Evolutionary Studies

Naruya Saitou *Editor*

Evolution of the Human Genome II

Human Evolution Viewed from Genomes

Springer

Evolutionary Studies

Series Editor

Naruya Saitou, National Institute of Genetics, Mishima, Japan

Everything is history, starting from the Big Bang or the origin of the universe to the present time. This historical nature of the universe is clear if we look at evolution of organisms. Evolution is one of most basic features of life which appeared on Earth more than 3.7 billion years ago. Considering the importance of evolution in biology, we are inaugurating this series. Any aspect of evolutionary studies on any kind of organism is a potential target of the series. Life started at the molecular level, thus molecular evolution is one important area in the series, but non-molecular studies are also within its scope, especially those studies on evolution of multicellular organisms. Evolutionary phenomena covered by the series include the origin of life, fossils in general, Earth–life interaction, evolution of prokaryotes and eukaryotes, viral and protist evolution, the emergence of multicellular organisms, phenotypic and genomic diversity of certain organism groups, and more. Theoretical studies on evolution are also covered within the spectrum of this new series.

More information about this series at http://www.springer.com/series/15220

Naruya Saitou
Editor

Evolution of the Human Genome II

Human Evolution Viewed from Genomes

Springer

Editor
Naruya Saitou
Department of Genomics and Evolutionary
Biology
National Institute of Genetics
Mishima, Shizuoka, Japan

ISSN 2509-484X ISSN 2509-4858 (electronic)
Evolutionary Studies
ISBN 978-4-431-56902-2 ISBN 978-4-431-56904-6 (eBook)
https://doi.org/10.1007/978-4-431-56904-6

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Japan KK part of Springer Nature.
The registered company address is: Shiroyama Trust Tower, 4-3-1 Toranomon, Minato-ku, Tokyo 105-6005, Japan

Preface

This book is the second volume of the *Evolution of the Human Genome* and is the third book of the Evolutionary Studies Springer Series. I am the series editor and the editor of this book. As I already wrote in the Preface of the first volume, published in 2017, its plan started in 2011. The subtitle of the first volume is "The Genome and Genes," and the subtitle of this volume is "Modern Human Dispersal." Although most of the manuscript for the second volume was completed by the end of 2015, the publication was delayed until 2021. I originally planned to contribute one or two chapters for the second volume; however, I could not make it during the five years. I thus abandoned my contribution and decided to publish the completed 12 chapters. Some authors tried to decline to publish their five-year-old manuscripts. I persuaded them to publish their manuscripts as they are. Any book or paper starts to become outdated immediately after its publication, but some of them may become "classic." I am sure that all the 12 chapters of this book stand during the five years and hope that many readers will appreciate them.

This volume consists of two parts: Part I is "Non-neutral Evolution on Human Genes," which includes three chapters. Chapter 1, titled "Anthropogeny," was written by Pascal Gagneux, and he reviewed various facets of human evolution study. Chapter 2 by Pierre Luisi et al. is about methods to detect natural selection on protein-coding genes. Genes related to high altitude adaptation are reviewed by Lian Deng and Shuhua Xu in Chap. 3. Readers can grasp various facets of the adaptation process caused by positive selection.

Part II of this volume includes nine chapters under the title "Evolution of Modern Human Populations." Jun Gojobori reviewed the research history of mitochondrial DNA diversity of human populations in Chap. 4. Mitochondrial DNA is maternally inherited, while Y chromosome is paternally inherited. The research history of this chromosome is discussed by Francesc Calafell and David Comas in Chap. 5. Researches on human DNA diversity for six geographical areas are discussed in Chaps. 6–12. These six areas and authors of the corresponding chapters are: Africa (David Comas and Francesc Carlafell) in Chap. 6, West and South Asia (Analabha Basu and Partha Majumder) in Chap. 7, Europe (Jaume Bertranpetit and Guido

Babruyani) in Chap. 8, Southeast Asia (Timothy A. Jinam) in Chap. 9, Oceania (Ana Duggan and Mark Stoneking) in Chap. 10, and America (Inaho Danjo, Hideaki Kanzawa-Kiriyama, and Naruya Saitou) in Chap. 11. Finally, Mathias Currat, Claudio S. Quilodrán, and Laurent Excoffier explained simulation studies on human dispersal in Chap. 12.

I would like to mention again that most of the chapters were written by 2015, and recent developments after 2015 are not included in these chapters. However, I hope this volume is still useful for many readers who are interested in a long history of DNA study of modern human populations.

Mishima, Japan Naruya Saitou

Contents

Part I Non-neutral Evolution on Human Genes

Part II Evolution of Modern Human Populations

Part I
Non-neutral Evolution on Human Genes

Chapter 1
Anthropogeny

Pascal Gagneux

Abstract Anthropogeny, "the study of the origin of humans" is an attempt to use all verifiable facts and ethical scientific methods to explain the origin of the species *Homo sapiens*. Only a transdisciplinary approach will allow to unravel the singularity that is the appearance of our species, the "planet-altering ape." Such transdisciplinarity will have to involve fields as varied as linguistics and psychology, biomedicine and neuroscience, physical and chemical sciences, comparative primatology, climate sciences and geology, archeology and paleontology with much support from computer science. Humans present a striking paradox as they combine an obvious mammalian and primate nature with a distinct combination of numerous biological and behavioral traits, making them spectacular outlier among the living world. The time depth of many of the processes that shaped our species represents a formidable obstacle. New fossils, archeological finds, ancient DNA technology, and comparative genomics are providing key new information. Anthropogenists are still facing a staggering list of humbling unknowns about the age of onset of key human innovations. These include but are not restricted to the following: symbolic capacity, personal name or kinship terms, language, home base use, fire use/cooking, pair bonding, awareness of paternal kinship networks, projectile weapon use, composite tool use, fiber use, bodily modifications, and death rituals. The human phenomenon reflects idiosyncratic concatenations of unlikely events. Key factors likely include both opportunities and constraints stemming from massive physical and cultural niche construction by our species that has increasingly taken its evolutionary fate in its own hands.

Keywords Human origins · Human evolution · Hominid · Hominin · Anthropogeny · Great ape · Niche construction · Theory of mind · Self-domestication

P. Gagneux (✉)
Department of Pathology and Anthropology, University of California San Diego, La Jolla, CA, USA
e-mail: pgagneux@health.ucsd.edu

N. Saitou (ed.), *Evolution of the Human Genome II*, Evolutionary Studies,
https://doi.org/10.1007/978-4-431-56904-6_1

1.1 Getting at the Origins of the Human Phenomenon

Questions about origins feature prominently in cosmologies of most human cultures. The study of the origin of our species or Anthropogeny has long fascinated philosophers and scientists.

***Anthropogeny** The investigation of the origin of man (humans) Oxford English Dictionary, 2006. First used in 1839 edition of Hooper's Med. Dict. and defined as "the study of the generation of man."

"Where do we come from?" and "How did we get here?" are the two questions driving anthropogeny. We have never been in a better position to attempt answering these questions. As we are approaching a clearer view of the timing and the location of our origin, we are still far from understanding the evolutionary singularity represented by the emergence of our question-asking species.

Huxley and Darwin initially predicted (Huxley 1863; Darwin 1871) that modern humans likely shared a last common ancestor with apes in Africa. Since then steady accumulation of hominid fossils in Africa, Asia, and Europe, combined with a wealth of molecular data now provides overwhelming evidence for a deeply rooted origin of our lineage in Africa. How did small Miocene apes evolve into bipedal, small-brained Pliocene hominids and then into tall, stone tool-using, running, and fire-controlling, ever larger-brained members of our genus *Homo*? A combination of factors, spanning the molecular, cellular, microbial, social, cultural, ecological, and climatic must have contributed to the peculiar trajectory of the hominin lineage. Modern Anthropogeny is focused on the circumstances and events that led to the appearance of our species between 200 and 100 kya in Africa and eventually followed by the almost complete replacement of all other hominin species in Europe and Asia.

To the best of our knowledge, the human species represents an evolutionary singularity. The study of human origins thus represents a historical exercise with a sample size of $n = 1$. As such, anthropogeny is an exercise in deep history (Smail 2008). For singular phenomena, one cannot rule out extremely rare or unusual events as causal factors. Getting at answers will require drawing on a large number of disciplines ranging from the molecular to the social and geophysical and benefitting most from comparative approaches (Fig. 1.1).

1.2 Our Evolutionary Roots

Humans are firmly rooted in the tree of life and share this planet with several closely related extant primate species. In stark contrast to their living relatives, humans can be characterized as a highly successful "weed species," having colonized the entire planet, replaced closely related species, and caused mass-extinctions everywhere we went (Diamond 1989). A combination of uniquely derived socio-cognitive

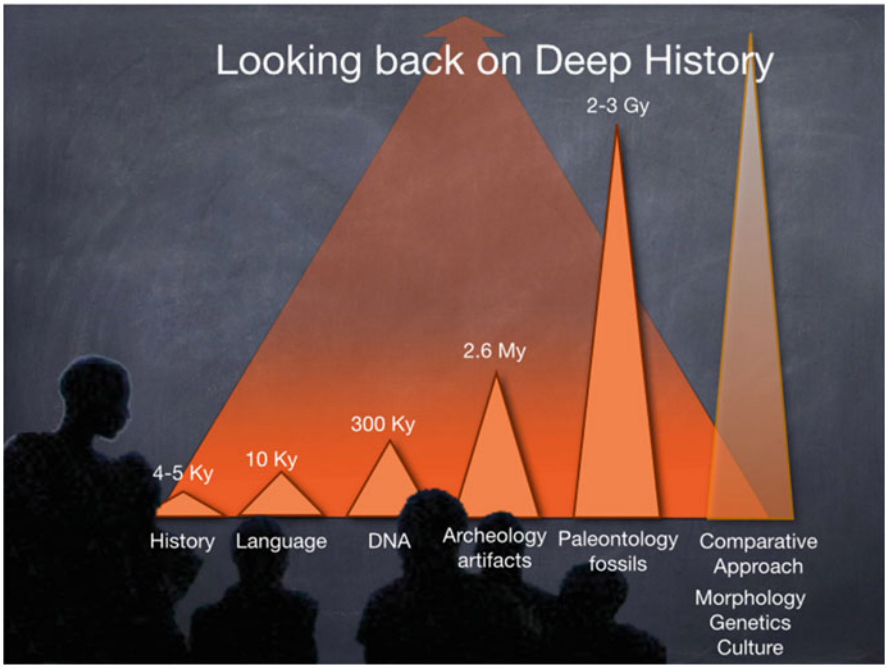


Fig. 1.1 The singularity of human evolution means that anthropogeny is first and foremost a historical enterprise. Different methodologies provide different time depth ranging from the <5 thousand years of written historical records to the billion year old fossil record. Comparative studies of genomes, phenotypes, and behavior represent our best chance at reconstructing the human story. Methods for deep history

adaptations, language, and technology catalyzed the powerful niche construction ability of our species and has directly contributed to this planetary take-over, also resulting in the endangered species status of all remaining non-human hominids (the "great apes") (Kondgen et al. 2008). Several other primate species have evolved remarkably flexible ecologies, including baboons in Africa and macaques in Asia, but all of these continue to coexist as multiple closely related species (Winder 2014; Morales and Melnick 1998).

1.2.1 Homo Sapiens: *The Paradoxical Ape*

Complete genome data from multiple individuals of all extant ape species clearly indicate that two non-human primate species are more closely related to humans than either is to any other extant primate species. Traditionally, the notion of great apes (pongid) vs humans was based on skeletal anatomy. The fact that humans share a common ancestor with bonobos (*Pan paniscus*) and chimpanzees (*Pan troglodytes*) after the divergence of the lineage leading to gorillas (*Gorilla, gorilla*) nullifies the biological validity of the term “great apes/pongids” (Prado-Martinez et al. 2013). Despite its biological fallacy, the term “great ape” continues to be used and is rather useful when discussing the human phenomenon, given the many ways in which human biology and behavior have come to diverge from that of other hominids (including all living great apes). The close phylogenetic proximity of humans and the two species of Pan have even been used to argue that all three species should share a genus (Goodman et al. 1989, 1998). Soft tissue anatomy also groups humans and chimpanzees into a monophyletic group excluding gorillas and orangutans (Gibbs et al. 2000). Conversely, the long list of human-unique specializations from cell biology to cognition to social structure makes it unlikely that our species will be renamed “*Pan sapiens*” or that the two chimpanzees will be renamed “*Homo troglodytes*” and “*Homo paniscus*” any time soon. Humans can be safely considered to simultaneously be genetic apes and ecological “ex-apes” (Marks 2012).

1.2.2 Measuring Genetic Distance

How can we best express genetic or genomic similarity and how are we to interpret the meaning of such figures? Despite the linear nature of DNA sequences, genomes are far from linear and thus linear comparisons in % genetic difference have serious limitations. The initial DNA hybridization experiments (Sibley and Ahlquist 1987) excluded most heterochromatin. Taking into account the entirety of the genetic material there is closer to 5% total difference between human and the closest extant non-human genome (Britten 2002; Mikkelsen et al. 2005). Much of the genetic variation consists of structural variation including changes in cytogenetic organization, segmental duplications, and lineage-specific expansion and/or deletions (Gazave et al. 2011). The exact changes that make us human remain painfully elusive. They likely include: point mutations and positive selection in structural and regulatory regions; gains of function: via recently duplicated or partially duplicated genes, change in gene copy numbers, de novo genes, accelerated regions including RNA genes; losses of function: including deletions or lost expression of otherwise conserved mammalian genes. Changes in expression and splicing: including both transcription levels and locations. Rapid transcription factor evolution: by segmental duplication or positive selection. Also relevant are changes in transposable elements: their type, abundance, activity, suppression, and locations. Prime candidates for the genetic basis of humanness are alterations to gene expression networks in the brain and factors affecting growth rate and life history timing. However, testing human-specific genetic changes for their phenotypic effects remains far from trivial, and understanding their adaptive importance in the face of natural, sexual, and social selection is more difficult still. The quest for the genetic bases of humanness remains a fantastic challenge (Varki and Altheide 2005; O’Bleness et al. 2012). For an attempt at ongoing enumeration of such traits, please see the Matrix of Anthropogeny website of the Center for Academic Research and

Training in Anthropogeny (CARTA) (http://carta.anthropogeny.org/content/about-moca).

