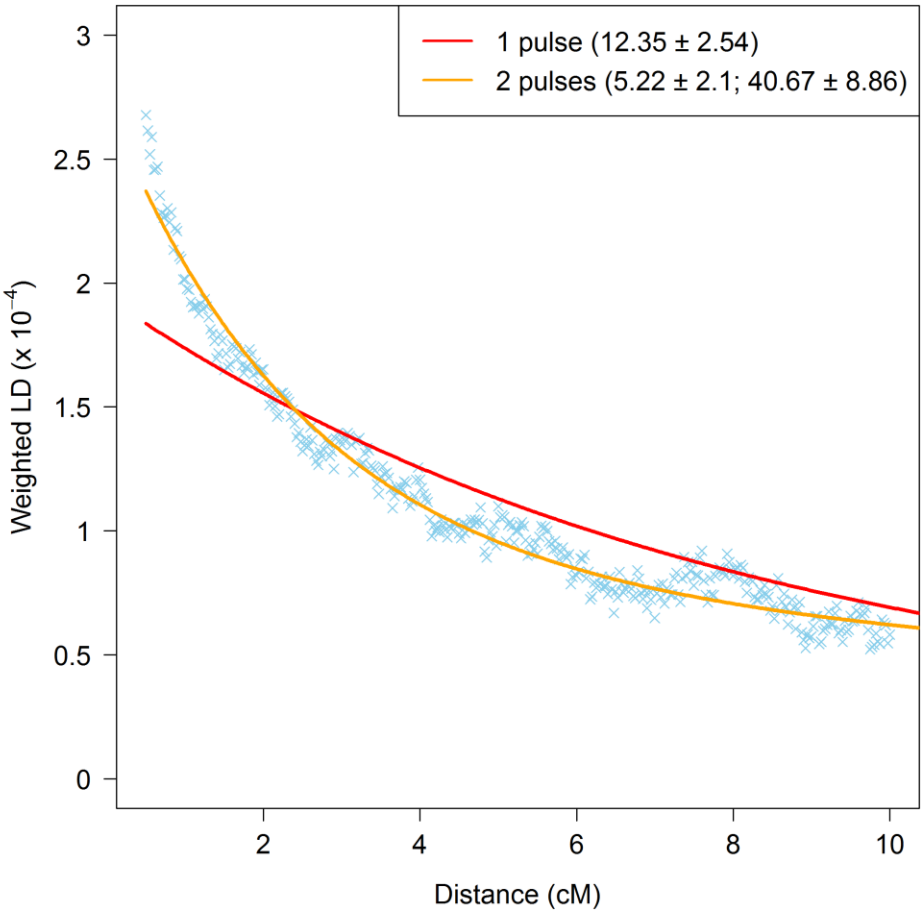
B



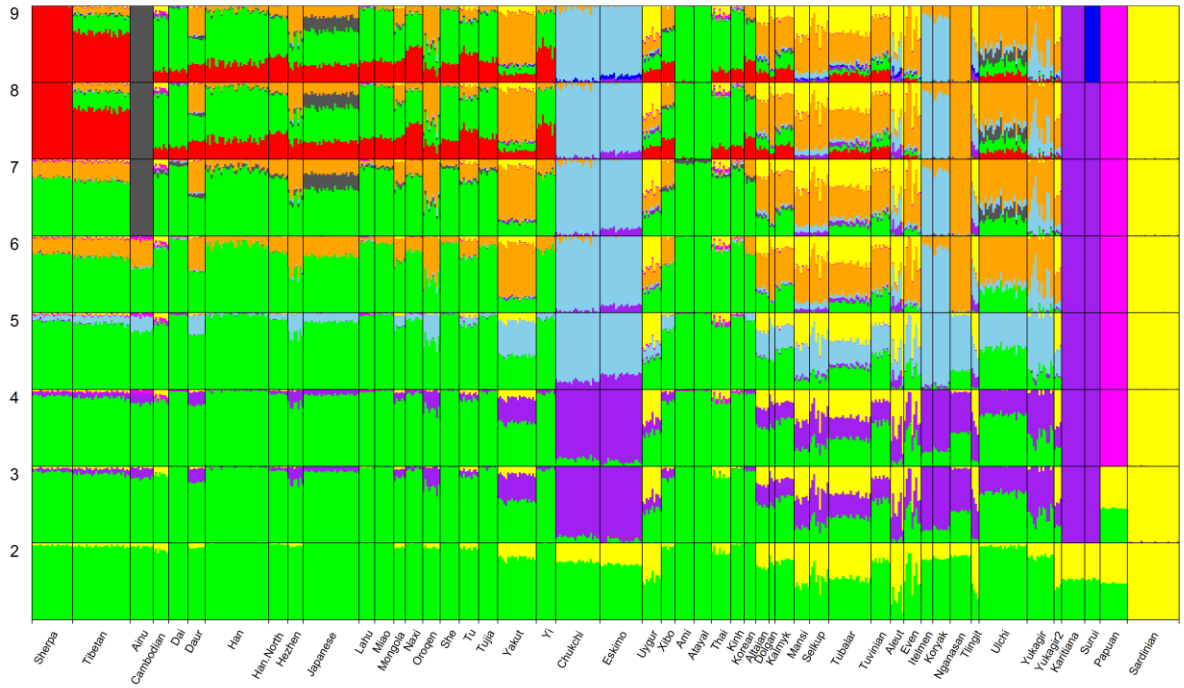
C



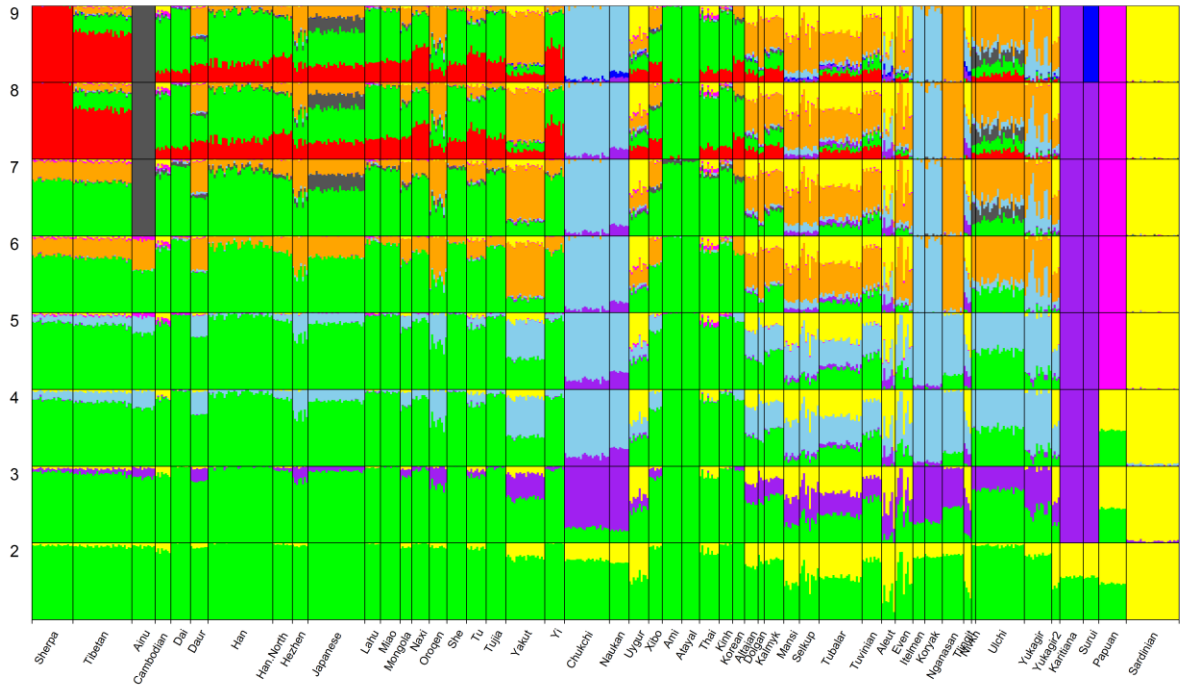
Supplementary Figure 4.3: Weighted admixture LD decay in the 10 admixed Ainu with the unadmixed Ainu and 1KG JPT as references. The estimated times (in generations) of the inferred admixture events and their standard deviations are shown in parenthesis.



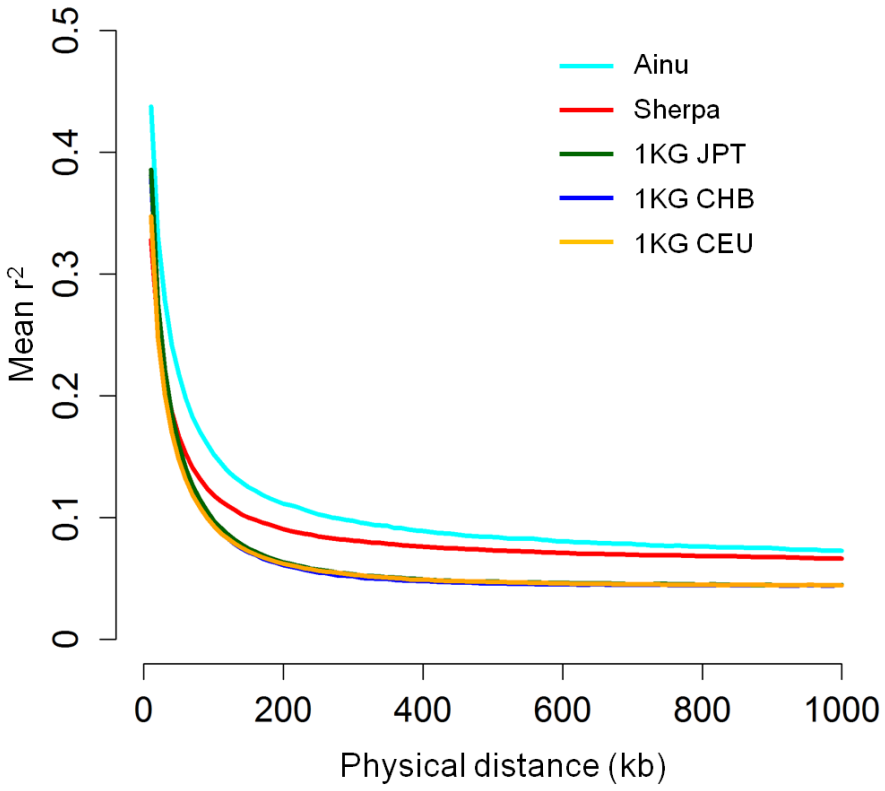
Supplementary Figure 4.4: *ADMIXTURE* analysis of East Asian and Siberian populations with K = 2 to 9. Ainu individuals are assigned to their own ancestry (dark grey) with K = 7, following separation of ancestry components concentrated in Sardinians (yellow), Native Americans (Karitiana and Surui; purple), Papuans (magenta), northeast Siberians (skyblue) and central Siberians and northeast Asians (orange; most concentrated in the Nganasan).



Supplementary Figure 4.5: *ADMIXTURE* analysis of East Asian and Siberian populations with K = 2 to 9. This analysis includes two Nivkh individuals, showing a substantial proportion of ancestry shared with the Ainu with K ≥ 7 (dark grey).

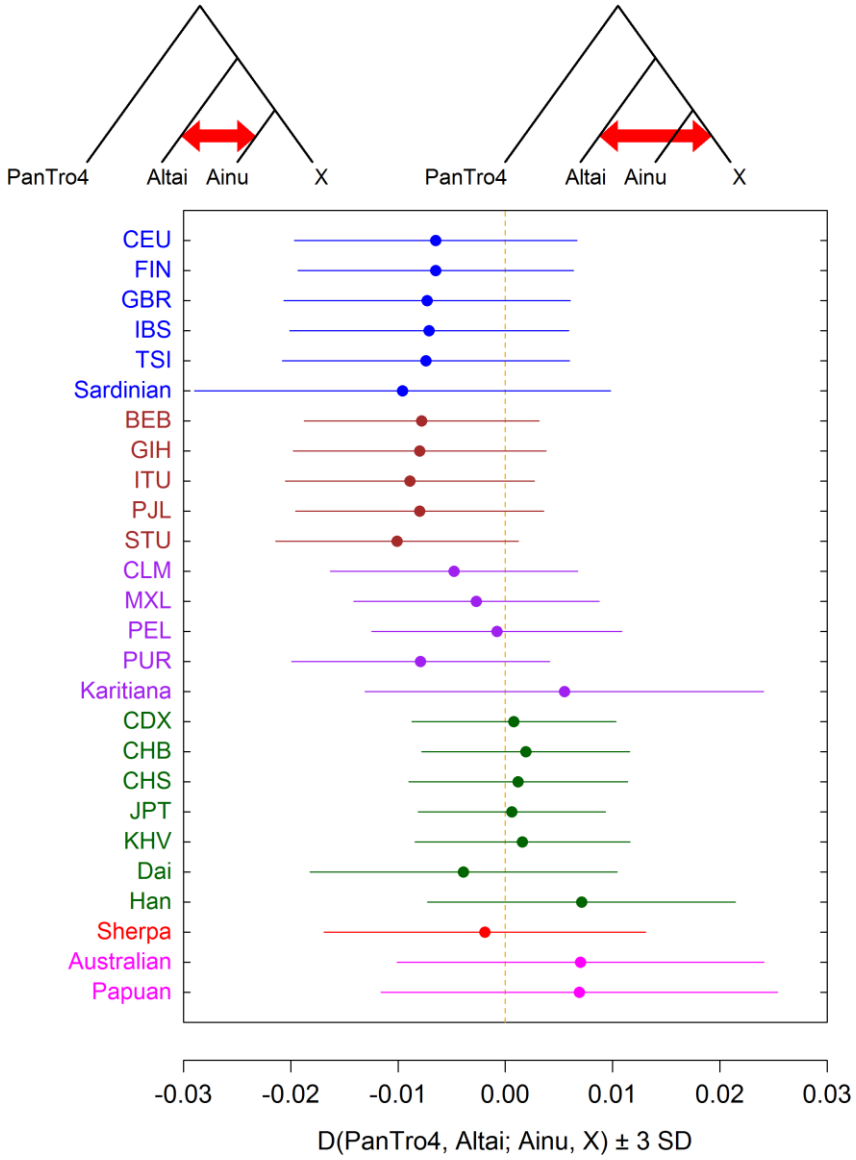


Supplementary Figure 4.6: LD decay across physical distance in the Ainu, Sherpa and 1KG populations. Mean r^2 values were calculated for all pairs of SNPs for 10 kb distance bins. For each population, we randomly chose 12 individuals and calculated LD to match the Ainu sample size.



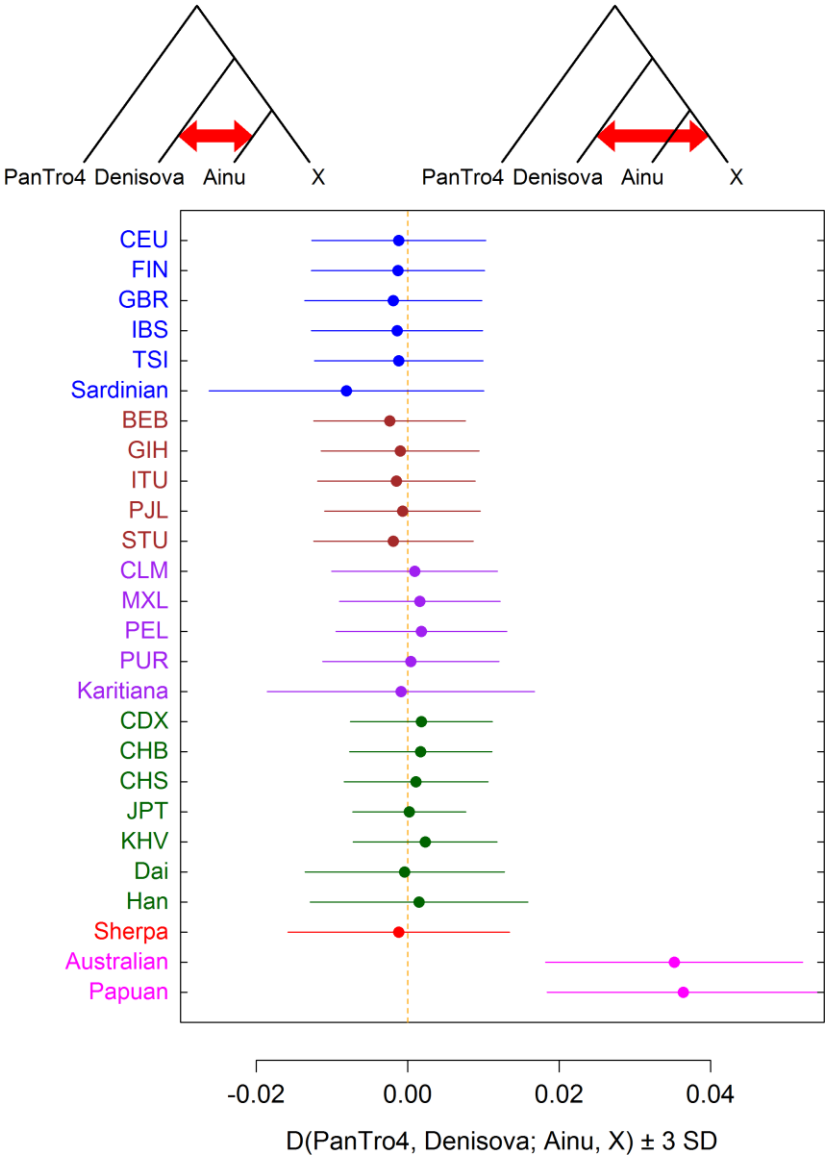
Supplementary Figure 4.7: Genetic affinity of the Ainu and other non-African populations to archaic hominins, (A) Altai Neandertal and (B) Denisovan, measured by Patterson's D(YRI, Archaic; Ainu, X). The Ainu show a similar level of archaic ancestry with the other East Asian populations (|D| < 1.5 SD). The "CND-1KG-Ainu" data set was used for this analysis. Horizontal bars around the value represent ± 3 standard deviations.