The fact that a few nucleotide changes at important functional sites of the genome, e.g. promoter region or transcription factor binding sites, can have drastic effects on development and phenotype has long prompted the hypothesis that relatively few regulatory changes would explain the drastic phenotypic differences between humans and apes (King and Wilson 1975). Comparative genomics have revealed ~2000 human accelerated regions (HARs) that seem enriched for functional elements such as enhancers (Pollard et al. 2006; Capra et al. 2013) and conserved regions uniquely deleted in humans (McLean et al. 2011; Lindblad-Toh et al. 2011). Similarly, human-specific duplication (HSDs) include a number of genes involved in neuronal proliferation, migration, and maturation (Nuttle et al. 2013). Of course, an important limitation remains in the uncertainty regarding the precise number and identity of functional elements and their interactions in the mammalian genome (The Encode Consortium 2012). Furthermore, given that even slight changes in the genome can profoundly affect function and with it the development of phenotypic traits, there is an obvious need for functional studies, which will mostly be limited to in vitro assays with hominid cells or studies of hominid DNA sequences in transgenic animal models (Sholtis and Noonan 2010; McLean et al. 2011).

1.2.3 Ancient Genome Data

More recently the access to ancient DNA from extinct and ancestral hominids found in temperate regions outside Africa has allowed advances in anthropogeny that few could have imagined just two decades ago (Shapiro and Hofreiter 2014). Paradoxically, anthropogeny has gone from an almost complete lack of fossils in Darwin's time, to fragmentary fossils but no DNA in most of the twentieth century, to more fossils and snippets of DNA, to thousands of complete genomes, including those of all living ape species and even fossil taxa like Neanderthals and Denisovan (the latter represented by a single finger bone and tooth). In an ironic twist of scientific history we now have whole genome data for taxa represented by only a single finger bone and a tooth (Meyer et al. 2012). Novel sources of ancient DNA include dental calculus, which also provides a wealth of information on ancient hominid diet and microbiomes (Warinner et al. 2015).

Clear evidence for limited introgression (Green et al. 2010; Reich et al. 2010; Prufer et al. 2014) combined with strong evidence for overall selection against most introgressed archaic DNA (Currat and Excoffier 2011) has had the few remaining multiregionalists claiming victory, while the out-of–Africa side feels confirmed by the rare exceptions of introgressed functional elements such as HLA alleles and EPAS1 in Tibetans (Abi-Rached et al. 2011; Huerta-Sanchez et al. 2014). The availability of two extinct hominid genomes is now also allowing the identification of very recent changes post-dating the divergence of the lineage leading toe

H. sapiens and the archaic Eurasian taxa of Neanderthal and Denisovan (Paabo 2014).

1.2.4 Limits to Detecting Ancient Selection

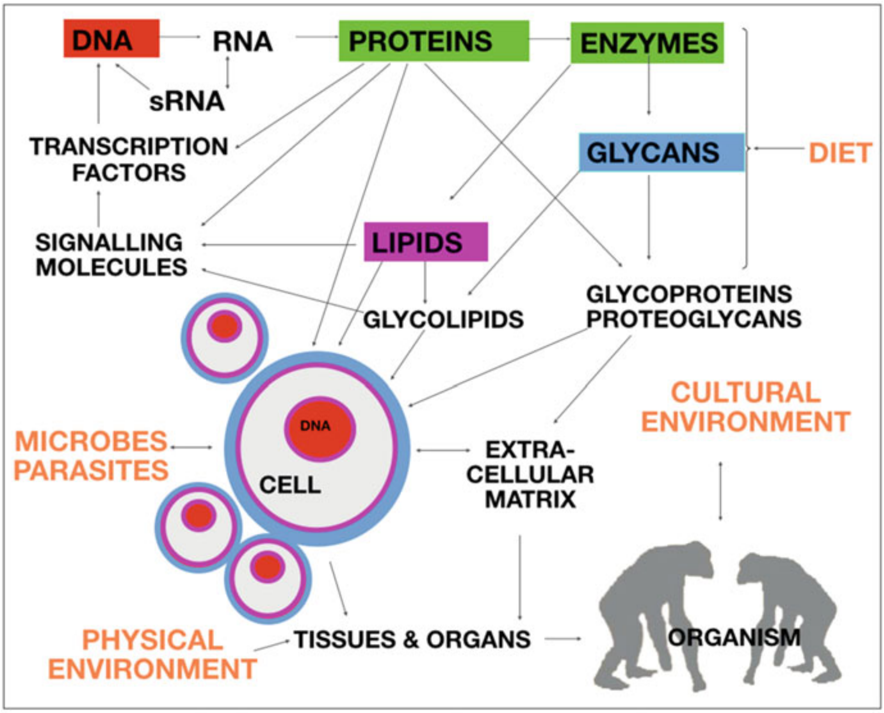
Ideally we would hope to find traces of past selection in areas of the genome responsible for unique modern human specializations. The irony is that the power to detect ancient selection in intra-specific sequence comparisons weakens substantially at just around the depth of time that modern humans appear on the scene (150 kya) (Granka et al. 2012; Voight et al. 2006).

Interpretations of this rapidly growing body of genomic data will crucially rely on independent investigation of countless phenomena ranging temporally from millisecond neuronal action potentials to million year geological epochs, and spatially from subatomic realms of stable isotopes to astronomical scales, affecting climate oscillation. It will also come to crucially rely on functional studies in cell culture and transgenic model animals. In addition, identifying the consequences in humans who carry deleterious mutations at human-specialized genetic loci will also help shed light on function in the absence of experiments.

The ongoing revelations about novel levels of complexity in genome organization and gene regulation make it difficult to clearly delineate where the genotype becomes phenotype. It is becoming clear that any stretch of DNA is already existing in a genomic environment comprising the location on a chromosome, chromatin structure, the identity of surrounding genetic elements (functional and non-functional "spacers"), and the presence of modifiers (Nadeau 2001). Non-coding RNAs and epigenetic modifications of DNA and histones are other dimensions blurring the genotype–phenotype boundary or forming a "code above the code." Classical phenotypes range from fossil teeth and bone to anatomy, physiology, development, and behavior of living individuals. Explanations of the human phenomenon will have to include all classes of biomolecules, their interactions during development of each organism as well as inter-organismal interactions starting with prenatal mother–offspring interface to social interactions within and between social groups (Fig. 1.2).

1.2.5 Phenotypes Are More Than Nucleic Acids and Proteins

Humans and all other organisms consist of four major classes of biomolecules: nucleic acids, proteins, lipids, and glycans. The latter two are not produced in a template driven manner, like the proteins encoded in genes, but rather are metabolically encoded and also influenced by the organism's diet and environment (Fig. 1.2). Lipids and glycans are also key components of extracellular tissues. Access to both is strongly affected by diet and the gut microbiome. An increased

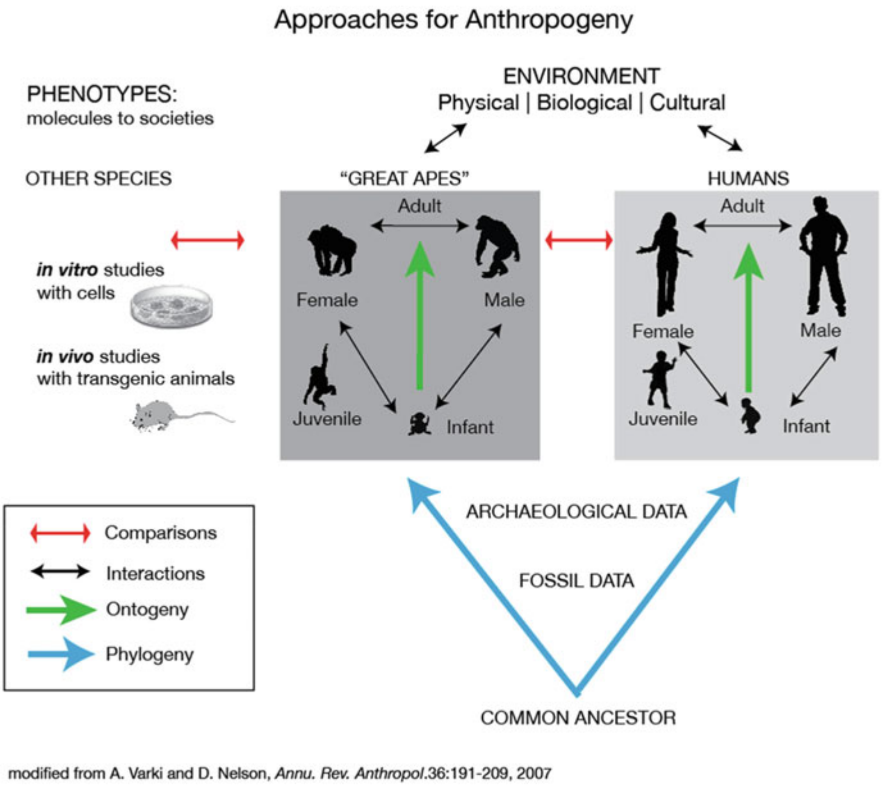


adapted from Varki, A. (Di Fiore, 2006)

Fig. 1.2 Molecules in context. Not all biomolecules are encoded in a template driven manner in the genome. Lipids and glycans come from the environment via diet and are then metabolically altered by enzymatic pathways of the organism. An understanding of uniquely human phenotypes must take into account all four major classes of biomolecules including nucleic acids, proteins, lipids, and glycans (modified from Varki, unpublished). Molecules in context

understanding of uniquely human phenotypes depends on an inclusive molecular approach, which appreciates how these four classes of biomolecules combine and interact (Marth 2008). For example, it was recently discovered that histones are modified by the addition of a single sugar (O-N-Acetylglucosamine or O-GlcNAc) to histone tails (Sakabe et al. 2010). Levels of sugar nucleotide substrate for this "histone code" modifications are heavily influenced by diet (Wells et al. 2003) providing a direct link between environmental/dietary conditions and histone post-translational modification.

Phenotypic information on non-human primates including the "great apes" will be crucial for interpreting genetic data. Comparing phenotypes of extant species at levels ranging from molecules to societies while taking into account phylogeny as well as ontogeny continues to reveal important facets of human specializations and point to systems and pathways where experimental work is warranted (Fig. 1.3). The hominin lineage consists after all of populations of reproducing individuals. A correct reconstruction of phylogeny requires correct inference about which



modified from A. Varki and D. Nelson, *Annu. Rev. Anthropol.*36:191-209, 2007

Fig. 1.3 Interdisciplinary approaches to anthropogeny. Paleontology provides crucial data on past life forms and their environments. Ancient DNA contributes key insights in the genetics of extinct and ancestral hominins. Comparative studies of human and great ape biology and behavior, including their ontogeny will continue to be key for identifying specializations of Homo sapiens. Such comparative studies should also include other species and be aided by in vitro studies using cells such as cell types derived from IPS cells obtained by minimally invasive ways from living individuals and experiments in transgenic model animals expressing manipulated to carry genetic material of interest (modified from Varki and Nelson 2007). Approaches for anthropogeny

populations continued to exchange genetic information. The classical view of phylogeny as a neat succession of bifurcations is complicated by the reality of hybridization (anastomosis/merging of lineages), which provides for networked phylogenies rather than neatly branched phylogenetic trees (Patterson et al. 2006; Reich et al. 2011; Jolly 2001). Availability of ancient DNA sequence from fossil and pre-fossil specimens is also opening up new avenues for directly measuring mutation rates (Fu et al. 2013; 2014).

1.3 Phenotypes: From Fossils to Past Behavior, Current Physiology, and Cognition

1.3.1 Fossil Data

Existing fossils clearly prove that bipedalism evolved early; that there were multiple bipedal lineages retaining excellent arboreal adaptations in Africa (White et al. 2009, Haile-Selassie et al. 2012); that the evolution of striding bipedalism came only more than two million years after the early bipedalism (Bramble and Lieberman 2004); that the expansion of cranial capacity came two million years ago (De Miguel and Henneberg 2001); that there was substantial anatomical variation among the first *Homo erectus (georgicus)* out of Africa (Lordkipanidze et al. 2013); that by 600 kya *Homo heidelbergensis* in Africa had reached cranial capacity comparable with modern humans (Conroy et al. 2000); that the continued expansion of the hominin cranium occurred despite increasing size constraints due to cephalo-pelvic disproportion at birth (Wells et al. 2012); that anatomically modern humans evolved 200–160 kya in East Africa (Fleagle et al. 2008, White et al. 2009). More fossils are badly needed but unfortunately very few field paleontology teams are enjoying stable financial support. The dearth of fossil representing the non-hominin (ape) lineages remains profoundly frustrating (McBrearty and Jablonski 2005; Suwa et al. 2007). The availability of powerful scanning technologies now allows studies on internal structures ranging from trabercular structure (Chirchir et al. 2015) to endocasts in breccia filled fossils such as *A. sediba* (Neubauer et al. 2012). Starch granules in tooth calculus (Henry et al. 2011) have allowed novel insights into the use of plant foods by Neanderthals and others.

Interpretation of the fossil record is also hampered by the tension between "lumpers," those who tend to group different fossils into the same taxon and "splitters" who tend to allocate each now fossil to its own taxon. A recent example is the reported large morphological variation among five *H. erectus* skulls from Dmanisi Georgia that all presumably represent a single species (Lordkipanidze et al. 2013) and the description of multiple taxa coexisting near Lake Turkana (Wood and Leakey 2011).

1.3.2 Archeological Data: Fossilized Behavior

Anthropogeny is first and foremost an African phenomenon. Behaviorally modern humans with fire, language, and projectile weapons colonized the entire planet and mostly replaced all other hominins (Henn et al. 2011). Much has been written about the "symbolic" watershed between "non-symbolic" ancestors and "fully symbolic," "behaviorally modern" humans. Tangible evidence for such a watershed remains very limited as there seem to be a few unequivocal tokens for "symbolic" behavior. Some of these, including shell beads and ochre for body paint long predate "behaviorally modern" humans (McBrearty and Brooks 2000), while others such as burials post-date their arrival (Gargett 1999). If burials are a clear sign of symbolic behavior, then such capacity might have evolved independently in Neanderthals and modern humans given the evidence for several Neanderthal burial sites across Europe and the Levant (Spikins et al. 2014).