A



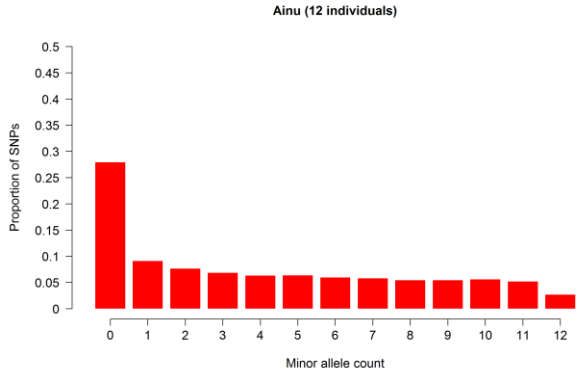
Supplementary Figure 4.7 – Continued.

B

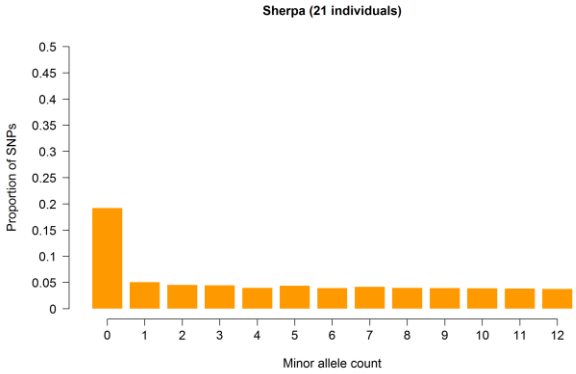


Supplementary Figure 4.8: Minor allele count distribution of SNPs in the "WHA" data set in (A) Ainu, (B) Sherpa, (C) Lahu, (D) Dai, (E) Atayal, (F) Ami, (G) Itelmen, (H) Nganasan, (I) Karitiana and (J) Surui. In all populations, minor allele count distribution was flat except for high numbers among the fixed SNPs (19% in the Sherpa to 47% in Surui). The Ainu do not have a particularly high proportion of fixed SNPs (28%) in comparison to the other East Asian, Siberian or Native American populations.

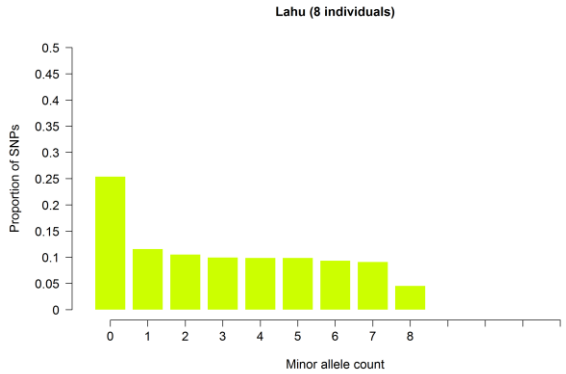
A



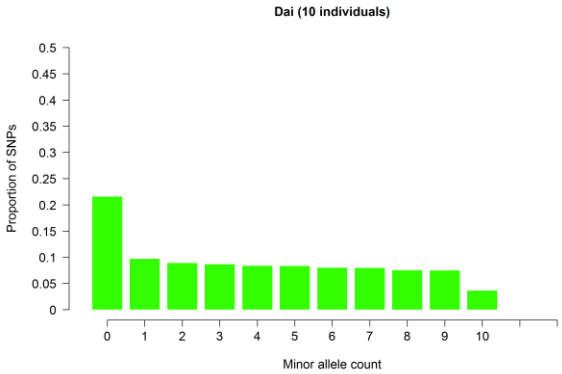
B



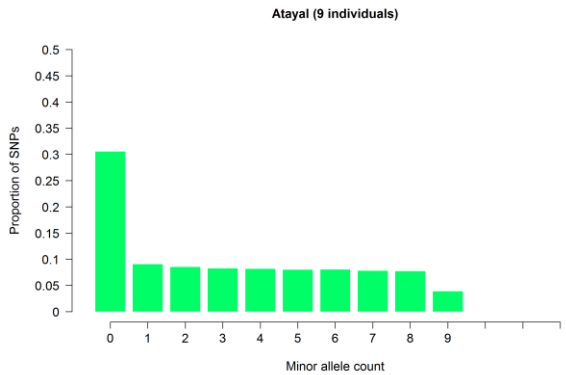
C



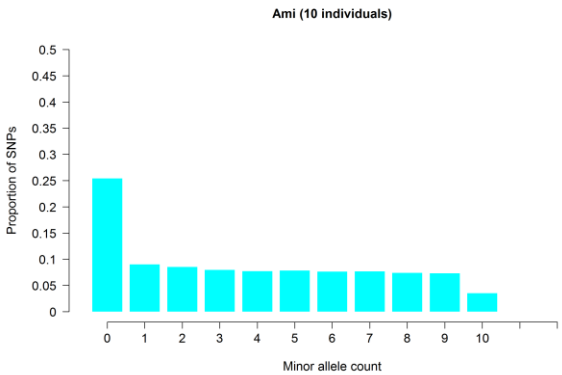
D



E

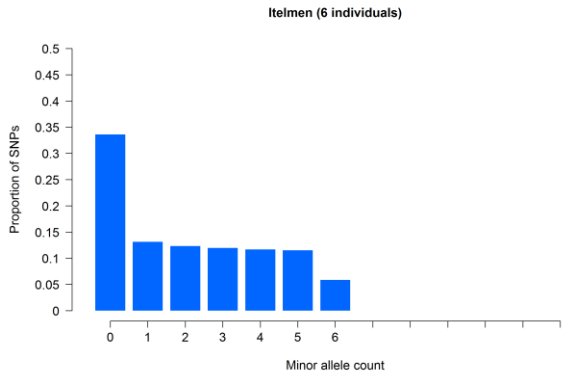


F

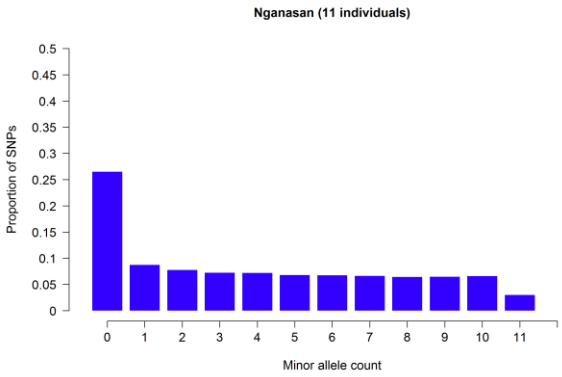


Supplementary Figure 4.8 – Continued.

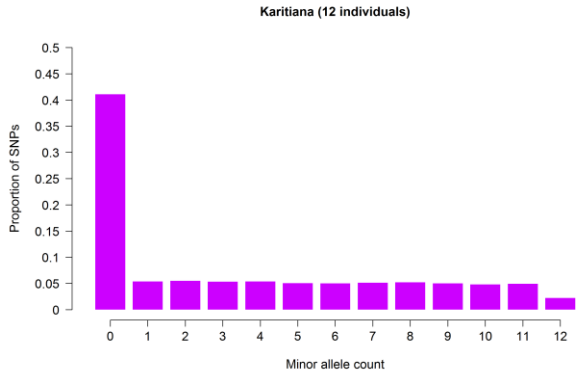
G



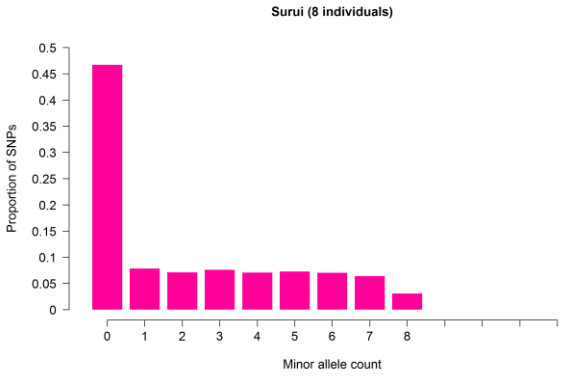
H



I

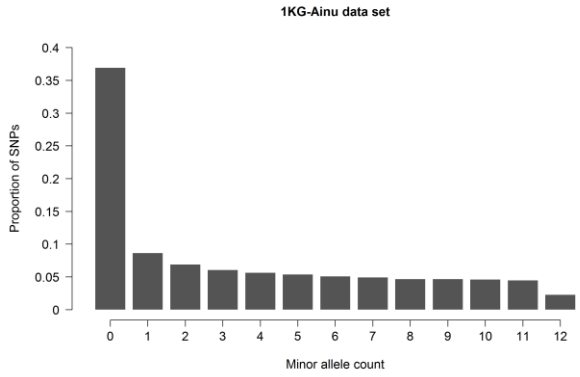


J

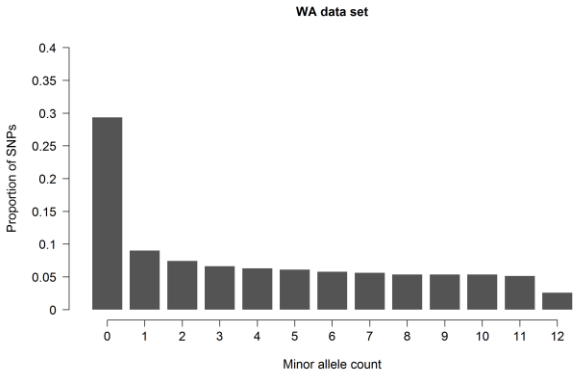


Supplementary Figure 4.9: Minor allele count distribution in the 12 Ainu individuals in (A) "1KG-Ainu" (540,304 SNPs), (B) "WA" (103,218 SNPs) and (C) "WHA" (45,513 SNPs) data sets.

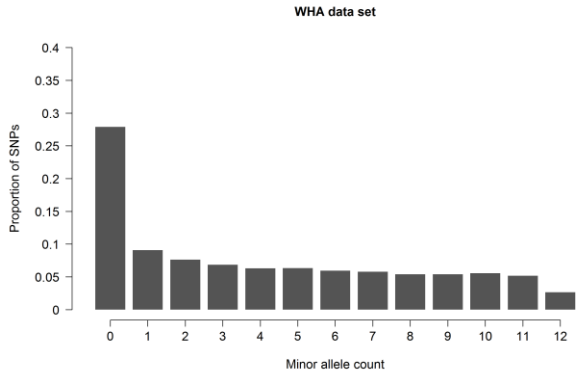
A



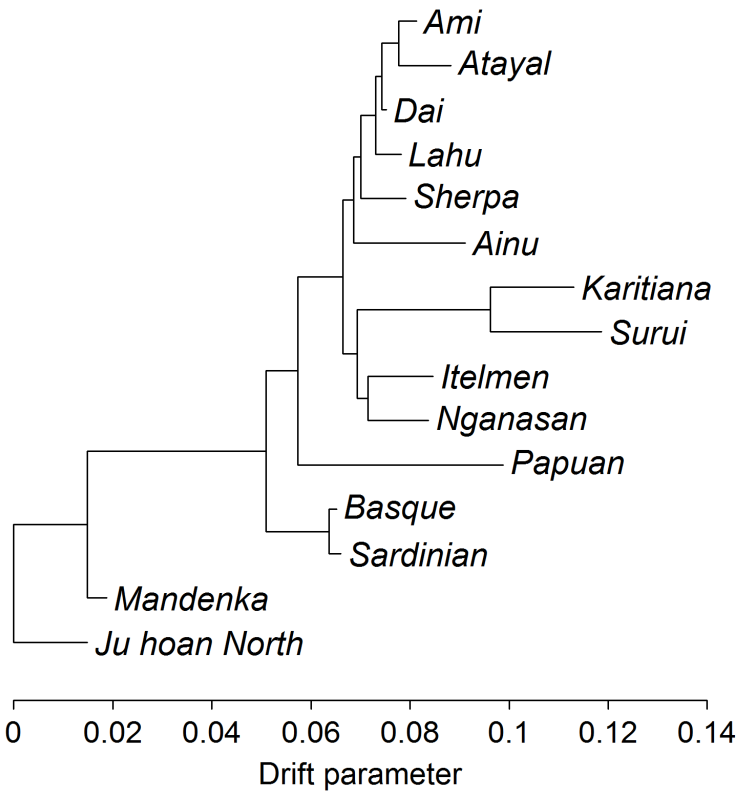
B



C

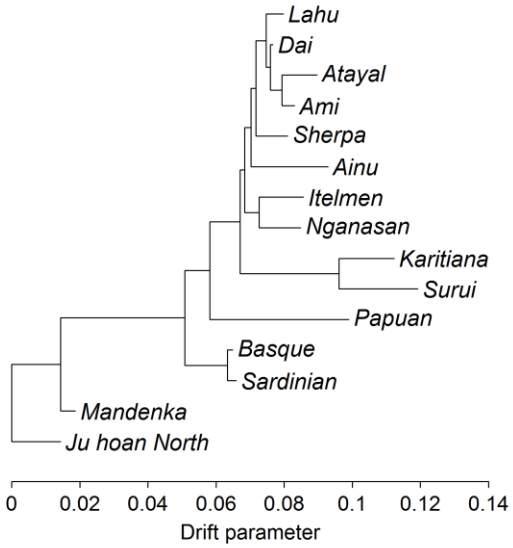


Supplementary Figure 4.10: A tree with minority topology of 15 world-wide populations inferred from 500 bootstrap replicates of maximum likelihood trees using *TreeMix*. This topology was supported in 23.2% of replicates. The only difference from a consensus tree in Figure 3 is the position of (Itelmen, Nganasan) clade as a sister group to Native Americans.

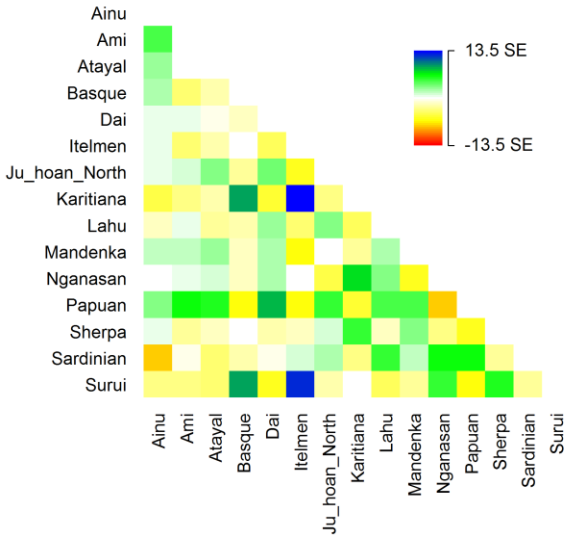


Supplementary Figure 4.11: *TreeMix* results with 0 to 5 migration edges. A single representative run was chosen from 100 bootstrap replicates. (A, C, E, G, I, K) Maximum likelihood tree with 0 to 5 migration edges. (B, D, F, H, J, L) Residual covariance matrices for the corresponding trees.

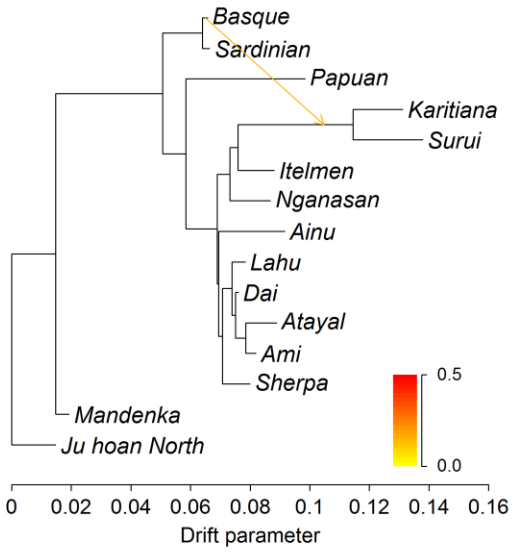
A



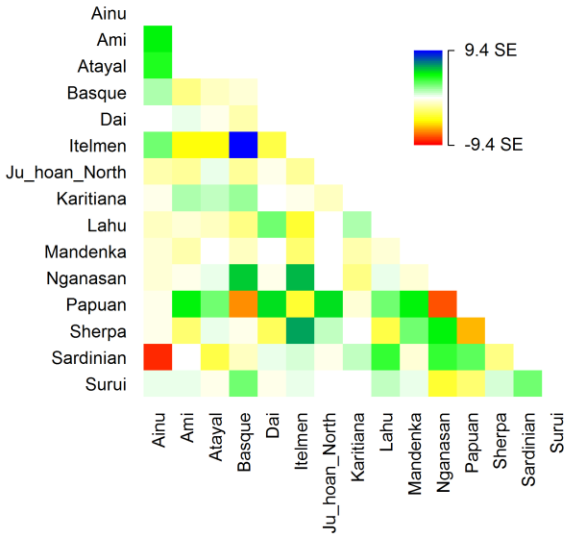
B



C

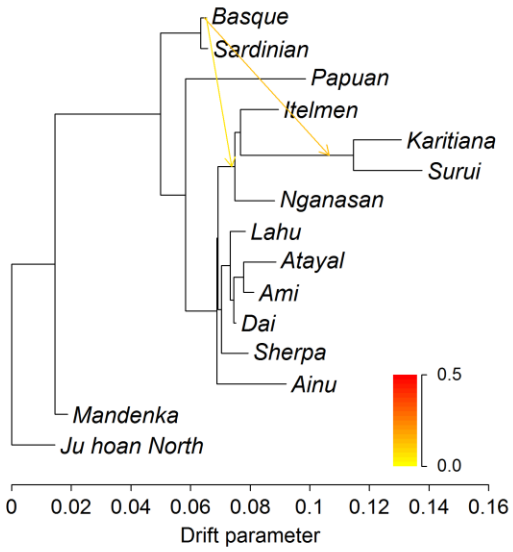


D

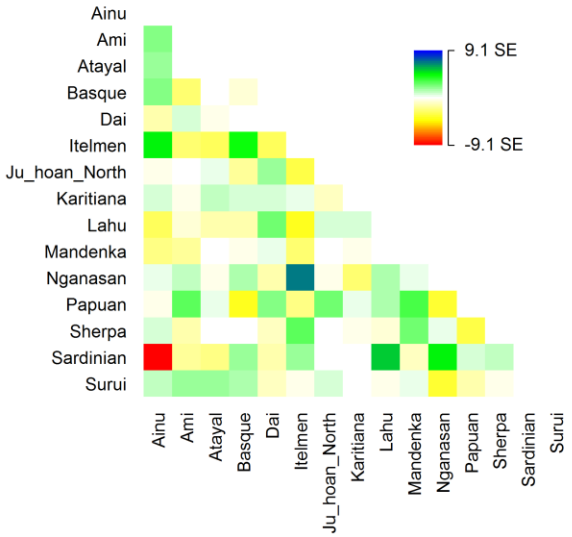


Supplementary Figure 4.11 – Continued.