Among the likely candidates for important impact on hominin genetics are the rise of the genus *Homo* to the place among top carnivores and its associated consumption of animal fat, use and reuse of home bases, the controlled use of fire, and cooking. Despite claims for the existence of home base use at more than one million years ago at Koobi Fora (Isaac et al. 1971), uncontested evidence is still pending (McBrearty and Brooks 2000; Brown et al. 2012). The cognitive specialization allowing for full theory of mind and language must include numerous genetic changes underlying neurodevelopment and brain function. These genetic changes contributed to a change in overall brain size and to complex reorganization and changes in overall connectivity.

1.3.3 Stable Isotopes, Paleoclimate, and Paleonutrition

Much information can be gleaned through the study of stable isotopes found in fossil material. Trophic levels, terrestrial versus marine diets, forest versus open grassland based diets are reflected in the ratio of stable isotopes of nitrogen and carbon and paleoclimate in reflected in oxygen and hydrogen isotopes (Schoeninger 2012; Bedaso et al. 2013). Plant wax biomarkers in sedimentary mud (sapropel), isotope composition of soil carbonate, and composition of fossil bovid fauna all point to an increase in more open vegetation over the last three million years (deMenocal 2011). The combination of stable isotope data with dental microwear data has allowed surprisingly detailed reconstruction of ancient hominid diets (Ungar and Sponheimer 2011). Finally microcharcoal in sediments from offshore or lake sediments can be powerful indicators of widespread burning, both natural and hominin in origins. Discerning the latter from the background rate of naturally ignited fires remains a big challenge (Bowman et al. 2011).

1.3.4 Learning from Living Foragers

There are a few remaining forager societies across the world, people who live entirely from gathering and hunting. Anthropologists and ethnographers studying these groups can glean powerful insights into the human condition prior to agriculture. Serious caveats include the important fact that these living societies do not represent ancient relics, and that they have routinely been pushed into degraded habitats not desirable for their pastoralist and agriculturalist neighbors. Nevertheless, behavioral patterns and cultural traits of these small-scale societies provide important information regarding past human life (Marlowe et al. 2014). Ethnographic, genetic, and linguistic data from such groups continues to inform our understanding of human biological and cultural evolution in crucial ways (Henn et al. 2011; Walker et al. 2011; Wiessner 2014; Tishkoff et al. 2007). Much interest exists in studying the microbiome of living foragers as examples of non-agricultural ecosystems. Life history studies on the Hadza of Tanzania were key for the development of the grandmother hypothesis (Kim et al. 2014). Information from non-agricultural societies also provides important insights about violent behavior and its determinants (Boehm 2012; Muller et al. 2009).

1.3.5 The Holocene Trap

Despite asking key questions about the Pleistocene, the wealth of data we have from the Holocene (including all of the neolithic behavioral records) unwittingly leads us to discuss more recent phenomena when trying to explain a much more ancient singularity. Our most cherished examples of cultural and behavioral adaptations shaping human genomes include lactase persistence and amylase copy number, both of which are firmly linked to neolithic times, i.e. the harvesting of milk from other mammals or feeding on grass grains, and do not shed light on the origins of our species (Tishkoff et al. 2007; Perry et al. 2007). These Holocene examples do however blatantly illustrate the power of behavioral and cultural adaptations for shaping human genetics. There is a need for considering cultural adaptations long predating the Neolithic, which might have similarly shaped our biology. How could the early use of fire and reuse of home bases have molded parts of our biology?

1.3.6 Biological Proxies for Past Behavior

DNA sequences encoding genes involved in reproductive behavior and their expression patterns allow a glimpse into past mating systems and are indicative of the presence of pair bonding (neotenous gene expression in male reproductive genes) (Saglican et al. 2014). This neotenous pattern mirrors clearly neotenous changes in gene expression networks in the brain (Somel et al. 2009). Similarly, sexual dimorphism or lack thereof in fossils can shed light on the degree of competition for mates. Recent work using digit ratios as proxy for in utero androgen exposure has generated interesting data with regard to past mating systems and the potential existence of alternative mating strategies in both sexes in humans (Wlodarski et al. 2015). The most unusual characteristics of our species, namely full theory of mind and language have no known biological counterparts that would survive in fossilized hard tissues (Povinelli and Preuss 1995).

Comparative medicine provides a long list of ailments apparently unique or unusual to our species (Varki et al. 2011). The ascertainment bias due to the two thousand plus years of history of human medicine but much younger medical knowledge of great apes needs to be kept in mind. A large number of diseases clearly affect humans differently. Major differences in immune system biology might underlie some of these disease differences, but how humans came to have such different immune systems remains an important unanswered question (Varki 2010). A further question is the degree to which derived genes represent a liability for disruption of proper neurodevelopment as evidenced by many human cognitive disorders. Studies of uniquely derived genes involved in human cognitive development and function and their disruption in individuals with intellectual disability are very promising in this respect (Hormozdiari et al. 2015).

1.3.7 The Crying Need for Phenotypic Data of Non-human Hominids

Natural selection operates mostly on individual phenotypes. Sadly, we only have limited information about the phenotypes of our closest living ape relatives. Opportunities for obtaining such information are rapidly vanishing with the closure of the last primate centers. The great ape sanctuaries across Africa offer some hope for continued access to great ape phenotypic studies (Farmer 2002). Common chimpanzees are the only “great ape” ever kept in captivity in significant numbers and even used for biomedical research. Captive chimpanzee populations are aging and biomedical research has come to near cessation with very few exceptions. Cell biology with induced pluripotent stem cells derived from minimally invasive samples (skin biopsies or milk teeth) offers some very promising avenues for studying cellular phenotypes including derived neuronal and other central nervous system cell types (Hrvoj-Mihic et al. 2014).

The use of non-human animal models also promises to produce important insights. Genetic changes including those in controlling regions such as HAR1 can be tested by transgenic expression in mouse (Capra et al. 2013; Prabhakar et al. 2008). Obviously, the lack of primate genomic background for such experiments in mice remains an important limitation. Recently established colonies of dwarf primates (marmosets) in China and Japan are intended to provide for transgenic primate experiments (Kishi et al. 2014).

Much of the human-specific biology takes place during development in utero, where experiments are ethically not possible in either apes or humans (Gagneux et al. 2005). Non-invasive imaging in humans and captive chimpanzees can provide extremely valuable insights such as the observation that the rate and velocity of chimpanzee brain decline well before birth while remaining steady in humans (Sakai et al. 2013). The schedule of myelination has also dramatically changed with an extension of mature myelination into the third decade of life for humans (Miller et al. 2012). Non-invasive imaging of human, chimpanzee, and macaque brains has revealed remarkable human-specific connectivity via the strongly lateralized arcuate fasciculus connecting Broca's and Wernicke's area (Rilling et al. 2008; Chen et al. 2013). Similarly, more detailed studies of cellular architecture in post-mortem brain samples are revealing striking differences between comparable regions in humans and "great apes" (Semendeferi et al. 2011). Most notably in regions involved in the limbic system and social cognition (Barger et al. 2014).

1.3.8 Niche Construction and Top-Down Effects

Complex neuro-behavioral phenotypes are subject to both, bottom-up regulation by genes affecting development and metabolism and top-down effects in the form of social and cultural input, which famously include diet, linguistic, and sociocultural input during a prolonged period of neuronal maturation in our species. How does the human genome encode propensity for language and a pattern of brain development that "expects" language input? Even more perplexing is the question about how such information underlying our linguistic capacities became internalized in the human germ line in the first place. Evidence for anatomical differences between brains of monolingual and second language learners would be further evidence of top-down effects (Klein et al. 2014; Mechelli et al. 2004).

A chimpanzee brain develops perfectly fine in the absence of language input, whereas a human brain does not reach its potential unless a child is spoken to (Greenough et al. 1987; Kuhl et al. 1992). It is striking that even apparently obvious biological traits such as bipedality appear to be subject to important learning and imitation for proper bipedal locomotion, despite the many anatomical adaptations to bipedality (Thelen 1995).

Among the top-down effects one could also consider the provocative idea of human "self-domestication," a form of social selection, against aggression within groups with important consequences for pro-social behavior and group function (Hare et al. 2012). Such a process could be in part responsible for the simultaneous selection of neotenous traits and shifts in developmental schedule typical of *Homo sapiens*. Delayed maturation, retention of juvenile characters, and heterochrony are all hallmarks of human development (Miller et al. 2012; Somel et al. 2009; Liu et al. 2012).

Humans are biologically dependent on cooked food as evidenced by the finding that female raw food eaters in modern societies frequently cease to ovulate (Carmody et al. 2011). Higher-level cognition cannot be explained outside the biological, social, and cultural contexts in which it evolved (Nunez et al. 2012). The cultural niche becomes a force in its own right, profoundly shaping human cognition and behavior. It is well conceivable that such higher-level niches have impacted the human genome differently than those of the "great apes" (Varki et al. 2008). To humans, fellow humans act as powerful "transcription factors" even more so than conspecifics do in other highly social species.

1.3.9 The Physical Niche

Exploitation of a large variety of landscapes was an early adaptation of hominins. It might have contributed to early bipedalism as a more efficient way of covering longer distances and gathering resources. It certainly contributed to a much wider set of food types consumed. Fire easily represents one of the most important cultural and technological breakthroughs of the hominin lineage. Vexingly, we still lack a firm evidence for the true age of this key innovation. Oldest evidence is currently at 1 mya (Wonderwerk Cave, South Africa) (Berna et al. 2012), but reasonably convincing arguments based on molar size reduction in the fossil record have been made for a role in the use of fire by early *Homo erectus* as early as 2 mya (Organ et al. 2011). The control of fire likely ushered in massive improvements in niche construction with profound effects on ecology, protection from predation, diet via cooking, and cognition via extended days and the social effects of social gatherings around fires (Wiessner 2014). It also provided novel technological opportunities by allowing altering of materials such as silcrete and compound adhesives (Brown et al. 2009, 2012). The phylogeny of head and body lice provides indirect evidence for early adoption of clothing by modern humans in Africa, a behavioral transition loaded with symbolic potential (Toups et al. 2011). The precise age of home base use is unknown, but the shift to repeated home base use (in contrast to daily new nests) would have dramatically altered the pathogen regime of our ancestors.

1.3.10 The Socio-Cognitive Niche

Human co-residence of multiple males and females but simultaneous widespread pair bonding is an arrangement not seen in any other primate. Pre-agricultural humans lived in small groups but these were likely part of extensive social networks linking such groups over generations. Pair bonding allows for increased confidence of paternity and through it to the establishment of male kinship networks that span individual social groups (Chapais 2013) allowing the evolution of “meta-group” social structure in humans as created by marriage patterns found across human cultures. Studies of marriage patterns in hunter-gatherers would indicate that human culture has been intricately implicated in marriage decisions from times long before agriculture (Walker et al. 2011). There is strong evidence for cultural effects on modern human genomes via different marriage rules in societies around the world with elevated lengths of runs of homozygosity in societies encouraging uncle-niece or first cousin marriages (Pemberton et al. 2012).

1.4 The Cultural Niche

Unlike the case with other, species human culture is ratcheting culture, whereby innovations are not only maintained across generations but can be further improved upon and even combined to form entire technologies (Dean et al. 2014). Human language is a powerful way of maintaining innovations and spreading these across social groups. Cognitive innovations such as beliefs about agency in nature and the supernatural become carried by language and are passed down through the generations. There is accumulating evidence that such beliefs can be very adaptive for both individuals and groups (Baumard and Boyer 2013). Rituals and norms can become powerfully anchored in local culture and enforced by institutions. Interestingly climatic and biological effects appear to affect both language and belief systems as the diversity is higher in the tropics for both religions and languages (Fincher and Thornhill 2008).

1.5 Language and Theory of Mind

Among the most important and species-specific social and cultural inputs in humans is language. This species-specific communication system simultaneously allows communication of experiences and ideas across individuals, time, and space, but also, rather paradoxiacally effectively precludes such communication across linguistic groups (Pagel 2009). Whether human language results from a saltationist event or from the combination of preexisting animal communication systems is a hotly debated issue. The same can be said about the question about the transition from no language to protolanguage via or with an important gestural component.

Much valuable insight continues to come from comparative animal psychology in the laboratory, the field, zoos or great ape sanctuaries (Herrmann et al. 2014; Subiaul et al. 2008), despite the fact that the extreme capacity of humans to envisage the mental life of others and to engage in shared “mental time travel” by using language has no, or at best very limited counterparts in other species.

Language allows for the establishment of widespread reputation, which introduces a completely novel factor in social behavior. Altruism towards non-kin, generosity, and even third-party punishment all can be highly favored by the existence and individual reputation and our awareness of it (Hardy and Van Vugt 2006). Chimpanzees can glean the generosity of other individual during observation of third-party interactions, but without language, they cannot spread that reputation beyond the actual observation (Subiaul et al. 2008). Efforts to find novel ways of studying language evolution based on syntax and phomenes are yielding some tantalizing insights similarities and differences between genetic and language evolution (Colonna et al. 2010; Creanza et al. 2015). While chimpanzee understands the psychology of others to a degree, they seem to lack a human-like theory of mind (Tomasello et al. 2003). The notion that there might have been a psychological threshold/barrier to full awareness of self and others (theory of mind) deserves special attention (Varki and Brower 2013), as it may help to explain why only one species of hominin was eventually left standing after exploiting its self-generated socio-cognitive niche to the fullest.

1.5.1 The Brain Needs the Body and the Group

The brainpower of our species undoubtedly underlies many of the cognitive specialization of *H. sapiens*. Currently, several genes with signatures of uniquely human changes and roles in neurodevelopment, including a handful dating to after the divergence of modern humans and Neanderthals are among the "hottest" candidates for getting at the genetics that make us humans (Paabo 2014).