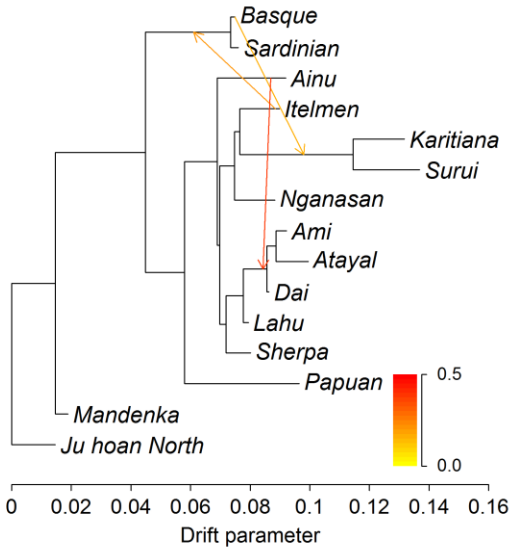
E



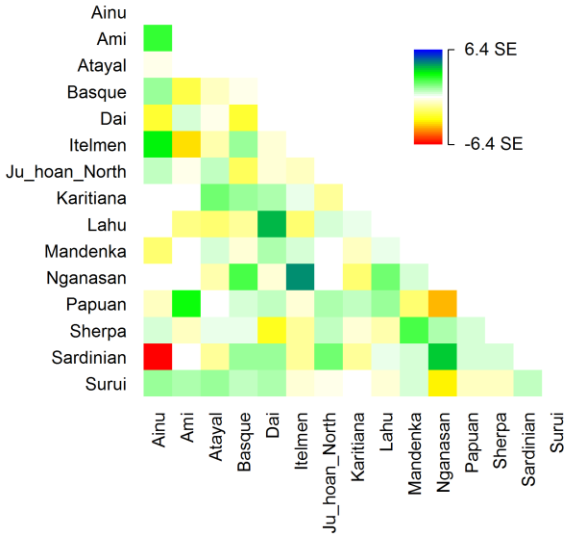
F



G

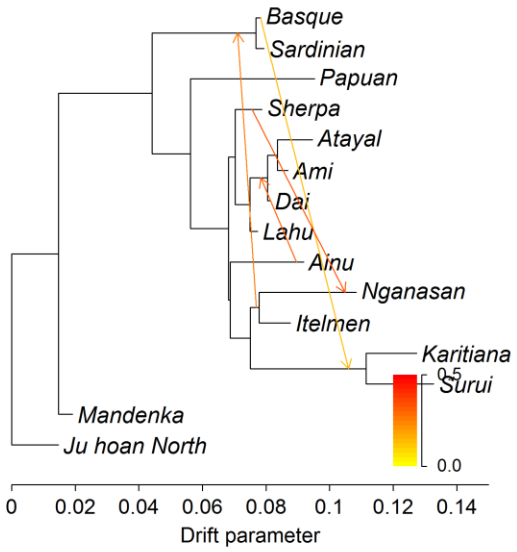


H

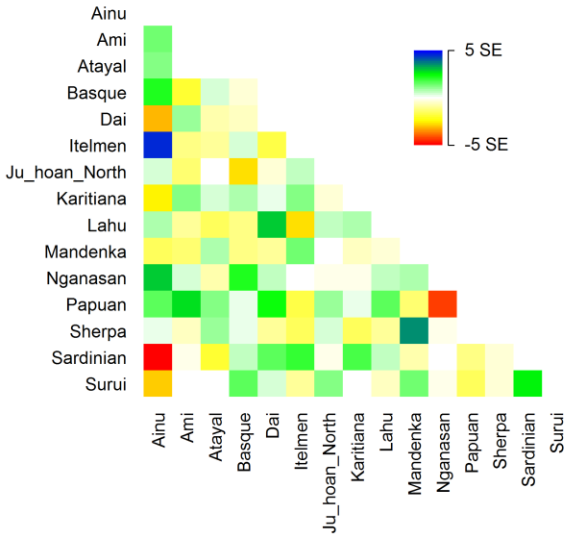


Supplementary Figure 4.11 – Continued.

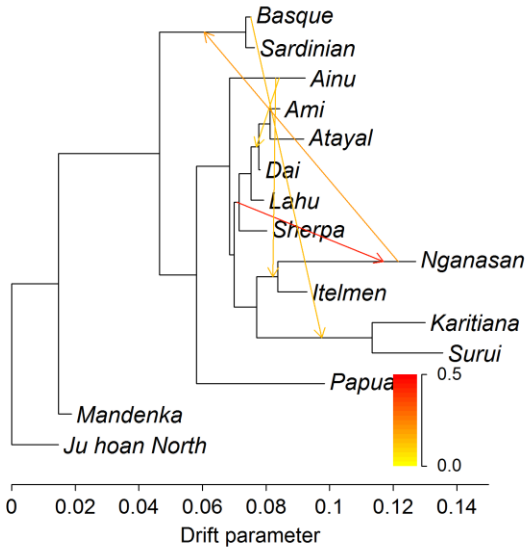
I



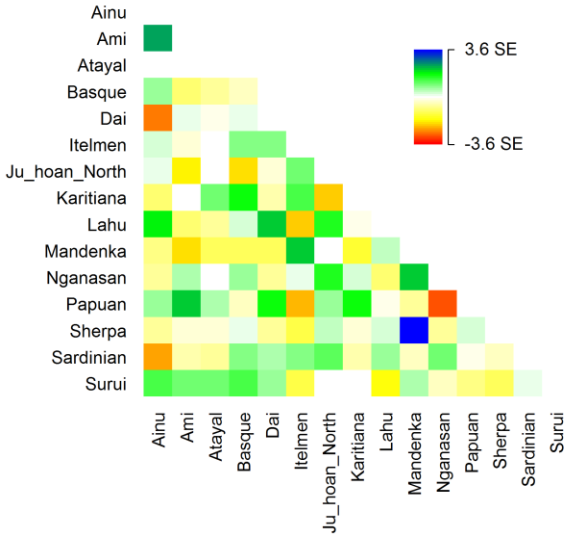
J



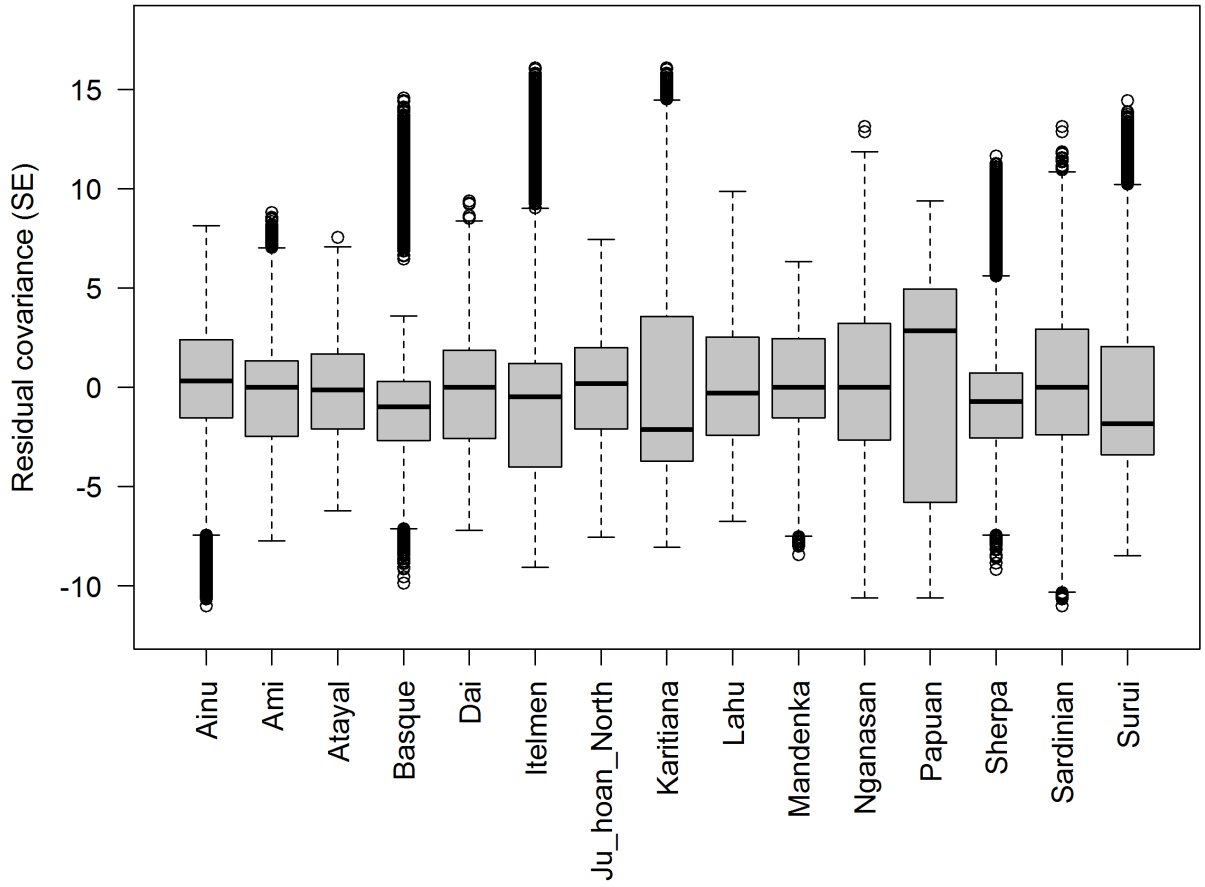
K



L

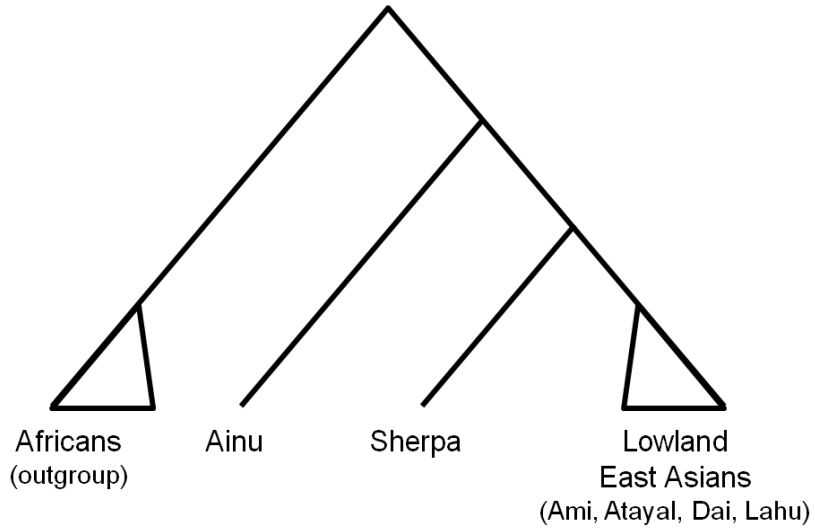


Supplementary Figure 4.12: Distribution of residual covariance (in standard error, SE) for each population across 500 bootstrap replicates of *TreeMix* with no migration edge allowed. All populations show small mean deviations from zero, suggesting that a population tree without migration edges does not fully explain the data. However, the Ainu show a similar level of deviation from zero, suggesting that they are not an unusual outlier in this analysis.

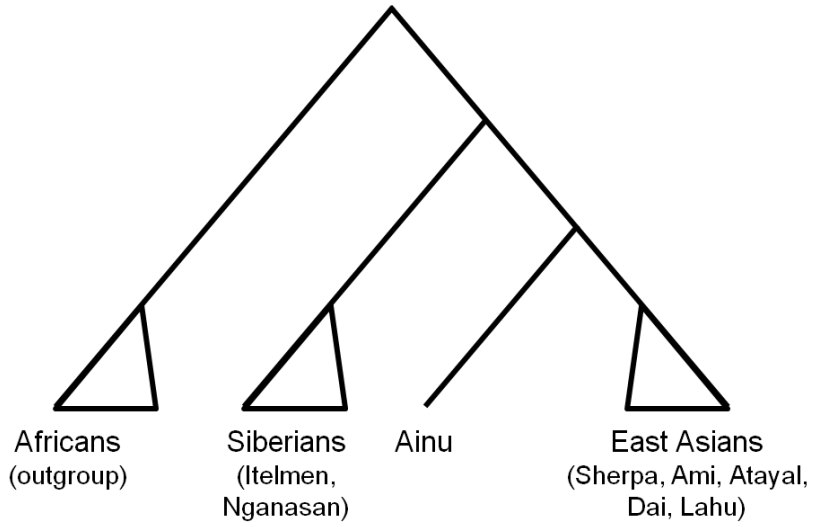


Supplementary Figure 4.13: Two hypothetical scenarios of population relationships based on the population trees. (A) The Sherpa and lowland East Asian populations (Ami, Atayal, Dai and Lahu) are equally related to the Ainu, their shared outgroup. (B) The Ainu and other East Asian populations including the Sherpa are equally related to Siberian populations, their shared outgroup.