Humans, however, are more than their large brains. Many of the genes involved in neurobiology have important functions for reproduction and immunity as well. It is important to consider in parallel that human mothers have to be able to gestate and give birth to large-headed babies against the constraints of a bipedal pelvis. The social system has to support mothers and their extremely altricial babies who carry on with a fetal rate of brain growth for a full year after birth. Such offspring are dependent on "mothers and others" especially once inter-birth interval shortens to where weaned offspring cannot find enough food on their own (Blaffer Hrdy 2009). Nutritional opportunities need to exist for the development of our expensive central nervous system with proposed necessary shift to higher trophic levels (top predator) (Hoberg et al. 2001), more marine resources, and/or cooking (Marean 2010; Organ et al. 2011).

1.6 Opportunities and Limitations

Anthropogenists are still facing a staggering list of humbling unknowns about the age of onset of key human innovations. These include but are not restricted to the following: symbolic capacity, personal name or kinship terms, language, home base use, fire use/cooking, pair bonding, awareness of paternal kinship networks, projectile weapon use, composite tool use, fiber use (strings, baskets, nets, hunting machines/traps, bow string, slingshot), bodily modifications (painting with pigments, scarifications, genital cutting, tattoos), and death rituals.

1.7 Open Minds, Closed Umbrellas

The human phenomenon likely reflects an idiosyncratic concatenation of unlikely events. Key factors likely include both opportunities and constraints stemming from massive physical and cultural niche construction by our species that has increasingly taken its own evolutionary fate in its own hands.

Famous umbrella hypotheses such as the “aquatic ape” or the “savannah ape” blatantly fail to account for the many human traits, which arose over a period of several million years (Langdon 1997).

Anthropogeny requires openness to least likely scenarios including ones not directly related to cognitive capacities, to name a few proposed candidates:

- Infection and immunity and their potential ramification for behavior and central nervous system development (Wang et al. 2012).
- Microbiomes, their establishment, evolutionary modification, and profound effects on the entire organism including mental function (Salvucci 2014).
- Climate and geophysical events (Mount Toba eruption) that can exert strong selection of human adaptability via culture and mental flexibility (Ambrose 1998; Calvin 2002).
- The interplay between stone tool manufacture, with strong lateralization of hand use and potential requirements for mental syntax (Stout and Chaminade 2012).
- The use of projectiles to hunt mobile prey and the importance of relative position, directionality, anticipation of motion as exaptation for syntax (Calvin 2001).
- Shifts in ecology allowing the lifting of energetic and nutrient limitations (Bradbury 2011; Organ et al. 2011) and opening of novel symbolic expressions (Duarte 2014; Henshilwood et al. 2011).

1.7.1 The Need for Transdisciplinarity

The human phenomenon includes wide ranges of spatial and temporal scales, from the subatomic (stable isotopes) to the astronomical (solar cycles and climate) and from the millisecond (neuronal action potential) to million years (paleontology). It also requires dealing with the deterministic when studying molecular mechanisms and the arbitrary when studying cultural attributes. Advances in understanding the human phenomenon will likely come from fresh perspectives originating from unexpected fields of research, involving researchers who do not shy away from difficult dialogues and collaborations. The lack of dialogue between sociocultural anthropologists and the physical and natural sciences represents an important hurdle for such transdisciplinary endeavors. The only hope forward is to promote interactions and willingness to refrain from mutual accusations over reductionism and scientism versus postmodernism and relativism. Human cultures and societies have played and continue to play important roles in shaping our biology, which in turn is part of any human cultural phenomenon.

1.8 Why Anthropogeny?

In these days of reduced funding for basic science and strong impetus on translational research, why engage in the exploration of the origins of the human phenomenon?

For one, questions about our origin are likely as old as our species and individuals not interested in their origins are very few. The study of our past deep history also promises to reveal important insights in how intricately involved humans themselves are in shaping their own biological destiny, long before the times of assisted reproduction. Such insights are bound to inform us in important ways about how we care for our young, how we make decisions and form moral views, how different societies use norms and sanctions to channel behaviors, and how different societies chose to interact. Anthropogeny will also provide important novel perspectives on human diseases and disabilities and potentially point to novel ways of preventing, treating, and managing these by understanding which factors in our past including some of the very features we celebrate as uniquely human achievements may predispose many of us for unnecessary suffering.

1.9 Note of Caution

In Ernst Haeckel's time, the lack of fossils prompted him to use fellow humans as "intermediary forms" between apes and Europeans in his Anthropogenie (Haeckel 1891). The study of our origins is loaded with such narcissistic bias and most of us will be tempted by findings with flattering implications about our own groups. Objectivity is heavily compromised when the object of study is our own origin (Marks 2012). Irrespective of whether this is a feeling of pride due to the perception of belonging to the more "original" groups or the more "derived" groups. Given the ugly history of early twentieth century anthropology and persisting attempts at classifying and ranking different human groups, it behooves any anthropogenist and student of the human genome to perpetually be on guard against Haeckel's specter of racist ideologies. After all, one of the hallmark characteristics of the human mind is how easily it adopts the parochial in-group/out-group paradigm (Bernhard et al. 2006). Luckily, that same characteristic can be exploited in advancing the effort to understand the human phenomenon as the hallmark of "our group" as a global species.

Acknowledgements I thank Ajit Varki for valuable comments on a draft of this chapter.

References

Abi-Rached L, Jobin MJ, Kulkarni S, McWhinnie A, Dalva K, Gragert L, Babrzadeh F, Gharizadeh B, Luo M, Plummer FA, Kimani J, Carrington M, Middleton D, Rajalingam R, Beksac M, Marsh SG, Maiers M, Guethlein LA, Tavoularis S, Little AM, Green RE, Norman PJ, Parham P (2011) The shaping of modern human immune systems by multiregional admixture with archaic humans. Science 334:89–94

Ambrose SH (1998) Late Pleistocene human population bottlenecks, volcanic winter, and differentiation of modern humans. J Hum Evol 34:623–651

Barger N, Hanson KL, Teffer K, Schenker-Ahmed NM, Semendeferi K (2014) Evidence for evolutionary specialization in human limbic structures. Front Hum Neurosci 8:277

Baumard N, Boyer P (2013) Explaining moral religions. Trends Cogn Sci 17:272–280

Bedaso ZK, Wynn JG, Alemseged Z, Geraads D (2013) Dietary and paleoenvironmental reconstruction using stable isotopes of herbivore tooth enamel from middle Pliocene Dikika, Ethiopia: implication for Australopithecus afarensis habitat and food resources. J Hum Evol 64:21–38

Berna F, Goldberg P, Horwitz LK, Brink J, Holt S, Bamford M, Chazan M (2012) Microstratigraphic evidence of in situ fire in the Acheulean strata of Wonderwerk cave, northern cape province, South Africa. Proc Natl Acad Sci U S A 109:E1215–E1220

Bernhard H, Fischbacher U, Fehr E (2006) Parochial altruism in humans. Nature 442:912–915

Blaffer Hrdy S (2009) Mother and others. Belknap Press, Cambridge

Boehm C (2012) Ancestral hierarchy and conflict. Science 336:844–847

Bowman DM, Balch J, Artaxo P, Bond WJ, Cochrane MA, D'Antonio CM, Defries R, Johnston FH, Keeley JE, Krawchuk MA, Kull CA, Mack M, Moritz MA, Pyne S, Roos CI, Scott AC, Sodhi NS, Swetnam TW, Whittaker R (2011) The human dimension of fire regimes on earth. J Biogeogr 38:2223–2236

Bradbury J (2011) Docosahexaenoic acid (DHA): an ancient nutrient for the modern human brain. Nutrients 3:529–554

Bramble DM, Lieberman DE (2004) Endurance running and the evolution of homo. Nature 432:345–352

Britten RJ (2002) Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels. Proc Natl Acad Sci U S A 99:13633–13635

Brown KS, Marean CW, Herries AI, Jacobs Z, Tribolo C, Braun D, Roberts DL, Meyer MC, Bernatchez J (2009) Fire as an engineering tool of early modern humans. Science 325:859–862

Brown KS, Marean CW, Jacobs Z, Schoville BJ, Oestmo S, Fisher EC, Bernatchez J, Karkanas P, Matthews T (2012) An early and enduring advanced technology originating 71,000 years ago in South Africa. Nature 491:590–593

Calvin WH (2001) The throwing Madonna: essays on the brain. iUniverse.com, Bengaluru

Calvin WH (2002) A brain for all seasons human evolution and abrupt climate change. University of Chicago Press, Chicago

Capra JA, Erwin GD, McKinsey G, Rubenstein JL, Pollard KS (2013) Many human accelerated regions are developmental enhancers. Philos Trans R Soc Lond Ser B Biol Sci 368:20130025

Carmody RN, Weintraub GS, Wrangham RW (2011) Energetic consequences of thermal and nonthermal food processing. Proc Natl Acad Sci U S A 108:19199–19203

Chapais B (2013) Monogamy, strongly bonded groups, and the evolution of human social structure. Evol Anthropol 22:52–65

Chen X, Errangi B, Li L, Glasser MF, Westlye LT, Fjell AM, Walhovd KB, Hu X, Herndon JG, Preuss TM, Rilling JK (2013) Brain aging in humans, chimpanzees (Pan troglodytes), and rhesus macaques (Macaca mulatta): magnetic resonance imaging studies of macro- and microstructural changes. Neurobiol Aging 34:2248–2260

Chirchir H, Kivell TL, Ruff CB, Hublin JJ, Carlson KJ, Zipfel B, Richmond BG (2015) Recent origin of low trabecular bone density in modern humans. Proc Natl Acad Sci U S A 112:366–371

Colonna V, Boattini A, Guardiano C, Dall'ara I, Pettener D, Longobardi G, Barbujani G (2010) Long-range comparison between genes and languages based on syntactic distances. Hum Hered 70:245–254

Conroy GC, Weber GW, Seidler H, Recheis W, Zur Nedden D, Mariam JH (2000) Endocranial capacity of the bodo cranium determined from three-dimensional computed tomography. Am J Phys Anthropol 113:111–118

Consortium The ENCODE Project (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74

Creanza N, Ruhlen M, Pemberton TJ, Rosenberg NA, Feldman MW, Ramachandran S (2015) A comparison of worldwide phonemic and genetic variation in human populations. Proc Natl Acad Sci U S A 112(5):1265–1272

Currat M, Excoffier L (2011) Strong reproductive isolation between humans and Neanderthals inferred from observed patterns of introgression. Proc Natl Acad Sci U S A 108:15129–15134

Darwin C (1871) The descent of man, and selection in relation to sex. D Appleton and company, New York

De Miguel C, Henneberg M (2001) Variation in hominid brain size: how much is due to method? Homo 52:3–58

Dean LG, Vale GL, Laland KN, Flynn E, Kendal RL (2014) Human cumulative culture: a comparative perspective. Biol Rev Camb Philos Soc 89:284–301

deMenocal PB (2011) Anthropology. Climate and human evolution. Science 331:540–542

Di Fiore A, Lawler RR, Gagneux P (2006) Molecular primatology. In: Campbell CJ, Fuentes A, McKinnon KC, Panger M, Bearder SK, Stumpf RM (eds) Primates in perspective, 2nd edn. Oxford University press, Oxford, pp 390–416

Diamond JM (1989) The present, past and future of human-caused extinctions. Philos Trans R Soc Lond Ser B Biol Sci 325:469–476; discussion 476

Duarte CM (2014) Red ochre and shells: clues to human evolution. Trends Ecol Evol 29:560–565

Farmer KH (2002) Pan-African sanctuary alliance: status and range of activities for great ape conservation. Am J Primatol 58:117–132

Fincher CL, Thornhill R (2008) Assortative sociality, limited dispersal, infectious disease and the genesis of the global pattern of religion diversity. Proc Biol Sci 275:2587–2594

Fleagle JG, Assefa Z, Brown FH, Shea JJ (2008) Paleoanthropology of the Kibish formation, southern Ethiopia: introduction. J Hum Evol 55:360–365

Fu Q, Mittnik A, Johnson PL, Bos K, Lari M, Bollongino R, Sun C, Giemsch L, Schmitz R, Burger J, Ronchitelli AM, Martini F, Cremonesi RG, Svoboda J, Bauer P, Caramelli D, Castellano S, Reich D, Paabo S, Krause J (2013) A revised timescale for human evolution based on ancient mitochondrial genomes. Curr Biol 23:553–559

Fu Q, Li H, Moorjani P, Jay F, Slepchenko SM, Bondarev AA, Johnson PL, Aximu-Petri A, Prufer K, de Filippo C, Meyer M, Zwyns N, Salazar-Garcia DC, Kuzmin YV, Keates SG, Kosintsev PA, Razhev DI, Richards MP, Peristov NV, Lachmann M, Douka K, Higham TF, Slatkin M, Hublin JJ, Reich D, Kelso J, Viola TB, Paabo S (2014) Genome sequence of a 45,000-year-old modern human from western Siberia. Nature 514:445–449

Gagneux P, Moore JJ, Varki A (2005) The ethics of research on great apes. Nature 437:27–29

Gargett RH (1999) Middle Palaeolithic burial is not a dead issue: the view from Qafzeh, saint-Cesaire, Kebara, Amud, and Dederiyeh. J Hum Evol 37:27–90

Gazave E, Darre F, Morcillo-Suarez C, Petit-Marty N, Carreno A, Marigorta UM, Ryder OA, Blancher A, Rocchi M, Bosch E, Baker C, Marques-Bonet T, Eichler EE, Navarro A (2011) Copy number variation analysis in the great apes reveals species-specific patterns of structural variation. Genome Res 21:1626–1639

Gibbs S, Collard M, Wood B (2000) Soft-tissue characters in higher primate phylogenetics. Proc Natl Acad Sci U S A 97:11130–11132

Goodman M, Koop BF, Czelusniak J, Fitch DH, Tagle DA, Slightom JL (1989) Molecular phylogeny of the family of apes and humans. Genome 31:316–335

Goodman M, Porter CA, Czelusniak J, Page SL, Schneider H, Shoshani J, Gunnell G, Groves CP (1998) Toward a phylogenetic classification of primates based on DNA evidence complemented by fossil evidence. Mol Phylogenet Evol 9:585–598

Granka JM, Henn BM, Gignoux CR, Kidd JM, Bustamante CD, Feldman MW (2012) Limited evidence for classic selective sweeps in African populations. Genetics 192:1049–1064

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH, Hansen NF, Durand EY, Malaspinas AS, Jensen JD, Marques-Bonet T, Alkan C, Prufer K, Meyer M, Burbano HA, Good JM, Schultz R, Aximu-Petri A, Butthof A, Hober B, Hoffner B, Siegemund M, Weihmann A, Nusbaum C, Lander ES, Russ C, Novod N, Affourtit J, Egholm M, Verna C, Rudan P, Brajkovic D, Kucan Z, Gusic I, Doronichev VB, Golovanova LV, Lalueza-Fox C, de la Rasilla M, Fortea J, Rosas A, Schmitz RW, Johnson PL, Eichler EE, Falush D, Birney E, Mullikin JC, Slatkin M, Nielsen R, Kelso J, Lachmann M, Reich D, Paabo S (2010) A draft sequence of the Neandertal genome. Science 328:710–722

Greenough WT, Black JE, Wallace CS (1987) Experience and brain development. Child Dev 58:539–559

Haeckel E (1891) Anthropogenie; oder, Entwickelungsgeschichte des menschen. Keimes- und stammesgeschichte. W. Engelmann, Leipzig

Haile-Selassie Y, Saylor BZ, Deino A, Levin NE, Alene M, Latimer BM (2012) A new hominin foot from Ethiopia shows multiple Pliocene bipedal adaptations. Nature 483:565–569

Hardy CL, Van Vugt M (2006) Nice guys finish first: the competitive altruism hypothesis. Personal Soc Psychol Bull 32:1402–1413

Hare B, Wobber V, Wrangham R (2012) The self-domestication hypothesis: evolution of bonobo psychology is due to selection against aggression. Anim Behav 83:573–585

Henn BM, Gignoux CR, Jobin M, Granka JM, Macpherson JM, Kidd JM, Rodriguez-Botigue L, Ramachandran S, Hon L, Brisbin A, Lin AA, Underhill PA, Comas D, Kidd KK, Norman PJ, Parham P, Bustamante CD, Mountain JL, Feldman MW (2011) Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. Proc Natl Acad Sci U S A 108:5154–5162

Henry AG, Brooks AS, Piperno DR (2011) Microfossils in calculus demonstrate consumption of plants and cooked foods in Neanderthal diets (Shanidar III, Iraq; Spy I and II, Belgium). Proc Natl Acad Sci U S A 108:486–491

Henshilwood CS, d’Errico F, van Niekerk KL, Coquinot Y, Jacobs Z, Lauritzen SE, Menu M, Garcia-Moreno R (2011) A 100,000-year-old ochre-processing workshop at Blombos cave, South Africa. Science 334:219–222

Herrmann E, Misch A, Hernandez-Lloreda V, Tomasello M (2014) Uniquely human self-control begins at school age. Dev Sci 18(6):979–993

Hoberg EP, Alkire NL, de Queiroz A, Jones A (2001) Out of Africa: origins of the Taenia tapeworms in humans. Proc Biol Sci 268:781–787

Hormozdiari F, Penn O, Borenstein E, Eichler EE (2015) The discovery of integrated gene networks for autism and related disorders. Genome Res 25:142–154

Hrvoj-Mihic B, Marchetto MC, Gage FH, Semendeferi K, Muotri AR (2014) Novel tools, classic techniques: evolutionary studies using primate pluripotent stem cells. Biol Psychiatry 75:929–935

Huerta-Sanchez E, Jin X, Asan BZ, Peter BM, Vinckenbosch N, Liang Y, Yi X, He M, Somel M, Ni P, Wang B, Ou X, Huasang LJ, Cuo ZX, Li K, Gao G, Yin Y, Wang W, Zhang X, Xu X, Yang H, Li Y, Wang J, Wang J, Nielsen R (2014) Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature 512:194–197

Huxley TH (1863) Evidence as to Man’s place in nature. Williams and Norgate, London

Isaac GL, Leakey RE, Behrensmeyer AK (1971) Archeological traces of early hominid activities, east of lake Rudolf, Kenya. Science 173:1129–1134

Jolly CJ (2001) A proper study for mankind: analogies from the Papionin monkeys and their implications for human evolution. Am J Phys Anthropol 33(Suppl):177–204

Kim PS, McQueen JS, Coxworth JE, Hawkes K (2014) Grandmothering drives the evolution of longevity in a probabilistic model. J Theor Biol 353:84–94

King MC, Wilson AC (1975) Evolution at two levels in humans and chimpanzees. Science 188:107–116

Kishi N, Sato K, Sasaki E, Okano H (2014) Common marmoset as a new model animal for neuroscience research and genome editing technology. Develop Growth Differ 56:53–62

Klein D, Mok K, Chen JK, Watkins KE (2014) Age of language learning shapes brain structure: a cortical thickness study of bilingual and monolingual individuals. Brain Lang 131:20–24

Kondgen S, Kuhl H, N'Goran PK, Walsh PD, Schenk S, Ernst N, Biek R, Formenty P, Matz-Rensing K, Schweiger B, Junglen S, Ellerbrok H, Nitsche A, Briese T, Lipkin WI, Pauli G, Boesch C, Leendertz FH (2008) Pandemic human viruses cause decline of endangered great apes. Curr Biol 18:260–264

Kuhl PK, Williams KA, Lacerda F, Stevens KN, Lindblom B (1992) Linguistic experience alters phonetic perception in infants by 6 months of age. Science 255:606–608

Langdon JH (1997) Umbrella hypotheses and parsimony in human evolution: a critique of the aquatic ape hypothesis. J Human Evol 33:479–494

Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, Ward LD, Lowe CB, Holloway AK, Clamp M, Gnerre S, Alfoldi J, Beal K, Chang J, Clawson H, Cuff J, Di Palma F, Fitzgerald S, Flicek P, Guttman M, Hubisz MJ, Jaffe DB, Jungreis I, Kent WJ, Kostka D, Lara M, Martins AL, Massingham T, Moltke I, Raney BJ, Rasmussen MD, Robinson J, Stark A, Vilella AJ, Wen J, Xie X, Zody MC, Baldwin J, Bloom T, Chin CW, Heiman D, Nicol R, Nusbaum C, Young S, Wilkinson J, Worley KC, Kovar CL, Muzny DM, Gibbs RA, Cree A, Dihn HH, Fowler G, Jhangiani S, Joshi V, Lee S, Lewis LR, Nazareth LV, Okwuonu G, Santibanez J, Warren WC, Mardis ER, Weinstock GM, Wilson RK, Delehaunty K, Dooling D, Fronik C, Fulton L, Fulton B, Graves T, Minx P, Sodergren E, Birney E, Margulies EH, Herrero J, Green ED, Haussler D, Siepel A, Goldman N, Pollard KS, Pedersen JS, Lander ES, Kellis M (2011) A high-resolution map of human evolutionary constraint using 29 mammals. Nature 478:476–482

Liu X, Somel M, Tang L, Yan Z, Jiang X, Guo S, Yuan Y, He L, Oleksiak A, Zhang Y, Li N, Hu Y, Chen W, Qiu Z, Paabo S, Khaitovich P (2012) Extension of cortical synaptic development distinguishes humans from chimpanzees and macaques. Genome Res 22:611–622

Lordkipanidze D, Ponce de Leon MS, Margvelashvili A, Rak Y, Rightmire GP, Vekua A, Zollikofer CP (2013) A complete skull from Dmanisi, Georgia, and the evolutionary biology of early homo. Science 342:326–331

Marean CW (2010) Pinnacle point cave 13B (Western Cape Province, South Africa) in context: the cape floral kingdom, shellfish, and modern human origins. J Hum Evol 59:425–443

Marks J (2012) Why be against Darwin? Creationism, racism, and the roots of anthropology. Am J Phys Anthropol 149(Suppl 55):95–104

Marlowe FW, Berbesque JC, Wood B, Crittenden A, Porter C, Mabulla A (2014) Honey, Hadza, hunter-gatherers, and human evolution. J Hum Evol 71:119–128

Marth JD (2008) A unified vision of the building blocks of life. Nat Cell Biol 10:1015–1016

Mcbrearty S, Brooks AS (2000) The revolution that wasn't: a new interpretation of the origin of modern human behavior. J Hum Evol 39:453–563

McBrearty S, Jablonski NG (2005) First fossil chimpanzee. Nature 437:105–108

McLean CY, Reno PL, Pollen AA, Bassan AI, Capellini TD, Guenther C, Indjeian VB, Lim X, Menke DB, Schaar BT, Wenger AM, Bejerano G, Kingsley DM (2011) Human-specific loss of regulatory DNA and the evolution of human-specific traits. Nature 471:216–219

Mechelli A, Crinion JT, Noppeney U, O'Doherty J, Ashburner J, Frackowiak RS, Price CJ (2004) Neurolinguistics: structural plasticity in the bilingual brain. Nature 431:757

Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prufer K, de Filippo C, Sudmant PH, Alkan C, Fu Q, Do R, Rohland N, Tandon A, Siebauer M, Green RE, Bryc K, Briggs AW, Stenzel U, Dabney J, Shendure J, Kitzman J, Hammer MF, Shunkov MV, Derevianko AP, Patterson N, Andres AM, Eichler EE, Slatkin M, Reich D, Kelso J, Paabo S

(2012) A high-coverage genome sequence from an archaic Denisovan individual. Science 338:222–226

Mikkelsen TJ et al (2005) Chimpanzee sequencing and analysis consortium, 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437:69–87

Miller DJ, Duka T, Stimpson CD, Schapiro SJ, Baze WB, McArthur MJ, Fobbs AJ, Sousa AM, Sestan N, Wildman DE, Lipovich L, Kuzawa CW, Hof PR, Sherwood CC (2012) Prolonged myelination in human neocortical evolution. Proc Natl Acad Sci U S A 109:16480–16485

Morales JC, Melnick DJ (1998) Phylogenetic relationships of the macaques (Cercopithecidae: Macaca), as revealed by high resolution restriction site mapping of mitochondrial ribosomal genes. J Hum Evol 34:1–23

Muller MN, Marlowe FW, Bugumba R, Ellison PT (2009) Testosterone and paternal care in East African foragers and pastoralists. Proc Biol Sci 276:347–354

Nadeau JH (2001) Modifier genes in mice and humans. Nat Rev Genet 2:165–174

Neubauer S, Gunz P, Weber GW, Hublin JJ (2012) Endocranial volume of Australopithecus africanus: new CT-based estimates and the effects of missing data and small sample size. J Hum Evol 62:498–510

Nunez R, Cooperrider K, Wassmann J (2012) Number concepts without number lines in an indigenous group of Papua New Guinea. PLoS One 7:e35662

Nuttle X, Huddleston J, O'Roak BJ, Antonacci F, Fichera M, Romano C, Shendure J, Eichler EE (2013) Rapid and accurate large-scale genotyping of duplicated genes and discovery of interlocus gene conversions. Nat Methods 10:903–909

O'Bleness M, Searles VB, Varki A, Gagneux P, Sikela JM (2012) Evolution of genetic and genomic features unique to the human lineage. Nat Rev Genet 13:853–866

Organ C, Nunn CL, Machanda Z, Wrangham RW (2011) Phylogenetic rate shifts in feeding time during the evolution of homo. Proc Natl Acad Sci U S A 108:14555–14559

Paabo S (2014) The human condition-a molecular approach. Cell 157:216–226

Pagel M (2009) Human language as a culturally transmitted replicator. Nat Rev Genet 10:405–415

Patterson N, Richter DJ, Gnerre S, Lander ES, Reich D (2006) Genetic evidence for complex speciation of humans and chimpanzees. Nature 441:1103–1108

Pemberton TJ, Absher D, Feldman MW, Myers RM, Rosenberg NA, Li JZ (2012) Genomic patterns of homozygosity in worldwide human populations. Am J Hum Genet 91:275–292

Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, Carter NP, Lee C, Stone AC (2007) Diet and the evolution of human amylase gene copy number variation. Nat Genet 39:1256–1260

Pollard KS, Salama SR, Lambert N, Lambot MA, Coppens S, Pedersen JS, Katzman S, King B, Onodera C, Siepel A, Kern AD, Dehay C, Igel H, Ares MJ, Vanderhaeghen P, Haussler D (2006) An RNA gene expressed during cortical development evolved rapidly in humans. Nature 443:167–172

Povinelli DJ, Preuss TM (1995) Theory of mind: evolutionary history of a cognitive specialization. Trends Neurosci 18:418–424

Prabhakar S, Visel A, Akiyama JA, Shoukry M, Lewis KD, Holt A, Plajzer-Frick I, Morrison H, Fitzpatrick DR, Afzal V, Pennacchio LA, Rubin EM, Noonan JP (2008) Human-specific gain of function in a developmental enhancer. Science 321:1346–1350

Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'Connor TD, Santpere G, Cagan A, Theunert C, Casals F, Laayouni H, Munch K, Hobolth A, Halager AE, Malig M, Hernandez-Rodriguez J, Hernando-Herraez I, Prufer K, Pybus M, Johnstone L, Lachmann M, Alkan C, Twigg D, Petit N, Baker C, Hormozdiari F, Fernandez-Callejo M, Dabad M, Wilson ML, Stevison L, Camprubi C, Carvalho T, Ruiz-Herrera A, Vives L, Mele M, Abello T, Kondova I, Bontrop RE, Pusey A, Lankester F, Kiyang JA, Bergl RA, Lonsdorf E, Myers S, Ventura M, Gagneux P, Comas D, Siegismund H, Blanc J, Agueda-Calpena L, Gut M, Fulton L, Tishkoff SA, Mullikin JC, Wilson RK, Gut IG, Gonder MK, Ryder OA, Hahn BH, Navarro A, Akey JM, Bertranpetit J, Reich D, Mailund T, Schierup MH, Hvilsom C, Andres AM, Wall JD, Bustamante CD, Hammer MF, Eichler EE, Marques-Bonet T (2013) Great ape genetic diversity and population history. Nature 499:471–475