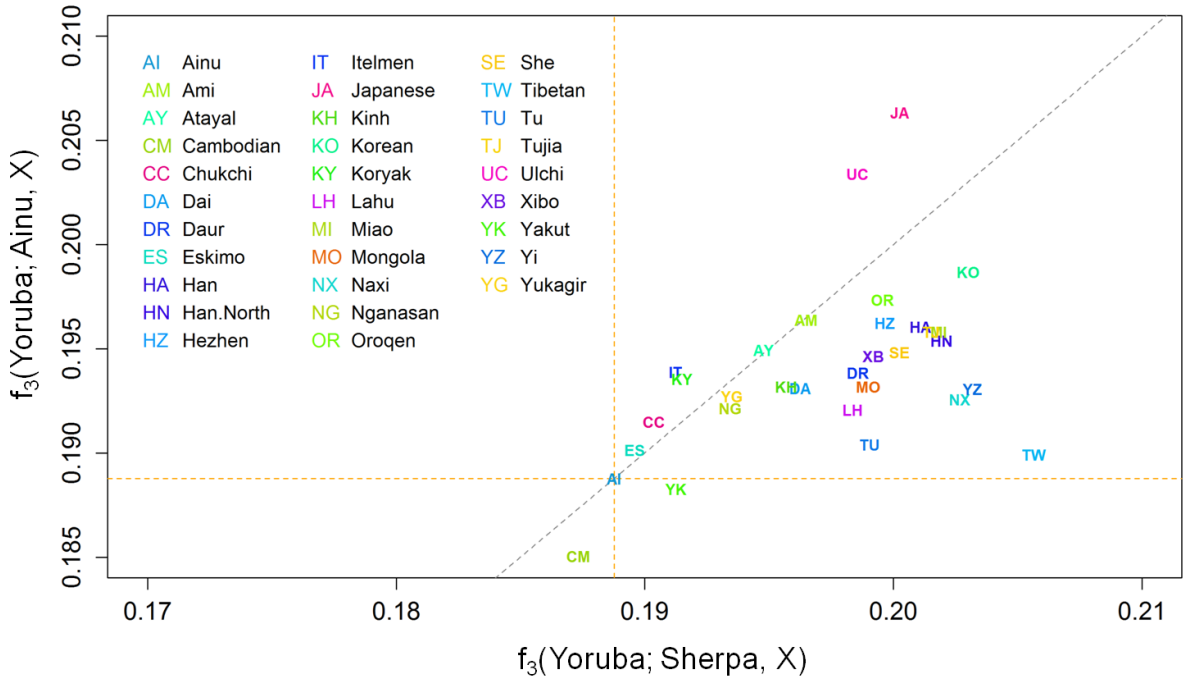
A



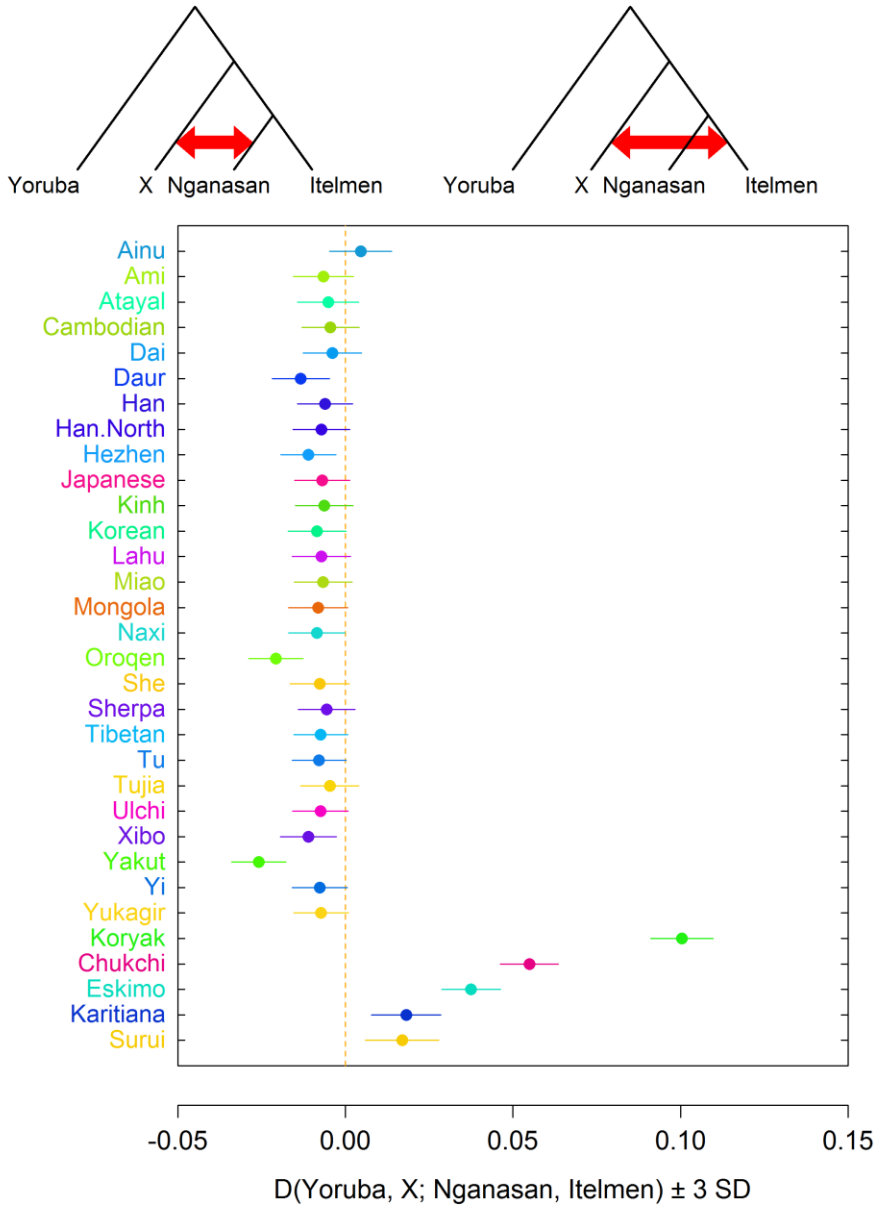
B



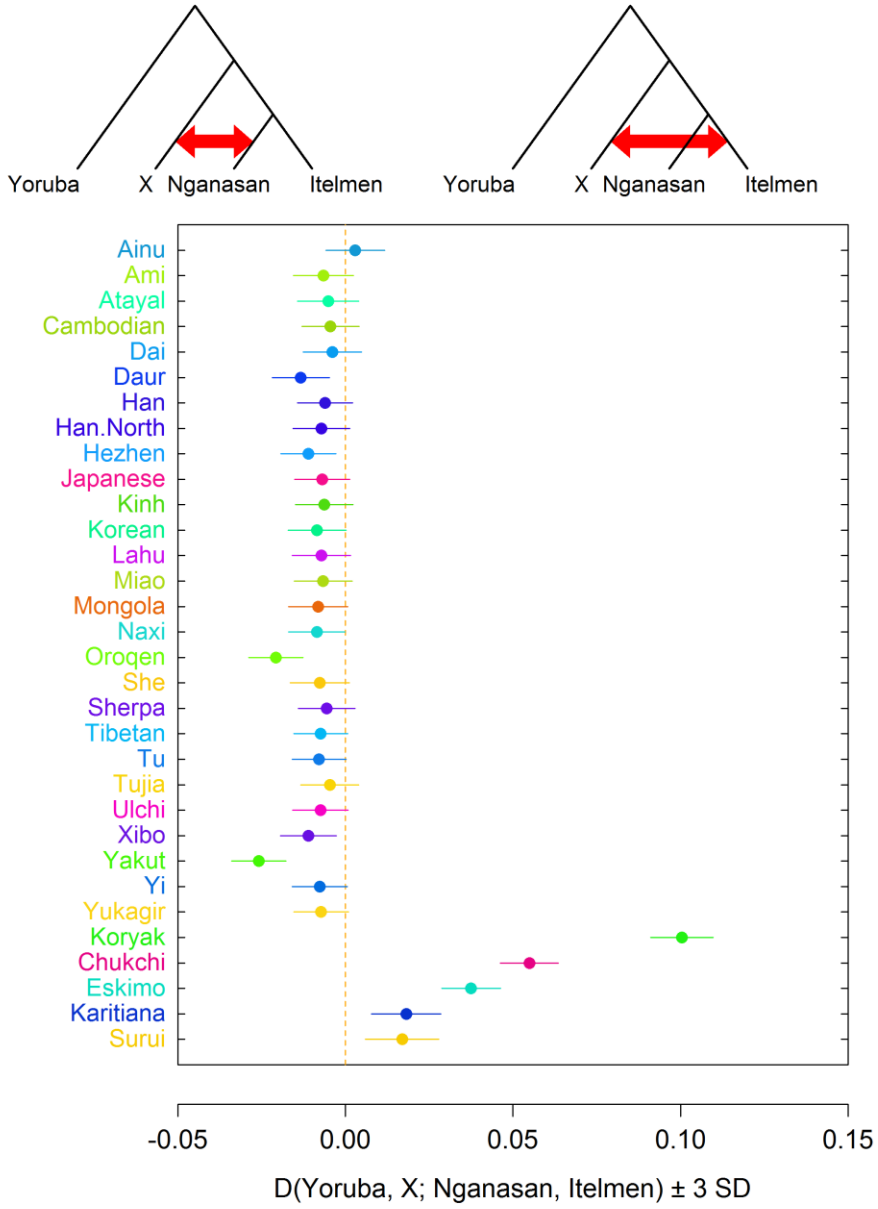
Supplementary Figure 4.14: Genetic affinity of East Asian and Siberian populations with the Ainu and the Sherpa measured by outgroup f_3 statistic. Most East Asian populations are closer to the Sherpa than to the Ainu, except for Japanese (JA), Ulchi (UC) and northeast Siberians (IT, KY, CC and ES). Sherpa and East Asians as well as Ainu and East Asians are in general closer to each other than Sherpa and Ainu (marked by dotted orange lines). The dotted grey line marks a line with slope of 1.



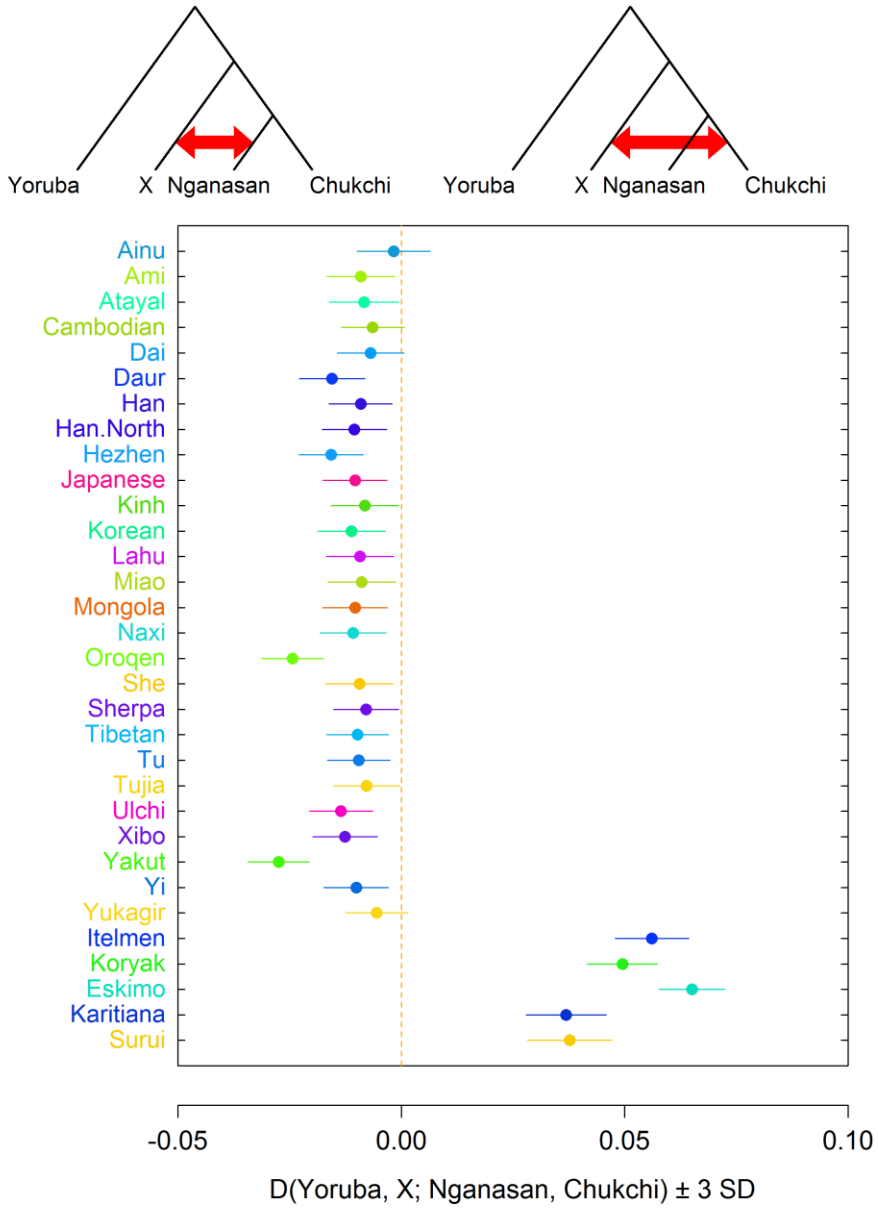
Supplementary Figure 4.15: The genetic affinity of East Asian and Siberian populations to Nganasan and Itelmen measured by Patterson's D(Yoruba, X; Nganasan, Itelmen). In this analysis, we used only the 12 Ainu individuals without non-Ainu ancestry. Both the Sherpa and Tibetans are closer to the Nganasan than to the Itelmen as the other East Asians are. The "WHA" data set used for this analysis includes the Sherpa and Tibetan samples. Horizontal bars around the value represent ± 3 standard deviations.



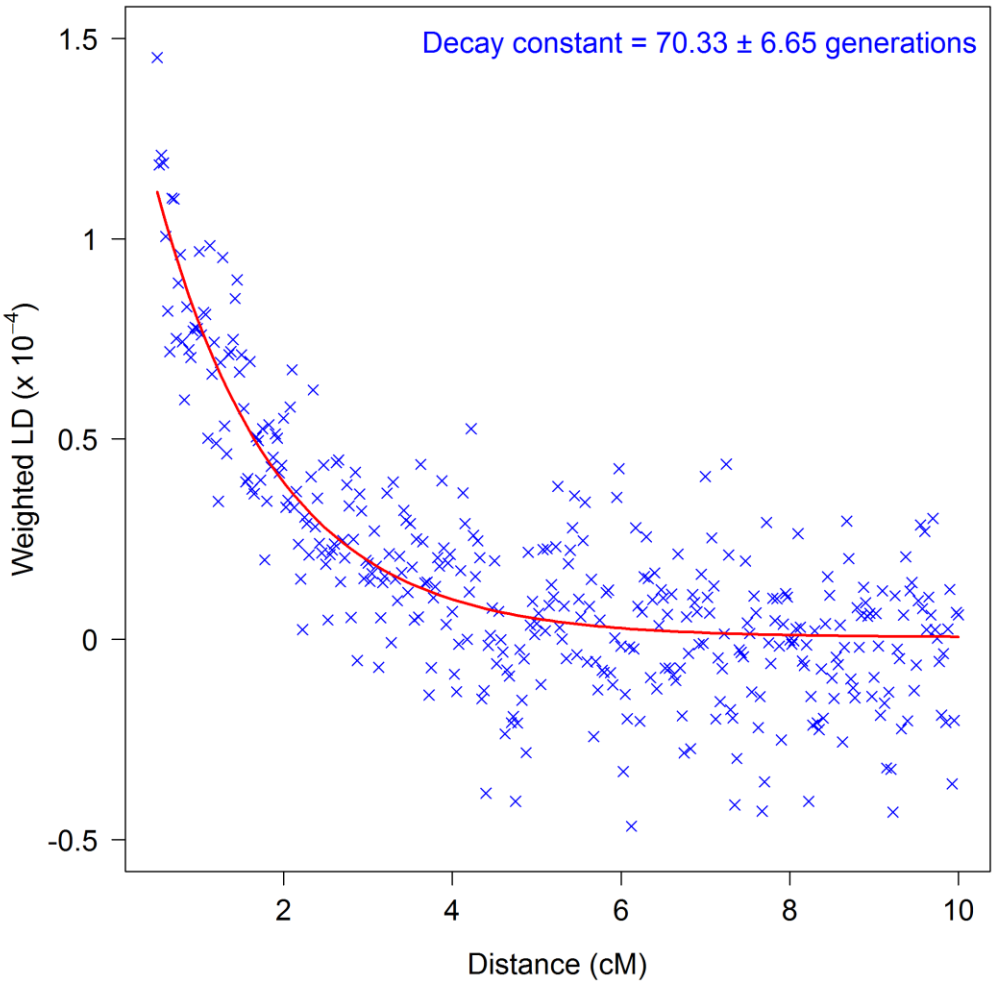
Supplementary Figure 4.16: Genetic affinity of East Asian and Siberian populations to Nganasan and Itelmen measured by Patterson's D(Yoruba, X; Nganasan, Itelmen). In this analysis, we used the 22 unrelated Ainu individuals including the recently admixed ones. Comparing these results to those in Figure S15, it is clear that the conclusions are not sensitive to the omission of the recently admixed Ainu. The "WHA" data set used for this analysis includes the Sherpa and Tibetan samples. Horizontal bars around the value represent ± 3 standard deviations.



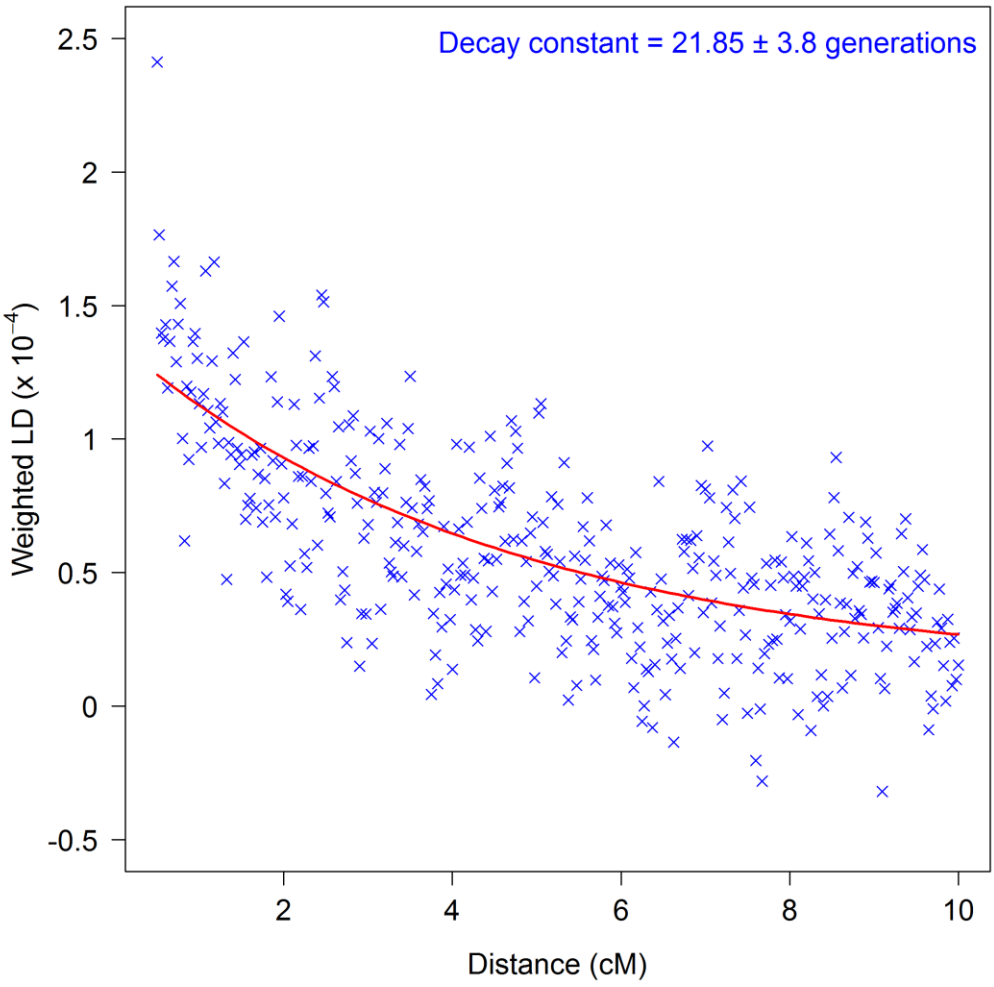
Supplementary Figure 4.17: Genetic affinity of East Asian and Siberian populations to Nganasan and Chukchi measured by Patterson's D(Yoruba, X; Nganasan, Chukchi). All East Asian populations except for the Ainu are significantly closer to the Nganasan. Native Americans (Surui and Karitiana) and northeast Siberians (Eskimo, Itelmen and Koryak) are significantly closer to the Chukchi. Horizontal bars around the value represent ± 3 standard deviations.



Supplementary Figure 4.18: Weighted admixture LD decay in the Japanese with the Ainu and the Han as references.

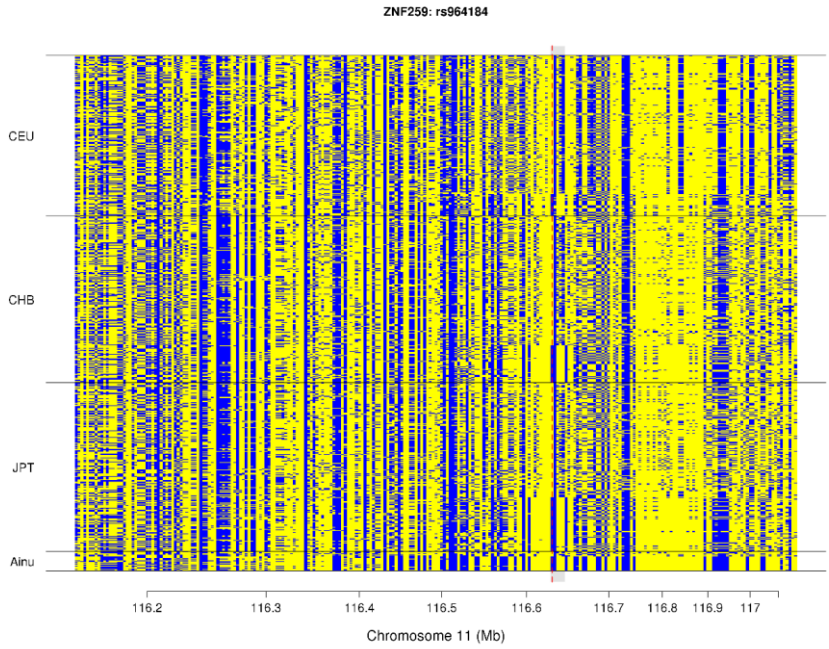


Supplementary Figure 4.19: Weighted admixture LD decay in Ulchi population with the Ainu and the Nganasan as references.