Prufer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, de Filippo C, Li H, Mallick S, Dannemann M, Fu Q, Kircher M, Kuhlwilm M, Lachmann M, Meyer M, Ongyerth M, Siebauer M, Theunert C, Tandon A, Moorjani P, Pickrell J, Mullikin JC, Vohr SH, Green RE, Hellmann I, Johnson PL, Blanche H, Cann H, Kitzman JO, Shendure J, Eichler EE, Lein ES, Bakken TE, Golovanova LV, Doronichev VB, Shunkov MV, Derevianko AP, Viola B, Slatkin M, Reich D, Kelso J, Paabo S (2014) The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505:43–49

Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PL, Maricic T, Good JM, Marques-Bonet T, Alkan C, Fu Q, Mallick S, Li H, Meyer M, Eichler EE, Stoneking M, Richards M, Talamo S, Shunkov MV, Derevianko AP, Hublin JJ, Kelso J, Slatkin M, Paabo S (2010) Genetic history of an archaic hominin group from Denisova cave in Siberia. Nature 468:1053–1060

Reich D, Patterson N, Kircher M, Delfin F, Nandineni MR, Pugach I, Ko AM, Ko YC, Jinam TA, Phipps ME, Saitou N, Wollstein A, Kayser M, Paabo S, Stoneking M (2011) Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. Am J Hum Genet 89:516–528

Rilling JK, Glasser MF, Preuss TM, Ma X, Zhao T, Hu X, Behrens TE (2008) The evolution of the arcuate fasciculus revealed with comparative DTI. Nat Neurosci 11:426–428

Saglican E, Ozkurt E, Hu H, Erdem B, Khaitovich P, Mehmet M (2014) Heterochrony explains convergent testis evolution in primates. biorxiv:010553

Sakabe K, Wang Z, Hart GW (2010) Beta-N-acetylglucosamine (O-GlcNAc) is part of the histone code. Proc Natl Acad Sci U S A 107:19915–19920

Sakai T, Matsui M, Mikami A, Malkova L, Hamada Y, Tomonaga M, Suzuki J, Tanaka M, Miyabe-Nishiwaki T, Makishima H, Nakatsukasa M, Matsuzawa T (2013) Developmental patterns of chimpanzee cerebral tissues provide important clues for understanding the remarkable enlargement of the human brain. Proc Biol Sci 280:20122398

Salvucci E (2014) Microbiome, holobiont and the net of life. Crit Rev Microbiol 42(3):485–494. 1–10

Schoeninger MJ (2012) Palaeoanthropology: the ancestral dinner table. Nature 487:42–43

Semendeferi K, Teffer K, Buxhoeveden DP, Park MS, Bludau S, Amunts K, Travis K, Buckwalter J (2011) Spatial organization of neurons in the frontal pole sets humans apart from great apes. Cereb Cortex 21:1485–1497

Shapiro B, Hofreiter M (2014) A paleogenomic perspective on evolution and gene function: new insights from ancient DNA. Science 343:1236573

Sholtis SJ, Noonan JP (2010) Gene regulation and the origins of human biological uniqueness. Trends Genet 26:110–118

Sibley CG, Ahlquist JE (1987) DNA hybridization evidence of hominoid phylogeny: results from an expanded data set. J Mol Evol 26:99–121

Smail DL (2008) On deep history and the brain. University of California Press, Berkeley

Somel M, Franz H, Yan Z, Lorenc A, Guo S, Giger T, Kelso J, Nickel B, Dannemann M, Bahn S, Webster MJ, Weickert CS, Lachmann M, Paabo S, Khaitovich P (2009) Transcriptional neoteny in the human brain. Proc Natl Acad Sci U S A 106:5743–5748

Spikins P, Hitchens G, Needham A, Rutherford H (2014) The cradle of thought: growth, learning, play and attachment in Neanderthal children. Oxf J Archaeol 33:111–134

Stout D, Chaminade T (2012) Stone tools, language and the brain in human evolution. Philos Trans R Soc Lond Ser B Biol Sci 367:75–87

Subiaul F, Vonk J, Okamoto-Barth S, Barth J (2008) Do chimpanzees learn reputation by observation? Evidence from direct and indirect experience with generous and selfish strangers. Anim Cogn 11:611–623

Suwa G, Kono RT, Katoh S, Asfaw B, Beyene Y (2007) A new species of great ape from the late Miocene epoch in Ethiopia. Nature 448:921–924

Thelen E (1995) Motor development. A new synthesis. Am Psychol 50:79–95

Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, Mortensen HM, Hirbo JB, Osman M, Ibrahim M, Omar SA, Lema G, Nyambo TB, Ghori J, Bumpstead S, Pritchard JK, Wray GA, Deloukas P (2007) Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet 39:31–40

Tomasello M, Call J, Hare B (2003) Chimpanzees understand psychological states - the question is which ones and to what extent. Trends Cogn Sci 7:153–156

Toups MA, Kitchen A, Light JE, Reed DL (2011) Origin of clothing lice indicates early clothing use by anatomically modern humans in Africa. Mol Biol Evol 28:29–32

Ungar PS, Sponheimer M (2011) The diets of early hominins. Science 334:190–193

Varki A (2010) Colloquium paper: uniquely human evolution of sialic acid genetics and biology. Proc Natl Acad Sci U S A 107(Suppl 2):8939–8946

Varki A, Altheide TK (2005) Comparing the human and chimpanzee genomes: searching for needles in a haystack. Genome Res 15:1746–1758

Varki A, Brower DL (2013) Denial: self-deception, false beliefs, and the origins of the human mind. Twelve, New York

Varki A, Nelson D (2007) Genomic comparisons of humans and chimpanzees. Ann Rev Anthropol 36:191–209

Varki A, Geschwind DH, Eichler EE (2008) Explaining human uniqueness: genome interactions with environment, behaviour and culture. Nat Rev Genet 9:749–763

Varki NM, Strobert E, Dick EJJ, Benirschke K, Varki A (2011) Biomedical differences between human and nonhuman hominids: potential roles for uniquely human aspects of sialic acid biology. Annu Rev Pathol 6:365–393

Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS Biol 4:e72

Walker RS, Hill KR, Flinn MV, Ellsworth RM (2011) Evolutionary history of hunter-gatherer marriage practices. PLoS One 6:e19066

Wang X, Mitra N, Secundino I, Banda K, Cruz P, Padler-Karavani V, Verhagen A, Reid C, Lari M, Rizzi E, Balsamo C, Corti G, De Bellis G, Longo L, Beggs W, Caramelli D, Tishkoff SA, Hayakawa T, Green ED, Mullikin JC, Nizet V, Bui J, Varki A (2012) Specific inactivation of two immunomodulatory SIGLEC genes during human evolution. Proc Natl Acad Sci U S A 109:9935–9940

Warinner C, Speller C, Collins MJ, Lewis CMJ (2015) Ancient human microbiomes. J Hum Evol 79:125–136

Wells L, Vosseller K, Hart GW (2003) A role for N-acetylglucosamine as a nutrient sensor and mediator of insulin resistance. Cell Mol Life Sci 60:222–228

Wells JC, DeSilva JM, Stock JT (2012) The obstetric dilemma: an ancient game of Russian roulette, or a variable dilemma sensitive to ecology? Am J Phys Anthropol 149(Suppl 55):40–71

White TD, Asfaw B, Beyene Y, Haile-Selassie Y, Lovejoy CO, Suwa G, WoldeGabriel G (2009) Ardipithecus ramidus and the paleobiology of early hominids. Science 326:75–86

Wiessner PW (2014) Embers of society: firelight talk among the Ju/'hoansi bushmen. Proc Natl Acad Sci U S A 111:14027–14035

Winder IC (2014) The biogeography of the Papio baboons: a GIS-based analysis of range characteristics and variability. Folia Primatol (Basel) 85:292–318

Wlodarski R, Manning J, Dunbar RI (2015) Stay or stray? Evidence for alternative mating strategy phenotypes in both men and women. Biol Lett 11(2):20140977

Wood B, Leakey M (2011) The Omo-Turkana Basin fossil hominins and their contribution to our understanding of human evolution in Africa. Evol Anthropol 20:264–292

Chapter 2
Positive Selection in Human Populations: Practical Aspects and Current Knowledge

Pierre Luisi, Marc Pybus, Hafid Laayouni, and Jaume Bertranpetit

Abstract Natural selection targets a heritable trait that provides greater or lower chances for an organism to reproduce, and/or to survive, in a given environment. This evolutionary process is therefore directional: while an advantageous trait will be selected for and, thus, increase in frequency in the population, a prejudicial phenotype will be selected against and purged from the population. This concept, introduced in 1858 simultaneously by Charles R. Darwin and Alfred Wallace ((Darwin and Wallace J Proc Linnean Soc London 3:46–50, 1858); (Darwin On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. John Murray, London, 1859)), has been at the core of the study of

P. Luisi
Institute of Evolutionary Biology (UPF-CSIC), CEXS-Universitat Pompeu Fabra, Barcelona, Spain

Microbial Paleogenomics Unit, Institut Pasteur, Paris, France

Facultad de Filosofía y Humanidades, Universidad Nacional de Córdoba, Córdoba, Argentina

M. Pybus
Institute of Evolutionary Biology (UPF-CSIC), CEXS-Universitat Pompeu Fabra, Barcelona, Spain

Fundació Puigvert, Barcelona, Spain

H. Laayouni (✉)
Institute of Evolutionary Biology (UPF-CSIC), CEXS-Universitat Pompeu Fabra, Barcelona, Spain

Bioinformatics Studies, ESCI-UPF, Barcelona, Spain
e-mail: hafid.laayouni@upf.edu

J. Bertranpetit (✉)
Institut de Biologia Evolutiva, Universitat Pompeu Fabra, Barcelona, Catalonia, Spain
e-mail: jaume.bertranpetit@upf.edu

N. Saitou (ed.), *Evolution of the Human Genome II*, Evolutionary Studies,
https://doi.org/10.1007/978-4-431-56904-6_2

evolution and biological research. However, since then there has been passionate debate concerning its relative importance among other evolutionary processes, the prevalence of adaptive traits, and how they are originated in natural populations.

Since the recent wealth in genomics data, population and evolutionary geneticists have been able to interrogate the genome to understand the molecular basis of natural selection. In this chapter, we will focus on a particular mode of natural selection: positive selection also referred as adaptive selection or Darwinian selection. We describe statistical approaches to identify signals of positive selection and their practical challenges using genomics data. Then, we give a review on the current knowledge on positive selection in the human genome.

Keywords Positive selection · Adaptive selection · Hard sweep · Haplotype · Polygenic adaptation · Genome-wide selection scans · Complex adaptive traits · Selection on regulatory elements

2.1 Statistical Approaches to Identify Signals of Positive Selection

Charles R. Darwin and Alfred Wallace introduced at the mid-nineteenth century the concept of natural selection, focusing on phenotypic variation (Darwin and Wallace 1858; Darwin 1859). Since then, natural selection has been also widely studied at the genomic level, with a particular interest for positive selection. Positive selection refers to the process through which an allele that determines an advantageous trait will increase rapidly in frequency in the population, potentially until it reaches fixation. The allele frequency trajectory in the population through the action of positive selection depends on two main factors: the strength of the selective pressure and the number of generations since it started. The strength of positive selection is measured by the selection coefficient defined as the increased percentage of offspring of the individuals carrying the advantageous genotype in each generation as compared to individuals with alternative genotypes. A higher selection coefficient allows the advantageous allele to increase quicker in frequency, and thereby, to reach fixation in a shorter time. The speed of the increase tends to decline with the rise of frequency of the advantageous allele in the population since the relative advantage of individuals carrying the advantageous genotype declines with the frequency of their competitors. As a consequence, the allele frequency trajectory is non-linear and depends on the number of generations since the allele began to increase in frequency through the action of positive selection.

The shift in allele frequency comes with some typical molecular footprints used to detect selective events in the genome. Usually, we distinguish between two method families according to the kind of data analyzed:

- **Using divergence data**, i.e. sequences from different species, one can identify substitutions in the genome that are different across the species due to past selective events that contributed to the species divergence.
- **Using polymorphism data**, i.e. sequence or genotype data from different populations within the same species, to explore the nucleotide and haplotype diversity within and among populations.

The different molecular patterns left by a selective event are not maintained forever in the genome, and those footprints allow inferring how many generations have passed since the selective events occurred. In this chapter we will focus only on methods of detecting positive selection using polymorphism and the recent advances in methods developed to detect selection using both genotyping and sequencing data.

2.1.1 Using Polymorphism Data

In 1974, Maynard-Smith and Haigh (Maynard-Smith and Haigh 1974) proposed a model to explain the molecular mechanisms at play when positive selection acts on a variant. In this model, now referred to as the hard sweep model, they described the phenomena of genetic hitchhiking which results from positive selection driving a quick increase in frequency of an initially rare and beneficial allele toward fixation. This selective sweep occurs so quickly that recombination is not efficient to cut the haplotype where the selected variant arose, and thus, most of the variants carried by this haplotype also increase in frequency (Fig. 2.1). Therefore, under the hard sweep model, one expects a decrease in genetic diversity in the surrounding genomic region. The size of the region affected by such a sweep is proportional to the ratio

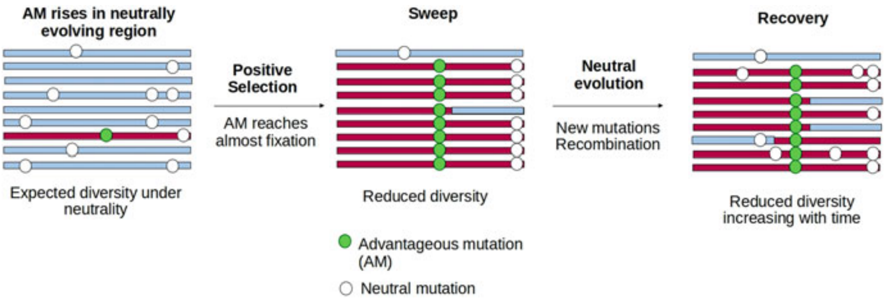

AM rises in neutrally evolving region
Expected diversity under neutrality
Positive Selection
AM reaches almost fixation
Sweep
Reduced diversity
Neutral evolution
New mutations Recombination
Recovery
Reduced diversity increasing with time
Advantageous mutation (AM)
Neutral mutation

Fig. 2.1 Molecular patterns in a genomic region suffering from a selective sweep. In a neutrally evolving region (before the selective sweep), an adaptive mutation (green circle) arises on one chromosome. During the selective sweep the frequency of the adaptive allele and its linked variants rapidly increase in frequency. After the sweep, the adaptive and linked alleles are fixed, and variability in the region is lost. During the recovery phase, new mutations begin to appear in different chromosome backgrounds by recombination and mutation restoring the diversity patterns

of the strength of selection and the rate of recombination (Barton 1998; Kaplan et al. 1989; Maynard-Smith and Haigh 1974). Thus, the reduction in levels of diversity within the genome is determined by the distribution of selection coefficients and the number of selective events in unlinked genomic regions. A selective sweep drives a quicker shift in allele frequency than what is expected under genetic drift. However, if recombination occurs, neutral alleles far from the selected site may not be driven to fixation, resulting in a temporary excess of high-frequency derived alleles at intermediate distance from the selected site (Fay and Wu 2000; Kim 2006; Przeworski 2002). Once the sweep is over, the genomic region enters a recovery phase during which it returns to neutral diversity levels through new mutations leaving a strong skew towards low frequency alleles persisting for many generations (Braverman et al. 1995; Kim 2006; Przeworski 2002). The strength and occurrence of sweeps can allow hitchhiking to dominate genetic drift, especially in large populations, and become the source of stochasticity for neutral alleles (Gillespie 2000; Kaplan et al. 1989; Maynard-Smith and Haigh 1974); this concept is known as genetic draft (Gillespie 2000). Maynard-Smith and Haigh formulated the theoretical background for most of the tests implemented thus far to detect signatures of selection at a molecular level using polymorphism data. A recently implemented database (Pybus et al. 2014) reports genome-wide scores for most of those tests ran on 1000 Genomes data in worldwide populations (The 1000 Genomes Project Consortium 2012), the latest publicly available polymorphism data. Those tests rely on three main features expected to be present in a genomic region surrounding a selected allele: long linkage disequilibrium (LD) haplotypes, a skewed Site Frequency Spectrum (SFS), and an excess of genetic differentiation among populations. The list of tests reported in this database (Pybus et al. 2014) is given in Table 2.1.

2.1.1.1 Tests Based on Long Haplotypes

Positive selection creates high levels of LD in the region surrounding the selected variant due to a quick shift in allele frequencies. For a given shift in allele frequency, less recombination events take place when there is a selective sweep than under genetic drift since the shift in allele frequency is much quicker in the former case. The Long Range Haplotype (LRH) test is commonly used to detect this signal (Sabeti et al. 2002a). However, this test does not take into account the recombination rate heterogeneity across the genome. To overcome this limitation, other tests have been implemented and are based on the Extended Haplotype Homozygosity decay (EHH, Sabeti et al. 2002a), which measures the decay of the haplotype homozygosity observed when moving away from the selected variant; this is caused by hitchhiking of a neutral allele (see Fig. 2.2 for a schematic representation of EHH decay calculation). The Cross-Population Extended Haplotype Homozygosity (XPEHH) compares the EHH decay observed in a population of interest to a reference (Sabeti et al. 2007). The integrated Haplotype Score (iHS; Voight et al. 2006) compares within the same population the EHH decay for the derived and ancestral alleles. Those two comparisons correct for recombination rate

Table 2.1 Statistics implemented by (Pybus et al. 2014) and are available in as UCSC tracks in the 1000 Genomes Selection Browser 1.0 at http://hsb.upf.edu/

Method family	Method	Reference
Site frequency Spectrum	Tajima's *D*	Tajima (1989)
	CLR	Nielsen et al. (2005)
	Fay and Wu's *H*	Fay and Wu (2000)
	Fu and Li's *D*	Fu and Li (1993)
	Fu and Li's *H*	Fu and Li (1993)
	R2	Ramos-Onsins and Rozas (2002)
Long haplotypes	XPEHH	Modified from Sabeti et al. (2007)
	ΔiHH	Modified from Voight et al. (2006)
	iHS	Modified from Voight et al. (2006)
	$\text{EHH}_{\text{average}}$	Modified from Sabeti et al. (2002a)
	EHH_{max}	Modified from Sabeti et al. (2002a)
	Wall's *B*	Wall (1999)
	Wall's *Q*	Wall (2000)
	Fu's *F*	Fu (1997)
	DH	Nei (1987)
	Za	Rozas et al. (2001)
	ZnS	Kelly (1997)
	ZZ	Rozas et al. 2001)
Population differentiation	F_{ST}	Weir and Cockerham (1984)
	XPCLR	Chen et al. (2010)
	ΔDAF	Hofer et al. (2009)
Descriptive statistics	Segregating sites	
	Singletons	
	π (nucleotide diversity)	Nei and Li (1979)
	DAF (Derived allele frequency)	
	MAF (Minor allele frequency)	

heterogeneity across the genome. Only recent selective sweeps (<30,000 years ago) can be characterized by the presence of long haplotype blocks because older sweeps have had time to shuffle the haplotype blocks, and are therefore not identifiable through this method.

2.1.1.2 Tests Based on Site Frequency Spectrum

The SFS is the representation of the number of alleles observed in a sample belonging to different frequency classes for a given set of polymorphic sites. Genetic hitchhiking around a selected allele will drive neutral alleles located nearby to high frequency leading to a reduced diversity, an excess of rare and derived alleles, and a

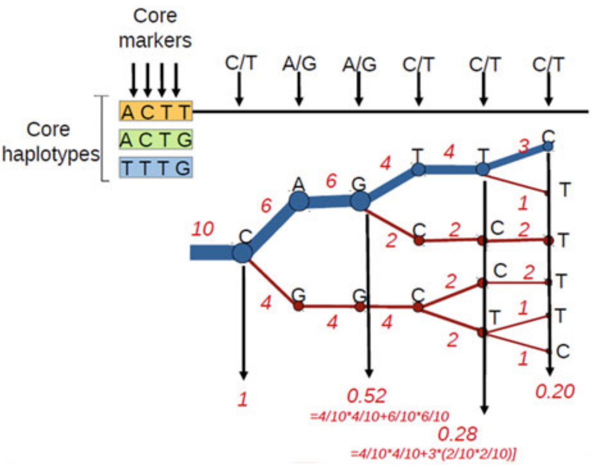


Fig. 2.2 Extended Haplotype Homozygosity decay. Moving away from the variant of interest, the haplotypes bifurcate and the haplotype carrying the core markers are less and less frequent. Thicknesses of the lines represent the frequency of the haplotype (haplotype counts in red). The haplotype homozygosity is given at the bottom

scarcity of alleles at intermediate frequency as compared to what is expected under neutrality (Fig. 2.3). The excess of rare alleles which persists for a long time during the recovery phase (up to ~250,000 years) can be formally tested by the famous statistic Tajima's *D* (Tajima 1989). Moreover, if the ancestral state of the variants is available, one can also test for the expected excess of high-frequency derived alleles (Fig. 2.3), with the Fay and Wu's *H* test (Fay and Wu 2000). This excess of rare alleles vanishes more rapidly as recombination allows neutral variants to evolve under genetic drift. This pattern can be detected for up to ~80,000 years after the sweep has occurred.

2.1.1.3 Tests Based on Genetic Differentiation

When a population faces a change in environment, positive selection may act on mutations that help the individual adapt better to this new environment. To detect the alleles responsible for local adaptation, one approach is to study genetic differentiation among populations. Traditionally the most used statistic is the fixation index, F_{ST}, first introduced by Sewall Wright which has been reformulated by multiple researchers. Using Cockerham and Weir's formula (Weir and Cockerham 1984), F_{ST} can be viewed as the proportion of genetic diversity due to allele frequency differences among populations:

$$F_{ST} = \frac{\sigma_{a^2}}{\sigma_{w^2} + \sigma_{b^2} + \sigma_{a^2}}$$

${\sigma_w}^2$, ${\sigma_a}^2$, and ${\sigma_b}^2$ are the intra-individual, inter-population, and within population inter-individual variances, respectively.

F_{ST} ranges from 0 to 1, with 0 signifying no differentiation (complete panmixia) and 1 indicating complete differentiation of the populations. Although high F_{ST} can

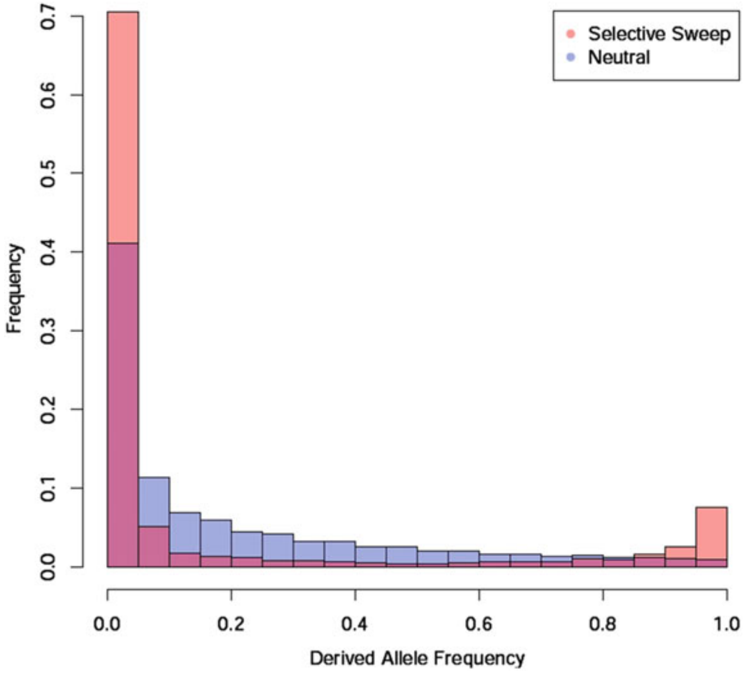


Fig. 2.3 Site Frequency Spectrum under different evolutionary models. The Unfolded SFS represents the number of derived alleles observed within different frequency classes. A region that has evolved under positive selection presents an excess of rare variants and of derived alleles at high frequency (red). During the recovery phase, the former pattern will remain due to new mutations arising in the region while the latter is lost more rapidly. Based on coalescent simulations of 100Kb regions evolving under neutrality (3000 neutral replicates in blue) and with a recent selective sweep driving an advantageous mutation up to fixation (300 selective sweep replicates in red), in an European-ancestry demographic model using COSI software (Schaffner et al. 2005)

putatively be attributed to the action of positive selection in one population, this approach is often criticized because of its sensitivity to population structure, demographic history, ascertainment bias, sample size, and minor allele frequency (for a review, see Holsinger and Weir 2009). The ΔDAF score (the differences of derived allele frequency between one population and a reference; Hofer et al. 2009) is another genetic differentiation index which suffers the same limitations. However, the use of the derived allele state allows identification of the population where positive selection has occurred. Further methods using genetic differentiation pattern have been developed. For example, the Cross-Population Composite Likelihood Ratio test (XPCLR) developed by Chen et al. (2010) relies on the comparison of a null model of genetic drift to one with a selective sweep by taking advantage of the genomic context around the selected allele in order to detect genomic regions with SFS differentiation among populations due to hitchhiking. This makes XPCLR more robust to demography and ascertainment bias than individual single nucleotide polymorphism (SNP) based methods such as F_{ST} and ΔDAF.

2.2 Practical Challenges in Detecting Positive Selection Using Polymorphism Data

Until recently, positive selection studies have been limited to sequence data from a restricted number of genes covering only a few thousands nucleotides. Now that detailed genetic maps are available in many human populations, it is possible to measure the signature of positive selection on a genomic scale using polymorphism data. Therefore, it is interesting to describe some potential challenges in detecting positive selection using polymorphism data and approaches to overcome them. First, detecting the different genomic footprints left by positive selection may be difficult in chip-array data. Second, those footprints may result from other mechanisms.

2.2.1 Distortions Due to Ascertainment Bias

Most genotype data used to study population diversity contain relatively important ascertainment bias. Ascertainment bias is the systemic distortion of the allele frequency spectrum due to a priori discovery of the polymorphisms segregating in a reduced sample. Thus, when genotyping individuals from other populations, especially those distant from the reference sample (the one where the initial genetic variants are described), it is not possible to catch all the genetic variation present in these populations.

Ascertainment bias is an intrinsic feature of genotyping technologies which are extensively used because they are simpler, cheaper, and much faster than sequencing approaches. The resulting genotype information for the population of interest will not be accurately produced for all the segregating sites but only for those present in the discovery sample. The probability of identifying a SNP is a function of its frequency, and as a consequence common SNPs are easier to detect. For example, many arrays use SNPs discovered in European samples, and, when used worldwide, the positions on the array are not polymorphic for the population of interest. Populations do not share all variation and some SNPs are private to particular populations (Casals and Bertranpetit 2012). The SNPs for newly designed arrays have been selected from public databases such as HapMap (www.hapmap.org) which in turn present an ascertainment bias of their own.

Usually, SNPs are selected to be genotyped in a population of interest with some of the following criteria: (1) having a Minor Allele Frequency (MAF) above a given threshold, usually relatively high in discovery samples representing either one or several populations of interest; (2) SNPs that are distant from one another by a given number of base pairs; (3) SNPs within targeted regions of interest; (4) SNPs maximizing the tagging of additional common SNPs that are in LD with them.

The criteria used affect the ascertainment bias, and it is difficult to assess a posteriori its extent when using genotyping arrays designed by others. Arrays with reduced ascertainment bias have been developed, for example, the Omni Family of Microarrays from Illumina which includes up to five million markers per sample and extensive coverage of new variants identified by the 1000 Genomes Project (The 1000 Genomes Project Consortium 2012), i.e. SNPs discovered through Next Generation Sequencing (NGS) in samples from worldwide populations; and Patterson and collaborators designed the Affymetrix Human Origins array with clearly documented ascertainment specifically for the study of population genetics (Patterson et al. 2012).