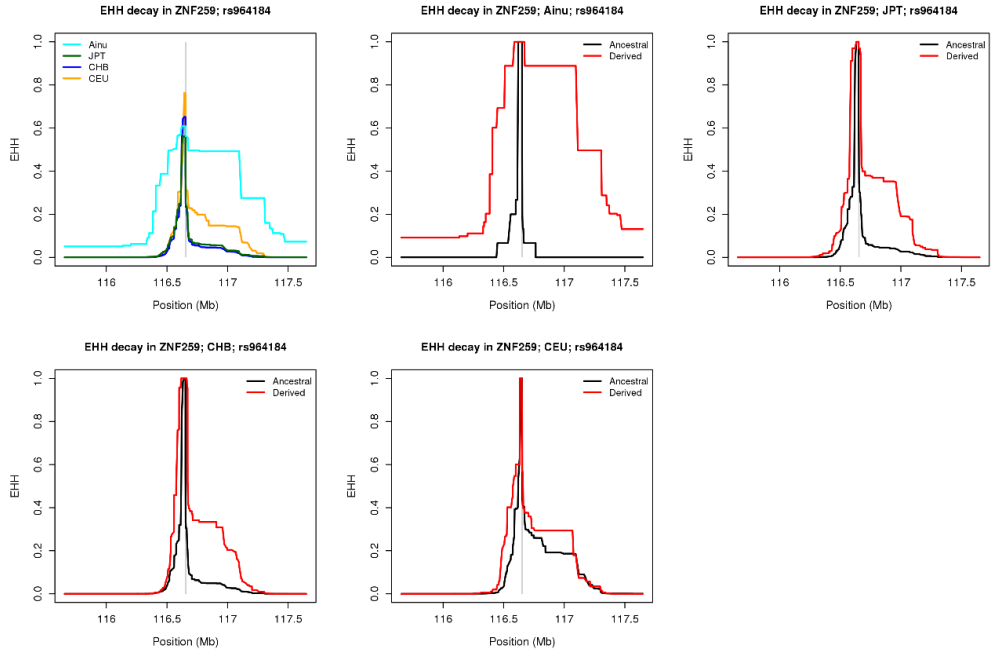


Supplementary Figure 4.20: Haplotype structure and EHH decay around rs964184 near the *APOA1* gene. (A) Haplotype structure around rs964184 in the Ainu and 1KG phase 3 JPT, CHB and CEU populations. Each column represents a variant and each row represents a phased haplotype. Yellow and blue colors represent ancestral and derived alleles, respectively. (B) EHH decay around rs964184. Ainu haplotypes harboring a derived allele at rs964184 show extended LD.

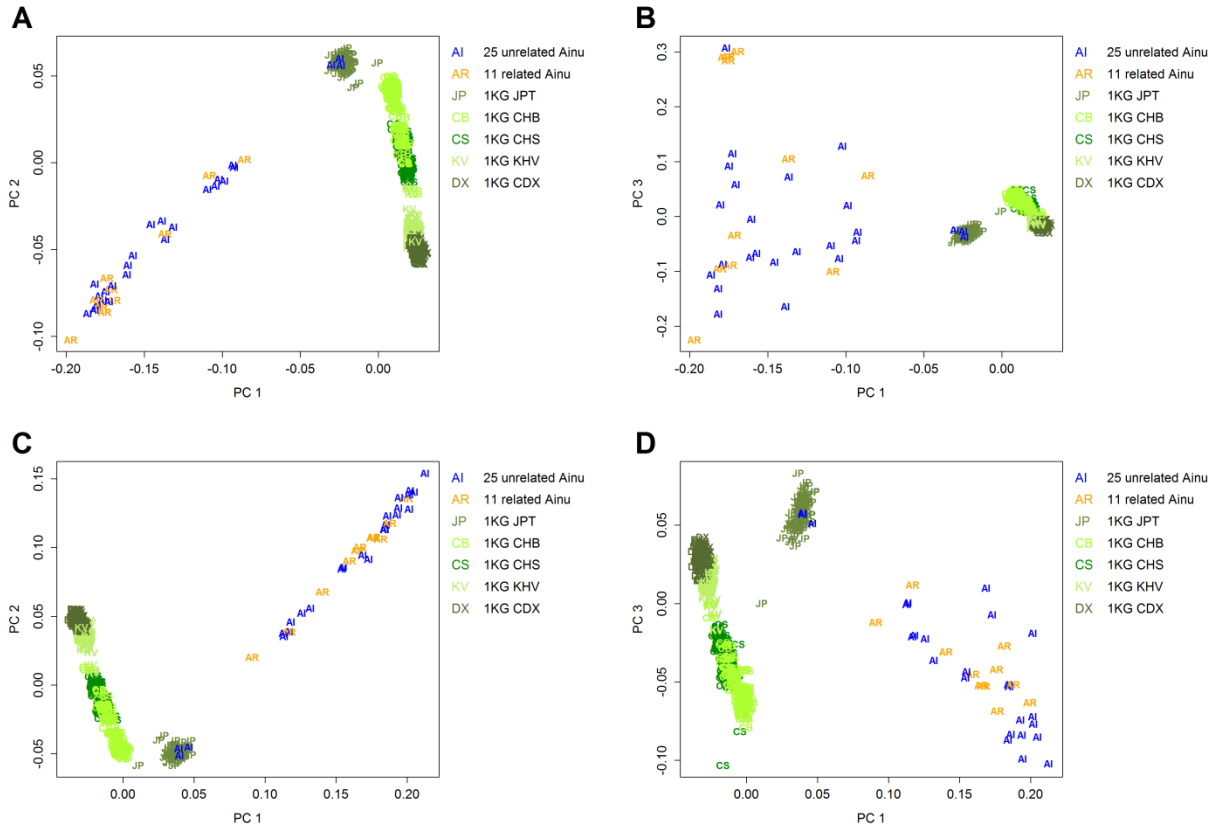
A



B



Supplementary Figure 4.21: The effect of including related Ainu individuals in PCA. Results with all 36 Ainu individuals and 1KG East Asians are presented in (A) PC1 vs. PC2 and (B) PC1 vs. PC3. Results with 25 unrelated Ainu individuals and 1KG East Asians are presented in (C) PC1 vs. PC2 and (D) PC1 vs. PC3.



Supplementary Table 4.1: A list of 71 populations used for calculating three-population (f_3) and Patterson's D statistics.

Chimp	Altai	Denisovan	LBK	Loschbour	MA1
BantuKenya	Biaka	Mandenka	Mbuti	Ju_hoan_North	Yoruba
BantuSA	Adygei	Basque	Bergamo	French	Orcadian
Russian	Sardinian	Tuscan	BedouinB	Druze	Mozabite
Palestinian	Balochi	Brahui	Kalash	Xibo	Cambodian
Dai	Daur	Han	Han_NChina	Hezhen	Japanese
Lahu	Miao	Mongola	Naxi	Oroqen	She
Tu	Tujia	Yakut	Yi	Chukchi	Eskimo
Sherpa	Tibetan	Ainu	Ami	Atayal	Kinh
Thai	Korean	Ulchi	Itelmen	Kalmyk	Koryak
Yukagir	Nganasan	Altaian	Mansi	Selkup	Tubalar
Tuvinian	Karitiana	Surui	Bougainville	Papuan	

Supplementary Table 4.2: Populations outside of East Asia have a symmetric relationship with the Ainu and East Asian farmer populations (Ami, Atayal, Dai, Lahu and Sherpa), suggesting that the Ainu do not harbor substantial non-East Asian ancestry. Patterson's D statistic was calculated in the form of D(Pop1, Pop2; Pop3, Pop4), using Yoruba as an outgroup (Pop1).

Pop1	Pop2	Pop3	Pop4	D	Z
Yoruba	Sardinian	Ainu	Ami	-0.0035	-1.331
			Atayal	-0.0043	-1.481
			Dai	-0.0057	-2.305
			Lahu	-0.0049	-1.907
			Sherpa	-0.0031	-1.249
	MA1		Ami	-0.0039	-0.714
			Atayal	-0.0028	-0.469
			Dai	-0.0009	-0.173
			Lahu	-0.0023	-0.423
			Sherpa	0.0037	0.696
	Karitiana		Ami	0.0045	1.269
			Atayal	0.0039	1.007
			Dai	0.0007	0.194
			Lahu	0.0026	0.704
			Sherpa	0.0017	0.480
	Papuan		Ami	0.0028	0.816
			Atayal	-0.0021	-0.597
			Dai	-0.0009	-0.286
			Lahu	-0.0021	-0.611
			Sherpa	-0.0033	-1.032

Supplementary Table 4.3: The Ainu form a clade with East Asian populations in comparison to contemporary populations outside of East Asia as well as to archaic hominins (|D| < 3 SD), which suggests that the Ainu do not harbor a substantial amount of non-East Asian ancestry. Based on the "CND-1KG-Ainu" data set, Patterson's D statistic was calculated in the form of D(Pop1, Pop2; Pop3, Pop4), using 1KG YRI as an outgroup (Pop1).

Pop1	Pop2	Pop3	Pop4	D	Z
YRI	Altai	Ainu	JPT	-0.0011	-0.491
	Neandertal		CHB	0.0000	-0.016
			CHS	-0.0002	-0.074
			KHV	-0.0006	-0.223
			CDX	-0.0013	-0.520
			Sherpa	0.0003	0.068
	Denisova		JPT	-0.0014	-0.679
			CHB	-0.0003	-0.115
			CHS	-0.0003	-0.124
			KHV	-0.0001	-0.064
			CDX	-0.0007	-0.322
			Sherpa	0.0008	0.219
	CEU		JPT	-0.0024	-1.539
			CHB	-0.0016	-0.933
			CHS	-0.0034	-1.993
			KHV	-0.0036	-2.108
			CDX	-0.0048	-2.819
			Sherpa	-0.0012	-0.488
	GIH		JPT	-0.0025	-1.749
			CHB	-0.0025	-1.593
			CHS	-0.0030	-1.891
			KHV	-0.0027	-1.757
			CDX	-0.0038	-2.457
			Sherpa	-0.0027	-1.275
	Papuan		JPT	-0.0010	-0.434
			CHB	-0.0020	-0.805
			CHS	-0.0009	-0.354
			KHV	-0.0013	-0.505
			CDX	-0.0015	-0.562
			Sherpa	-0.0075	-1.949

Supplementary Table 4.4: Major migration edges inferred from the *TreeMix* analyses, allowing 1 to 5 migration edges (m), suggest gene flow events between Europeans and Native Americans and/or Siberians, between the Ainu and lowland East Asian farmers, and to a lesser degree, between the Ainu and the Itelmen. Migration edges are counted across 100 bootstrap replicates. Here we tabulated migration edges inferred in more than 5% of replicates. When multiple populations are listed, the value refers to an internal branch that is the most recent common ancestor of all listed populations. Ainu related migration edges are highlighted in bold face. EA = East Asians, SB = Siberians, NA = Native Americans.

m	Source	Target	Count
1	Basque	(Karitiana, Surui)	77
	(Karitiana, Surui)	Basque	23
2	Basque or (Basque, Sardinian)	(Karitiana, Surui)	67
	(Karitiana, Surui)	(Basque, Sardinian)	33
	Itelmen	(Basque, Sardinian)	32
	Basque	(SB, NA)	31
	(EA, SB, NA)	(Karitiana, Surui)	10
	Sherpa	Nganasan	8
	Papuan	EA	14
3	Basque or (Basque, Sardinian)	(Karitiana, Surui)	74
	(Karitiana, Surui)	(Basque, Sardinian)	24
	Itelmen or (Itelmen, Nganasan)	(Basque, Sardinian)	41
	Basque	(SB, NA)	26
	Ainu	**(Ami, Atayal) or (Ami, Atayal, Dai)**	**43**
	Ami or (Ami, Atayal)	**Ainu**	**15**
	Sherpa	Nganasan	10
	Papuan	EA	13
	Ju_hoan_North	Papuan	12

(Continued in the next page)

Supplementary Table 4.4 – Continued.

m	Source	Target	Count
4	Basque or (Basque, Sardinian)	(Karitiana, Surui)	68
	(Karitiana, Surui)	(Basque, Sardinian)	28
	Itelmen, Nganasan or SB	(Basque, Sardinian)	36
	Basque	(SB, NA)	31
	Itelmen	Nganasan	9
	Ngansan	Itelmen	7
	Ainu	**(Ami, Atayal) or (Ami, Atayal, Dai)**	**52**
	Ami or (Ami, Atayal)	**Ainu**	**14**
	Itelmen	**Ainu**	**12**
	(EA, SB, NA)	(Karitiana, Surui)	14
	Sherpa	Nganasan	24
	(Ami, Atayal, Dai, Lahu, Sherpa)	Nganasan	7
	Papuan	EA	9
	Ju_hoan_North	Papuan	16
5	Basque or (Basque, Sardinian)	(Karitiana, Surui)	55
	(Karitiana, Surui)	Basque or (Basque, Sardinian)	46
	Itelmen, Nganasan or SB	(Basque, Sardinian)	39
	Basque	(SB, NA)	14
	Itelmen	Nganasan	8
	Ainu	**(Ami, Atayal) or (Ami, Atayal, Dai)**	**72**
	Ami or (Ami, Atayal)	**Ainu**	**17**
	Itelmen	**Ainu**	**17**
	Ainu	**Itelmen**	**6**
	Ainu	**(Itelmen, Nganasan)**	**12**
	(EA, SB, NA)	(Karitiana, Surui)	24
	Sherpa	Nganasan	24
	(Ami, Atayal, Dai, Lahu, Sherpa)	Nganasan	15
	Papuan	EA or (EA but Ainu)	18
	(Ami, Atayal) or Dai	Papuan	18
	Ju_hoan_North	Papuan	14

Supplementary Table 4.5: The Ainu are more closely related to lowland East Asian farmer populations (Ami, Atayal, Dai and Lahu) than to the Sherpa or to Tibetans, suggesting gene flow between the two groups after lowland East Asians split from the high-altitude East Asians. Choice of outgroup population did not affect results, suggesting that significant positive D statistic is not due to gene flow between the Sherpa or Tibetans and non-East Asian populations.

Pop1	Pop2	Pop3	D(Pop1, Pop2; Pop3, X) (Z)			
			Ami	Atayal	Dai	Lahu
Yoruba	Ainu	Sherpa	0.0192 (7.465)	0.0155 (5.532)	0.0109 (4.855)	0.0084 (3.334)
		Tibetan	0.0163 (7.064)	0.0126 (4.933)	0.0080 (4.051)	0.0054 (2.427)
Sardinian		Sherpa	0.0201 (8.473)	0.0170 (6.596)	0.0138 (6.449)	0.0104 (4.597)
		Tibetan	0.0158 (7.411)	0.0128 (5.342)	0.0095 (5.071)	0.0061 (3.007)
MA1		Sherpa	0.0257 (5.907)	0.0199 (3.903)	0.0157 (3.742)	0.0147 (3.444)
		Tibetan	0.0202 (5.097)	0.0144 (3.503)	0.0101 (2.835)	0.0091 (2.335)
Karitiana		Sherpa	0.0175 (6.112)	0.0143 (4.432)	0.0128 (4.835)	0.0080 (2.897)
		Tibetan	0.0144 (5.724)	0.0112 (3.817)	0.0097 (4.247)	0.0049 (1.943)
Papuan		Sherpa	0.0139 (4.933)	0.0150 (4.908)	0.0090 (3.578)	0.0075 (2.920)
		Tibetan	0.0093 (3.687)	0.0105 (3.825)	0.0044 (2.091)	0.0029 (1.264)

Supplementary Table 4.6: Northeast Siberians (Itelmen and Chukchi) are more closely related to the Ainu than to the other East Asians (Ami, Atayal, Dai, Lahu and the Sherpa). The Nganasan, a central Siberian population, do not show such an asymmetric relationship.

Pop1	Pop2	Pop3	Pop4	D	Z
Yoruba	Nganasan	Ainu	Ami	0.0021	0.670
			Atayal	-0.0008	-0.236
			Dai	-0.0021	-0.686
			Lahu	0.0004	0.120
			Sherpa	0.0032	1.032
	Itelmen		Ami	-0.0087	-2.710
			Atayal	-0.0102	-2.876
			Dai	-0.0102	-3.379
			Lahu	-0.0109	-3.318
			Sherpa	-0.0065	-2.060
	Chukchi		Ami	-0.0051	-1.639
			Atayal	-0.0074	-2.257
			Dai	-0.0072	-2.455
			Lahu	-0.0070	-2.222
			Sherpa	-0.0028	-0.955

Supplementary Table 4.7: Allele frequencies of three SNPs with selection signals in East Asians.

		Allele		Derived allele frequency					
Gene	SNP	Anc	Der	Ainu	JPT	CHB	AMR	EUR	AFR
EDAR	Rs3827760	A	G	0.250	0.803	0.937	0.392	0.011	0.003
OCA2	Rs1800414	T	C	0.875	0.572	0.592	0.000	0.000	0.001
ADH	Rs3811801	G	A	0.500	0.702	0.592	0.000	0.000	0.000

Anc = ancestral allele; Der = derived allele; AMR = 1KG phase 3 American populations; EUR = 1KG phase 3 European populations; AFR = 1KG phase 3 African populations

Supplementary Table 4.8: The list of 66 genomic regions harboring both XP-EHH and PBS signals in the Ainu, including top signal SNPs and the closest genes.

Region	XP-EHH				PBS			
	Top SNP	Top signal	Gene	Dist (kb)	Top SNP	Top signal	Gene	Dist (kb)
chr1:30817753-31217752	rs1188387	3.442	*MATN1*	34	rs4949290	1.470	*LAPTM5*	1
chr1:54467753-54967752	rs12734042	2.848	*SSBP3*	34	rs3753405	1.798	*SSBP3*	0
chr1:162017753-163067752	rs3927641	3.743	*NOS1AP*	0	rs10919316	1.280	*C1orf226*	3
chr1:175517753-175717752	rs10913052	3.039	*TNR*	0	rs12049604	1.065	*TNR*	0
chr1:232317753-232817752	rs12731562	2.990	*SIPA1L2*	0	rs10910538	1.219	*SIPA1L2*	0
chr2:13665703-13765702	rs2140525	2.852	*TRIB2*	794	rs6432398	1.108	*TRIB2*	814
chr2:48915703-49265702	rs4953650	3.031	*FSHR*	0	rs2268359	0.858	*FSHR*	0
chr2:133465703-133765702	rs7606532	3.125	*NCKAP5*	0	rs16841046	2.091	*NCKAP5*	0
chr2:134115703-134615702	rs1004045	3.316	*NCKAP5*	277	rs2012254	1.063	*NCKAP5*	0
chr2:139365703-139665702	rs12612135	3.312	*NXPH2*	107	rs10172094	1.110	*NXPH2*	113
chr2:150815703-151615702	rs2879927	2.945	*RND3*	129	rs16827946	1.128	*RND3*	287
chr3:11619244-11869243	rs301551	3.200	*TAMM41*	34	rs7618099	1.434	*TAMM41*	0
chr3:71269244-71369243	rs1522174	2.748	*FOXP1*	0	rs2037477	1.188	*FOXP1*	0
chr3:76769244-77219243	rs774590	3.290	*ROBO2*	268	rs1166980	1.150	*ROBO2*	160
chr3:77969244-78719243	rs1495598	3.624	*ROBO2*	407	rs4680999	2.217	*ROBO1*	33
chr3:80919244-81419243	rs17018503	3.151	*GBE1*	395	rs11714085	1.110	*GBE1*	189
chr3:95219244-96119243	rs6767219	3.676	*EPHA6*	439	rs9826768	2.510	*EPHA6*	906
chr3:136969244-137319243	rs1542535	2.632	*IL20RB*	336	rs12490164	1.566	*SOX14*	287
chr4:36418911-36518910	rs10001470	2.724	*DTHD1*	115	rs6531440	0.777	*DTHD1*	89
chr4:86718911-87218910	rs3796625	2.862	*ARHGAP24*	0	rs345326	1.910	*ARHGAP24*	0
chr4:111768911-112268910	rs12642421	2.965	*PITX2*	682	rs17042632	1.427	*PITX2*	352
chr4:188818911-189118910	rs7375901	3.959	*TRIML1*	13	rs12500011	1.102	*TRIML2*	0
chr5:17886344-18786343	rs6876500	2.789	*BASP1*	694	rs17631488	1.311	*CDH18*	695
chr5:77636344-79536343	rs784589	3.872	*ARSB*	0	rs16877259	1.451	*CMYA5*	44
chr6:4542106-5642105	rs12190412	3.696	*LYRM4*	0	rs7767658	1.756	*CDYL*	56
chr6:6792106-7592105	rs2764092	3.396	*SSR1*	0	rs9328401	1.753	*RREB1*	0
chr6:96142106-96692105	rs4839726	2.902	*MANEA*	106	rs4840245	1.594	*FUT9*	0
chr6:139742106-139992105	rs17406820	2.928	*CITED2*	74	rs9495561	1.430	*CITED2*	290
chr7:11293259-11893258	rs7782151	3.009	*THSD7A*	0	rs17165101	1.098	*THSD7A*	0
chr7:125743259-126093258	rs17149771	2.766	*GRM8*	2	rs671524	1.070	*GRM8*	280
chr7:134243259-134793258	rs6955887	3.554	*AKR1B15*	0	rs12666741	0.834	*BPGM*	9
chr7:152993259-153393258	rs84034	2.789	*ACTR3B*	475	rs6958821	1.259	*DPP6*	331
chr8:77365982-77515981	rs13259565	2.857	*ZFHX4*	151	rs16939290	1.083	*ZFHX4*	156

Supplementary Table 4.8 - Continued

Region	XP-EHH				PBS			
	Top SNP	Top signal	Gene	Dist (kb)	Top SNP	Top signal	Gene	Dist (kb)
chr9:7646587-8196586	rs4742455	3.164	*PTPRD*	196	rs2026463	1.078	*C9orf123*	48
chr9:71946587-72596586	rs11139898	3.103	*PTAR1*	9	rs1493048	1.160	*C9orf135*	42
chr9:93596587-93946586	rs192703	2.988	*SYK*	69	rs16907244	1.594	*AUH*	30
chr9:106346587-106596586	rs7034565	2.758	*SMC2*	326	rs10990903	1.741	*SMC2*	402
chr9:114246587-114896586	rs7030655	2.722	*C9orf84*	0	rs10981136	1.039	*UGCG*	14
chr9:121296587-121446586	rs7871273	3.055	*DBC1*	535	rs7024908	1.328	*DBC1*	585
chr9:133296587-133696586	rs1215988	2.796	*ASS1*	0	rs476067	1.497	*ASS1*	0
chr9:134596587-134996586	rs3012755	4.001	*MED27*	55	rs2987405	0.935	*MED27*	0
chr10:1604427-2054426	rs4457675	2.985	*ADARB2*	52	rs17156778	1.811	*ADARB2*	0
chr11:70798510-71548509	rs1792287	3.075	*DHCR7*	88	rs7125171	1.215	*DHCR7*	28
chr11:72198510-72698509	rs341078	2.692	*PDE2A*	30	rs11235622	1.415	*FCHSD2*	0
chr11:74498510-74998509	rs1111425	4.036	*XRRA1*	0	rs3824903	1.119	*SLCO2B1*	0
chr11:116498510-117048509	rs7123583	3.777	*BUD13 / APOA5*	19 / 60	rs2186670	1.697	*BUD13 / APOA5*	87 / 128
chr11:124598510-125198509	rs600702	3.077	*PKNOX2*	0	rs7129737	0.960	*ROBO4*	0
chr12:341619-1091618	rs524468	3.675	*SLC6A13*	0	rs2305164	2.294	*SLC6A13*	0
chr12:3641619-4041618	rs12370980	3.012	*EFCAB4B*	0	rs11062788	1.290	*EFCAB4B*	0
chr12:51791619-52641618	rs7974792	3.104	*SLC4A8*	0	rs17126441	0.841	*LOC283403*	0
chr12:52991619-53341618	rs694714	2.706	*KRT72*	0	rs17116681	2.050	*KRT72*	0
chr12:70991619-71491618	rs11178602	3.162	*TSPAN8*	27	rs2717420	1.090	*PTPRB*	0
chr12:108441619-108941618	rs1515633	2.670	*CMKLR1*	42	rs4964714	1.168	*FICD*	36
chr14:84274003-84674002	rs17119226	3.274	*FLRT2*	1639	rs4904162	1.168	*FLRT2*	1382
chr14:99624003-99924002	rs807734	2.791	*BCL11B*	0	rs17098384	1.503	*BCL11B*	0
chr15:28021673-28221672	rs4778192	2.629	*OCA2*	0	rs1800414	1.379	*OCA2*	0
chr15:34021673-34271672	rs1036004	2.720	*RYR3*	0	rs7496144	1.237	*AVEN*	0
chr15:50821673-51771672	rs17647084	3.183	*TNFAIP8L3*	0	rs2899473	1.444	*CYP19A1*	0
chr15:53571673-54121672	rs4776161	2.986	*WDR72*	0	rs518263	1.962	*WDR72*	0
chr15:92271673-92771672	rs17696427	3.387	*SLCO3A1*	12	rs8032332	1.006	*SLCO3A1*	75
chr16:51945481-52445480	rs2010842	3.459	*TOX3*	151	rs12597728	1.598	*TOX3*	119
chr16:84745481-85245480	rs8058723	3.215	*FAM92B*	99	rs16974564	0.999	*USP10*	0
chr18:49836305-50186304	rs12607660	3.027	*DCC*	0	rs919634	1.378	*DCC*	9
chr18:59486305-59586304	rs7505697	3.428	*RNF152*	0	rs12607798	1.123	*RNF152*	5
chr20:2961795-3461794	rs6133002	3.133	*PTPRA*	0	rs6107292	1.474	*ATRN*	11
chr20:43961795-44611794	rs459681	2.861	*SPINT4*	18	rs6032336	2.468	*WFDC8*	0

Supplementary Table 4.9: Top 10 genomic regions harboring extreme PBS signal in the Ainu in comparison to CHB.

Region	Top SNP	Anc[1]	Der[2]	Derived allele frequency		SNP location	GWAS catalogue[4] (based on the name of gene)	Empirical *P*-value[5]			
								N_e=2,219		N_e=1,000	
				EA[3]	Ainu			*T*=800	*T*=1,000	*T*=800	*T*=1,000
Chr1: 54467753−54967752	rs3753405	T	C	1.1%	54.2%	*SSBP3*	None	0.004	0.010	0.066	0.112
Chr2: 133465703-133765702	rs16841046	T	C	15.0%	87.5%	*NCKAP5*	Hypersomnia, glaucoma, and height	0.000	0.001	0.015	0.031
Chr3: 77969244-78719243	rs4680999	G	T	1.0%	70.8%	Upstream of *ROBO1*	Aspartate transaminase in liver, brain activity in schizophrenia patients	0.000	0.002	0.020	0.046
Chr3: 95219244-96119243	rs9826768	G	A	0.1%	70.8%	Upstream of *EPHA6*	Blood trace element (Cu and Zn levels), serum free IGF-1 level (obesity related)	0.000	0.001	0.017	0.040
Chr4: 86718911-87218910	rs345326	G	A	10.1%	75.0%	*ARHGAP24*	PR interval in electrocardiogram	0.002	0.004	0.028	0.050
Chr10: 1604427-2054426	rs17156778	G	A	12.2%	75.0%	*ADARB2*	Radiation response in LCL, Emphysema, BMI, % body fat	0.002	0.005	0.036	0.063
Chr12: 341619-1091618	rs2305164	C	T	2.8%	83.3%	*SLC6A13*	Serum metabolite level (pyroglutamine, deoxycarnitine, betaine,), glomerular filtration rate	0.000	0.000	0.008	0.020
Chr12: 52991619-53341618	rs17116681	G	A	18.4%	91.7%	*KRT72*	None	0.001	0.001	0.013	0.023
Chr15: 53571673-54121672	rs518263	T	C	1.6%	62.5%	*WDR72*	Blood urea nitrogen level, serum creatinine level, glomerular filtration rate, longevity, cognitive function	0.001	0.005	0.038	0.076
Chr20: 43961795-44611794	rs6032336	A	G	95.4%	20.8%	*WFDC8*	None	0.000	0.000	0.012	0.030

[1] Ancestral allele; [2] Derived allele; [3] 1KG phase 3 East Asians (excluding JPT); [4] Last date accessed to the database (https://www.genome.gov/26525384) is April 8th, 2015; [5] Empirical *p*-values were calculated as the proportion of simulations with their simulated frequency of the derived allele equal to or greater than the observed frequency in Ainu.

CHAPTER 5: CONCLUSIONS

The Tibetan plateau, sometimes referred to as “the roof of the world” or “the third pole”, is a land of extreme conditions for living, including hypobaric hypoxia, strong UV radiation and sparse resources. However, it is home to numerous populations and species of indigenous organisms, including humans, thriving in this seemingly barren landscape. For example, dozens of hand and footprints, found at 4,200 m.a.s.l. in Chusang (Tibet Autonomous Region, China) and dated between 11 to 32 kya (Zhang and Li 2002; Aldenderfer 2011; Brantingham et al. 2013), vividly witness a long history human presence in this part of the world. The Himalayas form the southern margin of the Tibetan plateau and coincide with a sharp cline of geography, climate and biodiversity. In human biogeography, the Himalayas mark the sharp transition area between East Asia and South Asia, in terms of genes, cultures and languages (Wang et al. 2012; Gayden et al. 2013). Such an abrupt change in the genetic profile over a short geographic distance is extremely rare in humans, whose biogeography is best explained by an isolation-by-distance pattern (Rosenberg et al. 2002; Li et al. 2008). Therefore, it is of common interest for human geneticists, archaeologists, biological anthropologists and evolutionary biologists to understand which environmental factors actually cause this barrier to persist and what kinds of genetic changes allow organisms to overcome it. In the studies presented in this dissertation, I tried to provide aspects of Tibetan genetic history, which eventually may help us understand how these remarkable people adapted to their native environments.

At least two distinct genetic ancestries are found in Tibetan populations

Results presented in Chapter 2 support the idea that contemporary Tibetans are a mixture of two genetic ancestries, a "high altitude" ancestry shared by all sampled Tibetans, but most enriched in the Sherpa, and a "low altitude" ancestry represented by various farming populations in lowland East Asia. The genetic profile of ancient Tibetans from the ACA region in Nepal, as presented in Chapter 3, also shows that contemporary Tibetans are their closest relatives among all contemporary East Asians used in the study. This suggests that the Tibetan gene pool had already diverged at least 2,500 BP, in contrast to a previous proposal (Yi et al. 2010). It is interesting that the ACA populations from all the periods tested have the adaptive *EGLN1* variant rs186996510 at high frequency, because it suggests that their ancestors were under altitudinal stress for a long enough time to allow this variant to evolve. This observation goes along well with estimates on the time of selection on this variant, ranging between 8,000 to 8,500 BP based on the LD pattern around it in contemporary Tibetans (Xiang et al. 2013; Lorenzo et al. 2014). Additional genetic data sampled across the Tibetan plateau and adjacent regions, especially ancient genomes of Tibetans and related Tibeto-Burman speaking groups, will be critical to refine the genetic history of the high altitude ancestry in East Asia.

Tibetan genetic adaptations to high altitudes may have evolved in multiple phases

It was a surprise to find evidence for a much more recent evolution of the *EPAS1* haplotype than the *EGLN1* one in the ancient genomes of the ACA samples, as presented in Chapter 3. Considering an archaic origin of the *EPAS1* haplotype, this haplotype must have stayed in very low frequency for about 50 kya without being lost, and then swept to high frequency in Tibetans over the last 2 kya. Why was there a delay of several thousand years for it

to be selected if the ancestral Tibetans arrived in the high altitude at least 8 kya as the *EGLN1* gene suggests? It must be pointed out that ancient samples from the ACA region may not be representative of the Tibetan gene pool as a whole. For example, the *EPAS1* haplotype might have arisen to high frequency much earlier in the north of the Himalayas, while it was either lost or drifted at low frequency in the isolated ACA population. Under this scenario, a later migration of Tibetans with this haplotype into the ACA region might have resulted in a recent increase in its frequency as observed in this study. Additional ancient genomes from the Tibetan plateau and the Himalayas are necessary to track the evolution of the *EPAS1* haplotype across time and space. However, if the pattern observed in the ACA region is indeed a proper representation of the evolutionary history of the *EPAS1* haplotype in Tibetans, more complicated scenarios may be worthy of consideration. These scenarios include, to name a few, that i) the *EPAS1* haplotype was introduced into Tibetans around the time of its selection through an unknown carrier population, ii) it was selected for a function different from that of the *EGLN1* haplotype, which became important later around that time, or iii) it works by interacting with other adaptive variants (such as the *EGLN1* one) so selection on these interacting variants must precede selection on it. All of the above scenarios, and others one can imagine, can be formulated into specific hypotheses to be tested. For all these scenarios, it is important to better understand the physiological effects of the *EPAS1* haplotype and the environmental factors conferring the fitness advantage.

Although the *EPAS1* gene encodes a transcription factor directly regulating transcription of the *EPO* gene, functions of the Tibetan *EPAS1* haplotype are still poorly understood because of multiple reasons. First of all, no protein coding variant was found on this haplotype and potentially important regulatory elements are in strong LD with each other. Therefore,

population genetics has a limited resolution in pinpointing the causal mutations in this case. Second, it is still not well understood how low hemoglobin level and oxygen content, for which strong evidence of statistical association with the Tibetan *EPAS1* haplotype exists, are connected to better survival and/or reproduction at altitude. It is possible that high hemoglobin level is particularly harmful at certain stages of life history, such as during pregnancy, as preeclampsia cases among Andean highlanders suggest. It is also a possibility that yet unknown phenotypes governed by the *EPAS1* gene are the actual targets of natural selection, and low hemoglobin is just a by-product of this selective event with no fitness effect. Functional genomics and genome editing approaches should be considered for future researches to resolve this issue.

Jomon and Tibetan ancestries represent two East Asian gene pools distinct from one of lowland farmer populations

The transition from foraging to agriculture made a huge impact on human history all around the world. The spread of agriculture, either in the form of sedentary farming or nomadic pastoralism, was frequently associated with a large-scale population expansion (a "demic" diffusion), as shown for the expansion of Neolithic farmers from Middle East to Europe (Lazaridis et al. 2014; Haak et al. 2015) and that of Bantu-speaking farmers from west to east and south Africa (Tishkoff et al. 2009; Hellenthal et al. 2014). East Asia was also heavily affected by the agricultural transition, which began as early as 9 kya (Molina et al. 2011; Yang et al. 2012; Callaway 2014). Several regions in East Asia have been hypothesized as "domestication centers" of many plant and animal species, including the lower basin of the Yangtze river for rice (Fuller et al. 2010; Molina et al. 2011; Callaway 2014), the Yellow river basin and western Manchuria for millets (Crawford 2006; Yang et al. 2012), and the Tibetan

plateau for Yaks (Qiu et al. 2012; Qiu et al. 2015). However, our understanding remains poor on important aspects such as the pre-agricultural genetic structure of East Asians and genetic changes caused by the agricultural expansion. The Japanese archipelago is a unique place in East Asia where archaeological and genetic evidence precisely track the spread of paddy field rice farming, brought by the "Yayoi" people. Here, the demic diffusion model provides a good explanation for the spread of farming, with increasing amounts of admixture with indigenous hunter-gatherers ("Jomon" people) in regions further away from northern Kyushu, where the Yayoi culture first appeared. Although the admixture process *per se* is well understood, the origins of Jomon and Yayoi people so far have not been well resolved. In the study presented in Chapter 4, I tried to illuminate the origin of the Jomon ancestry by using the contemporary Ainu people as its best representative. A long-held hypothesis of shared ancestry of Tibetans and the Ainu was not supported by genetic evidence analyzed here. Instead, the Ainu ancestry formed an outgroup of both Tibetans and other East Asian farmers. This suggests that Tibetans and the Ainu harbor two distinct ancestries in East Asia, differentiated yet from the third "lowland farmer" ancestry. In future studies, it will be interesting to see how these three ancestries are related to the first migration into northeast Asia and Siberia, which eventually led to the gene pool of Native Americans. Considering a complex pattern of gene flow between contemporary populations, ancient DNA could provide critical references to disentangle the evolution of East Asian gene pool across time and space.

Overall, this dissertation provides a body of new information to understand the genetic history of Tibetans and their altitude adaptations. Understanding the molecular changes enabling their long-term residence in the extreme environments of the Tibetan plateau, as well as the

specific selective pressures that shaped such changes, have wide implications not only on evolutionary genetics but also biomedical studies of hypoxia. Finally, additional population genetic studies in East Asia and Siberia may reveal how modern humans could have adapted to such an array of diverse environments, ranging from tropical rainforests in Southeast Asia to thin air in the Tibetan plateau to vast dry grasslands in Central Asia and finally to freezing dark nights in the Arctic Circle.

Future Directions: Mapping the genetic background of Tibetan adaptive phenotypes and functionally validating the identified loci

It is of great importance to identify environmental factors working as selection pressure, phenotypes conferring fitness advantages, and the genetic basis of such phenotypes in Tibetans. A straightforward approach is to collect a large number of candidate phenotypes in a sizeable Tibetan cohort together with genotype data for performing GWAS. I am currently involved in such a project in which about a thousand ethnic Tibetan women from Nepal were recruited. Together with hemoglobin and arterial oxygen saturation levels, measures of life-time reproductive success, such as the numbers of pregnancies and children born alive, survived to age 1, 5 or 15 years old, were collected as target phenotypes. Therefore, participants were limited to women of age 39 or older, to minimize errors due to unrealized reproductive potential. Genetic association with these phenotypes and search for signatures of positive selection for them are in progress. Once candidate genes and variants are identified, one may consider functional analyses of genetic association and population genetic signals. For example, identifying the causal mutations in the *EPAS1* haplotype is likely to be dependent on

experimental manipulation of a large set of linked variants in this haplotype. Functional genomics data, such as transcriptome or chromatin modification, can be employed to prioritize potentially causal elements and relevant tissue types. Eventually, emerging experimental techniques such as genome editing will enable a direct test of function in a relevant biological system.

REFERENCES

Adachi N, Shinoda Ki, et al. 2009. Mitochondrial DNA analysis of Jomon skeletons from the Funadomari site, Hokkaido, and its implication for the origins of Native American. *Am J Phys Anthropol* 138:255-265.

Aldenderfer M. 2006. Modelling plateau peoples: the early human use of the world's high plateaux. *World Archaeol* 38:357-370.

Aldenderfer M. 2011. Peopling the Tibetan plateau: insights from archaeology. *High Alt Med Biol* 12:141-147.

Aldenderfer M. 2013. Variation in mortuary practice on the early Tibetan plateau and the high Himalayas. *J Internat Assoc Bon Res* 1:293-318.

Aldenderfer M and Eng JT. 2016. Death and burial among two ancient high-altitude communities of Nepal. In: Schug GR and Walimbe SR, editors. *A Companion to South Asia in the Past*. Chichester, West Sussex, UK: John Wiley & Sons, Inc. p. 374-397.

Alexander DH, Novembre J, et al. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 19:1655-1664.

Alkorta-Aranburu G, Beall CM, et al. 2012. The genetic architecture of adaptations to high altitude in Ethiopia. *PLoS Genet* 8:e1003110.

Allentoft ME, Sikora M, et al. 2015. Population genomics of Bronze Age Eurasia. *Nature* 522:167-172.

Alt KW, Burger J, et al. 2003. Climbing into the past-first Himalayan mummies discovered in Nepal. *J Archaeol Sci* 30:1529-1535.

Altshuler DM, Gibbs RA, et al. 2010. Integrating common and rare genetic variation in diverse human populations. *Nature* 467:52-58.

Anderson E. 1949. *Introgressive hybridization*. New York, NY, USA: John Wiley and Sons, Inc.

Arnold ML. 2004. Transfer and origin of adaptations through natural hybridization: were Anderson and Stebbins right? *Plant Cell* 16:562-570.

Auwera GA, Carneiro MO, et al. 2013. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr Protoc Bioinformatics* 43:11.10.11-11.10.33.

Banskota K and Sharma B. 1995. *Tourism for mountain community development: Case Study report on the Annapurna and Gorkha regions of Nepal*. Kathmandu, Nepal: International Centre for Integrated Mountain Development.

Barton L. 2016. The cultural context of biological adaptation to high elevation Tibet. *Archaeol Res Asia* 5:4-11.

Beall C and Reichsman A. 1984. Hemoglobin levels in a Himalayan high altitude population. *Am J Phys Anthropol* 63:301-306.

Beall CM. 2006. Andean, Tibetan, and Ethiopian patterns of adaptation to high-altitude hypoxia. *Integr Comp Biol* 46:18-24.

Beall CM. 1982. A comparison of chest morphology in high altitude Asian and Andean populations. *Hum Biol* 54:145-163.

Beall CM. 2007. Two routes to functional adaptation: Tibetan and Andean high-altitude natives. *P Natl Acad Sci USA* 104:8655-8660.

Beall CM, Almasy LA, et al. 1999. Percent of oxygen saturation of arterial hemoglobin among Bolivian Aymara at 3,900-4,000 m. *Am J Phys Anthropol* 108:41-51.

Beall CM, Brittenham GM, et al. 1998. Hemoglobin concentration of high-altitude Tibetans and Bolivian Aymara. *Am J Phys Anthropol* 106:385-400.

Beall CM, Cavalleri GL, et al. 2010. Natural selection on *EPAS1* (*HIF2α*) associated with low hemoglobin concentration in Tibetan highlanders. *P Natl Acad Sci USA* 107:11459-11464.

Beall CM, Decker MJ, et al. 2002. An Ethiopian pattern of human adaptation to high-altitude hypoxia. *P Natl Acad Sci USA* 99:17215-17218.

Beall CM and Goldstein MC. 1987. Hemoglobin concentration of pastoral nomads permanently resident at 4,850–5,450 meters in Tibet. *Am J Phys Anthropol* 73:433-438.

Beall CM, Laskowski D, et al. 2001. Pulmonary nitric oxide in mountain dwellers. *Nature* 414:411-412.

Beall CM, Song K, et al. 2004. Higher offspring survival among Tibetan women with high oxygen saturation genotypes residing at 4,000 m. *P Natl Acad Sci USA* 101:14300-14304.

Beall CM, Strohl KP, et al. 1997. Ventilation and hypoxic ventilatory response of Tibetan and Aymara high altitude natives. *Am J Phys Anthropol* 104:427-447.

Befu H and Chard CS. 1964. A prehistoric maritime culture of the Okhotsk Sea. *Am Antiquity* 30:1-18.

Belwood P. 2014. *The Global Prehistory of Human Migration*. Chichester, West Sussex, UK: John Wiley & Sons.

Berg JJ and Coop G. 2014. A population genetic signal of polygenic adaptation. *PLoS Genet* 10:e1004412.

Bigham A, Bauchet M, et al. 2010. Identifying signatures of natural selection in Tibetan and Andean populations using dense genome scan data. *PLoS Genet* 6:e1001116.

Blanquart F, Kaltz O, et al. 2013. A practical guide to measuring local adaptation. *Ecol Lett* 16:1195-1205.

Brantingham PJ, Xing G, et al. 2013. Late Occupation of the High-Elevation Northern Tibetan Plateau Based on Cosmogenic, Luminescence, and Radiocarbon Ages. *Geoarchaeology* 28:413-431.

Briggs AW, Stenzel U, et al. 2007. Patterns of damage in genomic DNA sequences from a Neandertal. *P Natl Acad Sci USA* 104:14616-14621.

Bush WS and Moore JH. 2012. Genome-wide association studies. *PLoS Comput Biol* 8:e1002822.

Callaway E. 2014. Domestication: The birth of rice. *Nature* 514:S58-S59.

Chen F, Dong G, et al. 2015. Agriculture facilitated permanent human occupation of the Tibetan Plateau after 3600 BP. *Science* 347:248-250.

Chen H, Hey J, et al. 2015. A hidden Markov model for investigating recent positive selection through haplotype structure. *Theor Popul Biol* 99:18-30.

Chene P, Cechowska-Pasko M, et al. 2006. The effect of hypoxia on the expression of 150 kDa oxygen-regulated protein (ORP 150) in HeLa cells. *Cell Physiol Biochem* 17:89-96.

Chiaroni J, Underhill PA, et al. 2009. Y chromosome diversity, human expansion, drift, and cultural evolution. *P Natl Acad Sci USA* 106:20174-20179.

Childs G. 2012. Trans-Himalayan migrations as processes, not events: towards a theoretical framework. In: Huber T and Blackburn S, editors. *Origins and Migrations in the Extended Eastern Himalaya*. Leiden, The Netherlands: Koninklijke Brill NV. p. 11-32.

Chisholm B and Koike H. 1999. Reconstructing prehistoric Japanese diet using stable isotopic analysis. In: Omoto K, editor. *Interdisciplinary perspectives on the origins of the Japanese*. Kyoto, Japan: International Research Center for Japanese Studies. p. 199-222.

Cohen MN and Armelagos GJ. 1984. *Paleopathology at the origins of agriculture*. New York, NY, USA: Academic Press.

Crawford GW. 2006. East Asian plant domestication. In: Stark MT, editor. *Archaeology of Asia*. Oxford, UK: Blackwell Publishing Ltd. p. 77-95.

Daub JT, Hofer T, et al. 2013. Evidence for polygenic adaptation to pathogens in the human genome. *Mol Biol Evol* 30:1544-1558.

Delaneau O, Zagury J-F, et al. 2013. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods* 10:5-6.

Demeter F, Shackelford LL, et al. 2012. Anatomically modern human in Southeast Asia (Laos) by 46 ka. *P Natl Acad Sci USA* 109:14375-14380.

DePristo MA, Banks E, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43:491-498.

Dichgans M, Malik R, et al. 2014. Shared genetic susceptibility to ischemic stroke and coronary artery disease A genome-wide analysis of common variants. *Stroke* 45:24-36.

Dryden I. 2015. shapes: Statistical Shape Analysis. R package version 1.1-11. http://CRAN.R-project.org/package=shapes

Enattah NS, Jensen TG, et al. 2008. Independent introduction of two lactase-persistence alleles into human populations reflects different history of adaptation to milk culture. *Am J Hum Genet* 82:57-72.

Enattah NS, Sahi T, et al. 2002. Identification of a variant associated with adult-type hypolactasia. *Nat Genet* 30:233-237.

Epstein MP, Duren WL, et al. 2000. Improved inference of relationship for pairs of individuals. *Am J Hum Genet* 67:1219-1231.

Eshed V, Gopher A, et al. 2010. Paleopathology and the origin of agriculture in the Levant. *Am J Phys Anthropol* 143:121-133.

Fedorova SA, Reidla M, et al. 2013. Autosomal and uniparental portraits of the native populations of Sakha (Yakutia): implications for the peopling of Northeast Eurasia. *BMC Evol Biol* 13:127.

Field Y, Boyle EA, et al. 2016. Detection of human adaptation during the past 2,000 years. *bioRxiv* 052084; doi:http://dx.doi.org/10.1101/052084

Fornarino S, Pala M, et al. 2009. Mitochondrial and Y-chromosome diversity of the Tharus (Nepal): a reservoir of genetic variation. *BMC Evol Biol* 9:154.

Frichot E, Mathieu F, et al. 2014. Fast and efficient estimation of individual ancestry coefficients. *Genetics* 196:973-983.

Fu Q, Li H, et al. 2014. Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* 514:445-449.

Fu Q, Mittnik A, et al. 2013. A revised timescale for human evolution based on ancient mitochondrial genomes. *Curr Biol* 23:553-559.

Fuller DQ, Sato Y-I, et al. 2010. Consilience of genetics and archaeobotany in the entangled history of rice. *Archaeol Anthropol Sci* 2:115-131.

Gaillard C and Strauss Fo. 1990. Ethanol precipitation of DNA with linear polyacrylamide as carrier. *Nucleic Acids Res* 18:378.

Gamble C. 1994. *Timewalkers*. Cambridge, MA, USA: Harvard University Press.

Gavrilets S. 2003. Models of speciation: what have we learned in 40 years? *Evolution* 57:2197-2215.

Gayden T, Mirabal S, et al. 2009. Genetic insights into the origins of Tibeto-Burman populations in the Himalayas. *J Hum Genet* 54:216-223.

Gayden T, Perez A, et al. 2013. The Himalayas: Barrier and conduit for gene flow. *Am J Phys Anthropol* 151:169-182.

Gilbert-Kawai ET, Milledge JS, et al. 2014. King of the mountains: Tibetan and Sherpa physiological adaptations for life at high altitude. *Physiology* 29:388-402.

Ginolhac A, Rasmussen M, et al. 2011. mapDamage: testing for damage patterns in ancient DNA sequences. *Bioinformatics* 27:2153-2155.

Gleba M, Vanden Berghe I, et al. 2016. Textile technology in Nepal in the 5th-7th centuries CE: the case of Samdzong. *Sci Technol Archaeol Res* 2:1-11.

Global Lipids Genetics Consortium. 2013. Discovery and refinement of loci associated with lipid levels. *Nat Genet* 45:1274-1283.

Goebel T. 2002. The “microblade adaptation” and recolonization of Siberia during the late Upper Pleistocene. *Arch P Amer Ant Asso* 12:117-131.

Green RE, Krause J, et al. 2010. A draft sequence of the Neandertal genome. *Science* 328:710-722.

Gubin AN and Miller JL. 2001. Human erythroid porphobilinogen deaminase exists in 2 splice variants. *Blood* 97:815-817.

Guedes JdA, Lu H, et al. 2014. Moving agriculture onto the Tibetan Plateau: The archaeobotanical evidence. *Archaeol Anthropol Sci* 6:255-269.

Günther T, Valdiosera C, et al. 2015. Ancient genomes link early farmers from Atapuerca in Spain to modern-day Basques. *P Natl Acad Sci USA* 112:11917-11922.

Haak W, Lazaridis I, et al. 2015. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* 522:207-211.

Habu J. 2004. *Ancient Jomon of Japan*. Cambridge, UK: Cambridge University Press.

Hackinger S, Kraaijenbrink T, et al. 2016. Wide distribution and altitude correlation of an archaic high-altitude-adaptive *EPAS1* haplotype in the Himalayas. *Hum Genet* 135:393-402.

Hammer MF, Karafet TM, et al. 2006. Dual origins of the Japanese: common ground for hunter-gatherer and farmer Y chromosomes. *J Hum Genet* 51:47-58.

Hancock AM, Witonsky DB, et al. 2011. Adaptations to climate-mediated selective pressures in humans. *PLoS Genet* 7:e1001375.

Hanihara K. 1991. Dual structure model for the population history of the Japanese. *Japan Rev* 2:1-33.

Harris N. 2006. The elevation history of the Tibetan Plateau and its implications for the Asian monsoon. *Palaeogeogr Palaeoclimatol Palaeoecol* 241:4-15.

Hedrick PW. 2013. Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. *Mol Ecol* 22:4606-4618.

Hellenthal G, Busby GB, et al. 2014. A genetic atlas of human admixture history. *Science* 343:747-751.

Henn BM, Cavalli-Sforza LL, et al. 2012. The great human expansion. *P Natl Acad Sci USA* 109:17758-17764.

Hernandez RD, Kelley JL, et al. 2011. Classic selective sweeps were rare in recent human evolution. *Science* 331:920-924.

Hider JL, Gittelman RM, et al. 2013. Exploring signatures of positive selection in pigmentation candidate genes in populations of East Asian ancestry. *BMC Evol Biol* 13:150.

Houston J and Hartley AJ. 2003. The central Andean west-slope rainshadow and its potential contribution to the origin of hyper-aridity in the Atacama Desert. *Int J Climatol* 23:1453-1464.

Howells WW. 1997. *Getting here: The story of human evolution*. Washington, DC, USA: Compass Press.

Howie BN, Donnelly P, et al. 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* 5:e1000529.

Hublin J-J. 2012. The earliest modern human colonization of Europe. *P Natl Acad Sci USA* 109:13471-13472.

Huerta-Sánchez E, DeGiorgio M, et al. 2013. Genetic signatures reveal high-altitude adaptation in a set of Ethiopian populations. *Mol Biol Evol* 30:1877-1888.

Huerta-Sánchez E, Jin X, et al. 2014. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* 512:194-197.

Huey RB, Gilchrist GW, et al. 2000. Rapid evolution of a geographic cline in size in an introduced fly. *Science* 287:308-309.

Hüttel H. 1997. Archaologische Siedlungsforschung im Hohen Himalaya: Die Ausgrabungen der KAVA im Muktinath-Tal/Nepal 1994-1995. *Beitr Allgem Vergleich Archäol* 17:7-64.

Imo M, Maixner M, et al. 2013. Sympatric diversification vs. immigration: deciphering host-plant specialization in a polyphagous insect, the stolbur phytoplasma vector *Hyalesthes obsoletus* (Cixiidae). *Mol Ecol* 22:2188-2203.

Ingram CJ, Elamin MF, et al. 2007. A novel polymorphism associated with lactose tolerance in Africa: multiple causes for lactase persistence? *Hum Genet* 120:779-788.

Jablonski NG and Chaplin G. 2000. The evolution of human skin coloration. *J Hum Evol* 39:57-106.

Jablonski NG and Chaplin G. 2010. Human skin pigmentation as an adaptation to UV radiation. *P Natl Acad Sci USA* 107:8962-8968.

Jensen JD, Foll M, et al. 2016. The past, present and future of genomic scans for selection. *Mol Ecol* 25:1-4.

Jeong C, Alkorta-Aranburu G, et al. 2014. Admixture facilitates genetic adaptations to high altitude in Tibet. *Nat Commun* 5:3281.

Jinam T, Nishida N, et al. 2012. The history of human populations in the Japanese Archipelago inferred from genome-wide SNP data with a special reference to the Ainu and the Ryukyuan populations. *J Hum Genet* 57:787-795.

Jinam TA, Kanzawa-Kiriyama H, et al. 2015. Unique characteristics of the Ainu population in Northern Japan. *J Hum Genet* 60:565-571.

Jónsson H, Ginolhac A, et al. 2013. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* 29:1682-1684.

Kaelin Jr WG and Ratcliffe PJ. 2008. Oxygen sensing by metazoans: the central role of the HIF hydroxylase pathway. *Mol Cell* 30:393-402.

Kamberov YG, Wang S, et al. 2013. Modeling recent human evolution in mice by expression of a selected *EDAR* variant. *Cell* 152:691-702.

Kawecki TJ and Ebert D. 2004. Conceptual issues in local adaptation. *Ecol Lett* 7:1225-1241.

Keller A, Graefen A, et al. 2012. New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. *Nat Commun* 3:698.

Kimura R, Yamaguchi T, et al. 2009. A common variation in *EDAR* is a genetic determinant of shovel-shaped incisors. *Am J Hum Genet* 85:528-535.

Kircher M. 2012. Analysis of high-throughput ancient DNA sequencing data. In: Shapiro B and Hofreiter M, editors. *Ancient DNA: methods and protocols*. New York, NY, USA: Humana Press. p. 197-228.

Kloss-Brandstätter A, Pacher D, et al. 2011. HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. *Hum Mutat* 32:25-32.

Knörzer K-H. 2000. 3000 years of agriculture in a valley of the High Himalayas. *Veg Hist Archaeobot* 9:219-222.

Koganebuchi K, Katsumura T, et al. 2012. Autosomal and Y-chromosomal STR markers reveal a close relationship between Hokkaido Ainu and Ryukyu islanders. *Anthropol Sci* 120:199-208.

Krüttli A, Bouwman A, et al. 2014. Ancient DNA analysis reveals high frequency of European lactase persistence allele (T-13910) in medieval central europe. *PLoS One* 9:e86251.

LaPolla RJ. 2001. The role of migration and language contact in the development of the Sino-Tibetan language family. In: Aikhenvald AY and Dixon R, editors. *Areal diffusion and genetic inheritance: problems in comparative linguistics*. Oxford, UK: Oxford University Press. p. 225-254.

Lazaridis I, Patterson N, et al. 2014. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* 513:409-413.

León-Velarde F, Rivera-Ch M, et al. 2014. Chronic Mountain Sickness. In. *High Altitude: Human Adaptation to Hypoxia*. New York, NY, USA: Springer. p. 429-447.

Lettre G, Lange C, et al. 2007. Genetic model testing and statistical power in population-based association studies of quantitative traits. *Genet Epidemiol* 31:358-362.

Levene H. 1953. Genetic equilibrium when more than one ecological niche is available. *Am Nat* 87:331-333.

Lewontin R and Birch L. 1966. Hybridization as a source of variation for adaptation to new environments. *Evolution* 20:315-336.

Li H and Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25:1754-1760.

Li H and Durbin R. 2011. Inference of human population history from individual whole-genome sequences. *Nature* 475:493-496.

Li H, Gu S, et al. 2011. Diversification of the *ADH1B* gene during expansion of modern humans. *Ann Hum Genet* 75:497-507.

Li H, Handsaker B, et al. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* 25:2078-2079.

Li JZ, Absher DM, et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319:1100-1104.

Lipson M, Loh P-R, et al. 2013. Efficient moment-based inference of admixture parameters and sources of gene flow. *Mol Biol Evol* 30:1788-1802.

Loh P-R, Lipson M, et al. 2013. Inferring admixture histories of human populations using linkage disequilibrium. *Genetics* 193:1233-1254.

Lorenzo FR, Huff C, et al. 2014. A genetic mechanism for Tibetan high-altitude adaptation. *Nat Genet* 46:951-956.

Majmundar AJ, Wong WJ, et al. 2010. Hypoxia-inducible factors and the response to hypoxic stress. *Mol Cell* 40:294-309.

Marchini J, Cardon LR, et al. 2004. The effects of human population structure on large genetic association studies. *Nat Genet* 36:512-517.

Massa G. 2013. The funerary metals of Samdzong. MSc thesis (London: University College London).

Matsukusa H, Oota H, et al. 2010. A genetic analysis of the Sakishima islanders reveals no relationship with Taiwan aborigines but shared ancestry with Ainu and main-island Japanese. *Am J Phys Anthropol* 142:211-223.

Matthews L, Gopinath G, et al. 2009. Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res* 37:D619-D622.

McKenna A, Hanna M, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20:1297-1303.

Metspalu M, Kivisild T, et al. 2004. Most of the extant mtDNA boundaries in south and southwest Asia were likely shaped during the initial settlement of Eurasia by anatomically modern humans. *BMC Genet* 5:26.

Meyer M, Kircher M, et al. 2012. A high-coverage genome sequence from an archaic Denisovan individual. *Science* 338:222-226.

Mira A, Pushker R, et al. 2006. The Neolithic revolution of bacterial genomes. *Trends Microbiol* 14:200-206.

Miyata T, Hayashida H, et al. 1987. Male-driven molecular evolution: a model and nucleotide sequence analysis. *Cold Spring Harb Symp Quant Biol* 52:863-867.

Molina J, Sikora M, et al. 2011. Molecular evidence for a single evolutionary origin of domesticated rice. *P Natl Acad Sci USA* 108:8351-8356.

Moorjani P, Thangaraj K, et al. 2013. Genetic evidence for recent population mixture in India. *Am J Hum Genet* 93:422-438.

Mummert A, Esche E, et al. 2011. Stature and robusticity during the agricultural transition: evidence from the bioarchaeological record. *Econ Hum Biol* 9:284-301.

Nakagome S, Sato T, et al. 2015. Model-based verification of hypotheses on the origin of modern Japanese revisited by Bayesian inference based on genome-wide SNP data. *Mol Biol Evol* 32:1533-1543.

Niermeyer S. 2014. Reproduction and Growth. In. *High Altitude: Human Adaptation to Hypoxia*. New York, NY, USA: Springer. p. 341-355.

Novembre J, Johnson T, et al. 2008. Genes mirror geography within Europe. *Nature* 456:98-101.

Ohyi H. 1975. The Okhotsk culture, a maritime culture of the southern Okhotsk Sea region. In: Fitzhugh W, editor. *Prehistoric Maritime Adaptations of the Circumpolar Zone*. Chicago, IL, USA: Aldine Publishing Company. p. 123-158.

Oota H and Stoneking M. 2011. Effect of human migration on genome diversity in East Asia. In. *Racial representations in Asia*. Kyoto, Japan: Kyoto University Press. p. 173-187.

Oppitz M. 1974. Myths and facts: Reconsidering some data concerning the clan history of the Sherpas. *Kailash* 2:121-131.

Osborn A. 2014. Eye of the needle: cold stress, clothing, and sewing technology during the Younger Dryas cold event in North America. *Am Antiquity* 79:45-68.

Owsley DW and Jantz RL. 2014. *Kennewick Man: The Scientific Investigation of an Ancient American Skeleton*. College Station, TX, USA: Texas A&M University Press.

Paradis E, Claude J, et al. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20:289-290.

Patterson N, Moorjani P, et al. 2012. Ancient admixture in human history. *Genetics* 192:1065-1093.

Patterson N, Price AL, et al. 2006. Population structure and eigenanalysis. *PLoS Genet* 2:e190.

Peng M-S, Palanichamy GM, et al. 2011. Inland post-glacial dispersal in East Asia revealed by mitochondrial haplogroup M9a'b. *BMC biology* 9:1.

Peng Y, Shi H, et al. 2010. The ADH1B Arg47His polymorphism in East Asian populations and expansion of rice domestication in history. *BMC Evol Biol* 10:15.

Peng Y, Yang Z, et al. 2011. Genetic variations in Tibetan populations and high-altitude adaptation at the Himalayas. *Mol Biol Evol* 28:1075-1081.

Petkova D, Novembre J, et al. 2016. Visualizing spatial population structure with estimated effective migration surfaces. *Nat Genet* 48:94-100.

Petousi N, Croft QPP, et al. 2014. Tibetans living at sea level have a hyporesponsive hypoxia-inducible factor system and blunted physiological responses to hypoxia. *J Appl Physiol* 116:893-904.

Pickrell JK, Coop G, et al. 2009. Signals of recent positive selection in a worldwide sample of human populations. *Genome Res* 19:826-837.

Pickrell JK, Patterson N, et al. 2012. The genetic prehistory of southern Africa. *Nat Commun* 3:1143.

Pickrell JK and Pritchard JK. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet* 8:e1002967.

Pinhasi R, Thomas MG, et al. 2012. The genetic history of Europeans. *Trends Genet* 28:496-505.

Pitulko VV, Nikolsky PA, et al. 2004. The Yana RHS site: humans in the Arctic before the last glacial maximum. *Science* 303:52-56.

Pitulko VV, Tikhonov AN, et al. 2016. Early human presence in the Arctic: Evidence from 45,000-year-old mammoth remains. *Science* 351:260-263.

Price AL, Patterson NJ, et al. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38:904-909.

Price AL, Tandon A, et al. 2009. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet* 5:e1000519.

Pritchard JK and Di Rienzo A. 2010. Adaptation–not by sweeps alone. *Nat Rev Genet* 11:665-667.

Pritchard JK, Pickrell JK, et al. 2010. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr Biol* 20:R208-R215.

Prüfer K, Racimo F, et al. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* 505:43-49.

Purcell S, Neale B, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81:559-575.

Qi X, Cui C, et al. 2013. Genetic evidence of Paleolithic colonization and Neolithic expansion of modern humans on the Tibetan Plateau. *Mol Biol Evol* 30:1761-1778.

Qiu Q, Wang L, et al. 2015. Yak whole-genome resequencing reveals domestication signatures and prehistoric population expansions. *Nat Commun* 6:10283.

Qiu Q, Zhang G, et al. 2012. The yak genome and adaptation to life at high altitude. *Nat Genet* 44:946-949.

Raghavan M, Skoglund P, et al. 2014. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature* 505:87-91.

Ralf A, Oven M, et al. 2015. Simultaneous analysis of hundreds of Y-chromosomal SNPs for high-resolution paternal lineage classification using targeted semiconductor sequencing. *Hum Mutat* 36:151-159.

Rasmussen M, Guo X, et al. 2011. An Aboriginal Australian genome reveals separate human dispersals into Asia. *Science* 334:94-98.

Rasmussen M, Li Y, et al. 2010. Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* 463:757-762.

Rasmussen M, Sikora M, et al. 2015. The ancestry and affiliations of Kennewick Man. *Nature* 523:455-458.

Reich D, Green RE, et al. 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* 468:1053-1060.

Reich D, Patterson N, et al. 2011. Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. *Am J Hum Genet* 89:516-528.

Reich D, Thangaraj K, et al. 2009. Reconstructing Indian population history. *Nature* 461:489-494.

Rosenberg NA, Pritchard JK, et al. 2002. Genetic structure of human populations. *Science* 298:2381-2385.

R Core Team. 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Sabeti PC, Varilly P, et al. 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature* 449:913-918.

Sanson M, Ingueneau C, et al. 2008. Oxygen-regulated protein-150 prevents calcium homeostasis deregulation and apoptosis induced by oxidized LDL in vascular cells. *Cell Death Differ* 15:1255-1265.

Savolainen O, Lascoux M, et al. 2013. Ecological genomics of local adaptation. *Nat Rev Genet* 14:807-820.

Scally A and Durbin R. 2012. Revising the human mutation rate: implications for understanding human evolution. *Nat Rev Genet* 13:745-753.

Schadt EE, Molony C, et al. 2008. Mapping the genetic architecture of gene expression in human liver. *PLoS Biol* 6:e107.

Scheet P and Stephens M. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* 78:629-644.

Scheinfeldt LB, Soi S, et al. 2012. Genetic adaptation to high altitude in the Ethiopian highlands. *Genome Biol* 13:R1.

Schiffels S and Durbin R. 2014. Inferring human population size and separation history from multiple genome sequences. *Nat Genet* 46:919-925.

Semenza GL. 2012. Hypoxia-inducible factors in physiology and medicine. *Cell* 148:399-408.

Shi H, Zhong H, et al. 2008. Y chromosome evidence of earliest modern human settlement in East Asia and multiple origins of Tibetan and Japanese populations. *BMC biol* 6:45.

Simons A and Schön W. 1998. Cave systems and terrace settlements in Mustang, Nepal. Settlement periods from prehistoric times to the present day. *Beitr Allgem Vergleich Archäol* 18:27-47.

Simons A, Schön W, et al. 1998. Archaeological research in Mustang: Report on the field work of the years 1994 and 1995 done by the Cologne University Team. *Ancient Nepal* 140:65-83.

Simonson TS, Yang Y, et al. 2010. Genetic evidence for high-altitude adaptation in Tibet. *Science* 329:72-75.

Singh A. 1999. Cist Burials in Kinnaur, western Himalayas: A preliminary report on recent discovery. *Cent Asiatic J* 43:249-258.

Skoglund P, Mallick S, et al. 2015. Genetic evidence for two founding populations of the Americas. *Nature* 525:104-108.

Skoglund P, Malmström H, et al. 2014. Genomic diversity and admixture differs for Stone-Age Scandinavian foragers and farmers. *Science* 344:747-750.

Skoglund P, Malmström H, et al. 2012. Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. *Science* 336:466-469.

Skoglund P, Storå J, et al. 2013. Accurate sex identification of ancient human remains using DNA shotgun sequencing. *J Archaeol Sci* 40:4477-4482.

Skotte L, Korneliussen TS, et al. 2013. Estimating individual admixture proportions from next generation sequencing data. *Genetics* 195:693-702.

Stacul G. 1968. Excavation near Ghaligai and chronological sequence of protohistorical cultures in the Swat Valley. *East and West* 19:44-91.

Stebbins GL. 1959. The role of hybridization in evolution. *Proceedings of the American Philosophical Society* 103:231-251.

Stinson S. 1985. Chest dimensions of European and Aymara children at high altitude. *Ann Hum Biol* 12:333-338.

Stoneking M and Delfin F. 2010. The human genetic history of East Asia: weaving a complex tapestry. *Curr Biol* 20:R188-R193.

Storz JF, Scott GR, et al. 2010. Phenotypic plasticity and genetic adaptation to high-altitude hypoxia in vertebrates. *J Exp Biol* 213:4125-4136.

Summerhayes GR, Leavesley M, et al. 2010. Human adaptation and plant use in highland New Guinea 49,000 to 44,000 years ago. *Science* 330:78-81.

Sun Y-j, Fang M-w, et al. 2010. Endothelial nitric oxide synthase gene polymorphisms associated with susceptibility to high altitude pulmonary edema in Chinese railway construction workers at Qinghai-Tibet over 4 500 meters above sea level. *Chin Med Sci J* 25:215-221.

Swallow DM. 2003. Genetics of lactase persistence and lactose intolerance. *Annu Rev Genet* 37:197-219.

Tajima A, Pan I-H, et al. 2002. Three major lineages of Asian Y chromosomes: implications for the peopling of east and southeast Asia. *Hum Genet* 110:80-88.

Taylor CT and McElwain JC. 2010. Ancient atmospheres and the evolution of oxygen sensing via the hypoxia-inducible factor in metazoans. *Physiology* 25:272-279.

Tenesa A, Navarro P, et al. 2007. Recent human effective population size estimated from linkage disequilibrium. *Genome Res* 17:520-526.

Thangaraj K, Chaubey G, et al. 2005. Reconstructing the origin of Andaman Islanders. *Science* 308:996-996.

The HUGO Pan-Asian SNP Consortium. 2009. Mapping human genetic diversity in Asia. *Science* 326:1541-1545.

The 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491:56-65.

Tishkoff SA, Reed FA, et al. 2009. The genetic structure and history of Africans and African Americans. *Science* 324:1035-1044.

Tishkoff SA, Reed FA, et al. 2007. Convergent adaptation of human lactase persistence in Africa and Europe. *Nat Genet* 39:31-40.

Tito RY, Belknap SL, et al. 2011. DNA from early Holocene American dog. *Am J Phys Anthropol* 145:653-657.

Tiwari DN. 1985. Cave Burials From Western Nepal, Mustang. *Ancient Nepal* 85:1-21.

Tokunaga K, Ohashi J, et al. 2001. Genetic link between Asians and native Americans: evidence from HLA genes and haplotypes. *Hum Immunol* 62:1001-1008.

Van Oven M and Kayser M. 2009. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum Mutat* 30:E386-E394.

Voight BF, Kudaravalli S, et al. 2006. A map of recent positive selection in the human genome. *PLoS Biol* 4:e72.

Volodko NV, Starikovskaya EB, et al. 2008. Mitochondrial genome diversity in arctic Siberians, with particular reference to the evolutionary history of Beringia and Pleistocenic peopling of the Americas. *Am J Hum Genet* 82:1084-1100.

Wang B, Zhang Y-B, et al. 2011. On the origin of Tibetans and their genetic basis in adapting high-altitude environments. *PLoS One* 6:e17002.

Wang C, Zöllner S, et al. 2012. A quantitative comparison of the similarity between genes and geography in worldwide human populations. *PLoS Genet* 8:e1002886.

Weiner JS and Lourie JA. 1969. *Human Biology, A Guide to Field Methods*. Oxford, UK: Blackwell Scientific Publications.

Weir BS and Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. *Evolution* 38:1358-1370.

Welter D, MacArthur J, et al. 2014. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 42:D1001-D1006.

Wen B, Li H, et al. 2004. Genetic evidence supports demic diffusion of Han culture. *Nature* 431:302-305.

Wright KM, Lloyd D, et al. 2013. Indirect evolution of hybrid lethality due to linkage with selected locus in *Mimulus guttatus*. *PLoS Biol* 11:e1001497.

Wright S. 1931. Evolution in Mendelian populations. *Genetics* 16:97-159.

Wu H, Caffo B, et al. 2010. Redefining CpG islands using hidden Markov models. *Biostatistics* 11:499-514.

Xiang K, Peng Y, et al. 2013. Identification of a Tibetan-specific mutation in the hypoxic gene *EGLN1* and its contribution to high-altitude adaptation. *Mol Biol Evol* 30:1889-1898.

Xing J, Watkins WS, et al. 2010. Toward a more uniform sampling of human genetic diversity: a survey of worldwide populations by high-density genotyping. *Genomics* 96:199-210.

Xu S, Li S, et al. 2011. A genome-wide search for signals of high-altitude adaptation in Tibetans. *Mol Biol Evol* 28:1003-1011.

Yang X, Wan Z, et al. 2012. Early millet use in northern China. *P Natl Acad Sci USA* 109:3726-3730.

Yi X, Liang Y, et al. 2010. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* 329:75-78.

Yoneda M, Tanaka A, et al. 2002. Radiocarbon marine reservoir effect in human remains from the Kitakogane site, Hokkaido, Japan. *J Archaeol Sci* 29:529-536.

Yoon D, Ponka P, et al. 2011. Hypoxia. 5. Hypoxia and hematopoiesis. *Am J Physiol-Cell Ph* 300:C1215-C1222.

Zeller T, Wild P, et al. 2010. Genetics and beyond–the transcriptome of human monocytes and disease susceptibility. *PLoS One* 5:e10693.

Zhang DD and Li S. 2002. Optical dating of Tibetan human hand-and footprints: An implication for the palaeoenvironment of the last glaciation of the Tibetan Plateau. *Geophys Res Lett* 29:16-11-16-13.

Zhao M, Kong Q-P, et al. 2009. Mitochondrial genome evidence reveals successful Late Paleolithic settlement on the Tibetan Plateau. *P Natl Acad Sci USA* 106:21230-21235.

Zhou X and Stephens M. 2012. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* 44:821-824.

Ziesemer KA, Mann AE, et al. 2015. Intrinsic challenges in ancient microbiome reconstruction using 16S rRNA gene amplification. *Sci Rep* 5:16498.