Ascertainment bias has a direct effect on many statistics that detect positive selection using polymorphism data (Thornton and Jensen 2007). First, and the most straightforward, SFS-based statistics are distorted by the artifact of the excess of common variants in genotyping arrays. Second, the tests based on genetic differentiation, such as the F_{ST} index, rely on a measure of genetic variance within and among the populations. Hence, if the SNPs genotyped within different populations present different ascertainment bias, the distribution of the index of genetic differentiation will be distorted. Haplotype-based statistics were developed in the first decade of this century with the goal to implement other methods less sensitive to ascertainment bias (Sabeti et al. 2007, Voight et al. 2006). These methods rely on an accurate estimate of LD patterns within a genomic region in order to infer whether there is a pattern of EHH (Granka et al. 2012). If the genotyping array only contains common variants and particularly chosen to *tag* the variability from another population (González-Neira et al. 2006), the observed LD patterns in the studied population are unlikely to be real. For example, in the data from the Human Genetic Diversity Panel (HGDP; Cann et al. 2002), for African populations the genotyping tag only 67% of SNPs with MAF above 5% and the power to detect positive selection is lower than for European samples, where 90% of such SNPs are tagged. It has been proved that haplotype diversity is more representative than individual SNP heterozygosity in the HGDP data (Conrad et al. 2006), suggesting that the ascertainment schemes affect more individual variants than haplotypes.

Nowadays, more studies obtain genotype information through NGS which does not suffer any ascertainment bias. However, the SFS is highly dependent on the coverage (read depth) used for sequencing. The power to detect rare variants increases with coverage (The 1000 Genomes Project Consortium 2012). Moreover, the genotype information may also depend on the sequencing center, its technology, and the SNP calling algorithm used. Therefore, for population genetics studies, one should be cautious when merging data from different datasets and control for the coverage across the genome.

2.2.2 The Confounding Factor of Background Selection

Background selection (BGS) is a process by which neutral variation is removed from the population when linked to deleterious variants (Charlesworth et al. 1993). Therefore, BGS reduces levels of polymorphisms in regions with many functional elements and low recombination. The lower level of polymorphisms in an extended region is often attributed as a result of positive selection because it is a molecular pattern expected under the hard sweep model. It is consequently important to correct for BGS. One straightforward approach when analyzing protein-coding regions is to look for lower levels of neutral variation near functional substitutions, i.e. at functional sites where a mutation has been fixed in a set of species, which is evidence for positive selection while not being expected under BGS. However, this approach is biased towards protein-coding regions, and would just detect events of positive selection acting on mutations with a priori known function. An alternative to this approach would be to correct for several genomic variables that correlate with BGS, such as levels of recombination rate and functional constraint. Measuring functional constraint is not straightforward but one can use the density of coding sequences (CDS), conserved coding sequences (CCDS), conserved non-coding sequences, and untranslated regions (UTRs). Moreover, Enard et al. recently found that GC content presents a strong correlation with levels of neutral diversity (Enard et al. 2014). Although BGS has been seen as mimicking positive selection at a molecular level, after the article by Charlesworth, Morgan, and Charlesworth (1993), tests based on LD—namely XPEHH and iHS—show insensitivity to BGS (Enard et al. 2014; Fagny et al. 2014), and therefore, their extreme deviations may directly be attributed to recent hard sweeps.

2.2.3 Demography Can Mimic Positive Selection

Many neutral mechanisms can affect the genetic diversity in populations or species, among which several demographic processes can lead to molecular patterns expected under a positive selection scenario (Table 2.2).

2.2.3.1 Migration and Structure

The neutral model assumes that any cross-gender individual pair has the same probability to reproduce in the population. However, there may be population subdivision due to geographic distance, social, linguistic, or economical barriers (e.g., in India with the caste system). Barriers to random mating are likely not to be absolute, and a number of migrants can move between subpopulations each generation. When hidden population subdivision is occurring and panmixia is improperly assumed, the genetic variability is higher than expected with an excess of variants at

Table 2.2 Some demographic processes can leave molecular patterns expected under positive selection

Process	Description	Molecular pattern
Migration	Individuals move from one population to another	Increased genetic variability within each population and lower genetic differentiation among populations
Isolation	One population is isolated from the others and drifts on its own	Increased genetic differentiation among populations
Population structure	The studied population is actually structured into several subpopulations	Higher variability than expected
Population expansion	The population increases rapidly in size	Increased number of rare variants and decreased variability
Population bottleneck	The population decreases rapidly in size and rebounds to its original size after several generations	Increased number of rare variants and derived alleles at high frequency with decreased variability
Founder effect	A new population is founded by a small number of individuals from a larger population and the new population then increases in size	Gene surfing: Mutations that occur on the frontier of a growing population are more likely to expand and get fixed since only a few individuals are founding the population

intermediate frequency. Migration from an external population causes a higher variability with an excess of rare variants is expected.

2.2.3.2 Population Expansion

During population expansion, a new generation has a greater number of individuals than the previous one. A well-described human population expansion event occurred after the Neolithic transition. One possible cause is that the agriculturist way of life may have provided a more reliable mode of sustenance and allowed settlements to increase in size. A population with expansion will show an excess of singletons at low frequency as compared to a population with constant size due to recent mutations which have not increased in frequency through genetic drift, and remain almost individual specific (Keinan and Clark 2012). This also implies a lower genetic variability than expected for the population size.

2.2.3.3 Population Bottleneck

A bottleneck is the phenomena through which population size decreases suddenly, followed by a recovery, or increase, of the original population size in a few generations. One striking example is the Black Death plague faced by Asian and European populations in the fourteenth century. Plague is thought to be responsible for several large epidemics with death rates of up to 30–50% of the European population and lingering thereafter in Europe for several centuries (McEvedy 1988). Many alleles from the original population, mostly at low frequency, will either disappear or become very frequent during the decreasing size phase, thus reducing the genetic variability. During the recovery phase, as in population expansion, an excess of rare variants will arise.

2.2.3.4 Founder Effect

A founder effect occurs when a small subpopulation leaves its former habitat to establish a new one. This can be seen as a particular case of a bottleneck. Modern humans likely colonized geographic areas out-of-Africa through several founder effects (Reich et al. 2001). One more recent example would be the colonization of Quebec, Canada ~400 years ago by ~8500 French settlers. Such event allows variants to rapidly reach fixation through genetic drift, a phenomena called gene surfing (Hallatschek and Nelson 2008), which mimics genetic hitchhiking.

2.2.4 Has a Region of Interest Evolved Under Positive Selection?

One major challenge in assessing whether a region of interest has evolved under the action of positive selection is to circumvent the confounding factors of past demographic processes as well as data ascertainment bias. For that purpose, one can compute the statistic designed to detect footprints of positive selection and estimate its significance by comparison to a reference distribution. This reference distribution must reflect the expected score under selectively neutral evolution with the data used. Indeed, values of statistics are not absolute but are relative to the studied population and to the kind of data analyzed. There are two main approaches to defining reference distributions: simulations and the outlier approach.

2.2.4.1 Using Simulations Accounting for Demography

Since the development of coalescence theory (Hudson 1991; Kingman 1982; Wakeley 2008) and the recent wealth in computational capacity, simulations have become a powerful approach in population genetics. It is now possible to generate large independent data sets through simulations of genetic data that mimic population demographics. Those data sets are, in turn, used to assess the statistical significance of empirical data accurately. Particularly, one can simulate sets of genetic data under a neutral model with appropriate demographic parameters to infer what the empirical data would look like without the action of positive selection, and then, a significance threshold at a given false positive rate (FPR) can be estimated. In this case, any putative biases from empirical data are eliminated. Furthermore, in order to evaluate the reliability of the estimated threshold, simulations can incorporate selective events to the neutral model to infer the power of the approach.

The simulation software that has been implemented so far can be divided into those based on coalescent theory and on forward simulation. Coalescent simulation is the first approach widely used to simulate genetic data at the sequence level and, as the name suggests, is based on coalescent theory. First introduced by John Kingman in 1982 (Kingman 1982), it relies on a backward model describing the characteristics of joining lineages back in time to the most recent common ancestor (MRCA). It represents the theoretical background for most neutral genetic models, as well as the estimation of many population genetic parameters. The coalescence theory provides computational efficiency with several coalescence simulation software available, such as *FastCoal* (Marjoram and Wall 2006), *CoaSim* (Mailund et al. 2005), *SelSim* (Spencer and Coop 2004), *cosi* (Schaffner et al. 2005), *ms* (Hudson 2002), and *msms* (Ewing and Hermisson 2010).

For many of the underlying coalescent models, parameters have been calibrated to fit empirical data in order to retrieve the past demographic history of human populations. For example, Schaffner et al. used HapMapIII data to infer the demographic history of three populations through the calibration of their model by making the simulated data match empirical data for pairwise F_{ST} values, LD decay (how LD for pairwise SNPs decreases with physical distance in the genome), and SFS (Schaffner et al. 2005). Further implementations used more complex empirical data features, such as the joint SFS across populations (Gravel et al. 2011). Those programs simulate genomic regions spanning a few megabases in hundreds of samples without large computational costs in time or resources. This is particularly useful when computing large simulated distributions of the statistics to estimate the statistical significance for a genomic region. However, coalescent simulations present several limitations. Most importantly they have limited accuracy in simulating the number of recombination and gene conversion events, and the ability to implement possible recombination patterns. As a consequence, a realistic recombination map incorporated into the model increases the computational cost and therefore reduces the size of the simulated region. With a simplistic recombination map, the simulated genomic regions can be longer but the model is unlikely to be accurate. Another traditional issue with coalescent simulations is the incorporation of selective events. Attempts to improve coalescent simulations (Ewing and Hermisson 2010; Grossman et al. 2010; Spencer and Coop 2004) have usually come at the cost of over-simplifying other aspects of the model such as recombination map, population changes, sample size, and length of the simulated genomic regions.

To circumvent the limitations, the forward simulation approach has been proposed as an alternate. Genomic data is simulated forward in time from an ancestral status, allowing more flexibility to the model including complex recombination patterns and other genomic features (gene content, background selection; for an example, see *SFS_CODE,* Hernandez 2008). The demographic processes included in the model can also present a much higher layer of complexity (e.g., see *dadi,* Gutenkunst et al. 2009; Uricchio and Hernandez 2014). However, these approaches require the simulation of whole populations and, therefore, are very computationally expensive, preventing the generation of large data sets. For a neutral model of human demography, Excoffier and colleagues implemented a coalescent model, *fastsimcoal2*, which allows for a high level of demographic complexity, with serial founder effects, range expansions, and admixture among populations (Excoffier et al. 2013). This model overpasses forward simulation models such as *dadi (*Gutenkunst et al. 2009) which is arguably the reference in the field. The models are calibrated to make the simulated data fit the empirical data. Therefore, when the empirical data contains ascertainment bias, it is important to either correct for it (Nielsen et al. 2004) or to take it into account in the estimation procedure (Pickrell et al. 2012; Wollstein et al. 2010). Although it is not an easy task the calibrated model can inaccurately reflect past demography if ignored (Excoffier et al. 2013). In addition, most models rely on a priori assumptions on demographic events and therefore accurate models are available for a reduced number of well-studied populations.

2.2.4.2 Outlier Approach

As mentioned before, constructing a neutral model using simulations is computational expensive and the model is not likely to incorporate all the layers of demographic and genomic complexity. One may prefer to use the outlier approach: an empirical distribution of statistics to detect positive selection built from a large number of loci across the genome. The loci located in the extreme tail(s) of the distribution, i.e. outliers, are considered as possible targets of positive selection. The assumption behind this framework is that demography stochastically affects the whole genome evenly while positive selection, a deterministic process, affects only a few loci and does not distort the distribution. This approach also allows correction for ascertainment bias and the confounding effect of background selection, as long as the reference loci are accurately sampled. It is important to note that the genome can be seen as a mosaic of several chunks, each with its own history, and although the population definition is accurate, the chunk demographics may be very different with some specific genomic regions exhibiting extreme molecular patterns that mimic positive selection. This may be inaccurately identified as under positive selection resulting in false positives (Kelley et al. 2006). Inaccuracies occur particularly in the case of positive selection targeting recessive alleles, standing variation, and population bottlenecks (Teshima et al. 2006). Another difficulty of the outlier approach is the arbitrary threshold used to consider a score's significance. Setting thresholds require a priori definition of the proportion of the genome expected to be under positive selection. For example, if the 5% most extreme scores are considered to be under putative positive selection, the underlying assumption is that 5% of the genome is expected to be under selection. However, no accurate estimate is available for many organisms and it remains one of the main questions in studying positive selection. Finally, the outlier approach only identifies the most extreme case of positive selection, many of the selected alleles, especially those with a relatively low selection coefficient, are likely to be false negatives.

2.2.4.3 Combination of Different Tests

Assessing statistical significance for a given score through either simulations or the outlier approach is necessary to determine whether a genomic region has been evolving under positive selection. However, it is delicate to make sure that a significant score is not actually a false positive and especially difficult when drawing conclusions from only a single method that a locus has been targeted by positive selection. To reduce the risk of false positives, it is wise to use different methods developed to detect the impact of positive selection at a molecular level. Particularly, one may use methods based on different kinds of molecular footprints left by a selective sweep (SFS, LD, and genetic differentiation). This way, the false discovery rate is likely to be reduced: the false positives from individual methods are unlikely to overlap, since each method is sensitive to different demographic processes. Zeng et al. implemented two compound tests, DH (Zeng et al. 2006) and DHEW (Zeng et al. 2007), which combine the SFS-based methods Fay and Wu's H and Tajima's D specifically the Ewens–Watterson test (Watterson 1978) for DHEW. The underlying idea of DH is that Fay and Wu's H and Tajima's D are sensitive to population bottlenecks and expansions, respectively (Zeng et al. 2006), while insensitive to the other demographic process. Thus, combining the two tests is robust to both demographic processes. The idea is very simple; using neutral simulations, a significance threshold for both tests is set for a given FPR. Afterwards, if a region of interest is significant for both tests, it is identified as a target of positive selection. The original method relies on neutral simulations with rather simplistic demography using *ms*, but the framework suggested by Zeng et al. can extend it to an outlier approach as in Luisi et al. (2015) where Fay and Wu's H and Tajima's D are computed in a large number of genomic regions to make the reference distribution and estimate of the join threshold significant.

A more simplistic method is to use any combination test, i.e. a test that combines K individual test's P-values, such as the Fisher combination's test:

$$\mathrm{Z_F} = -2\sum_{\mathrm{i=1}}^{\mathrm{K}} \log \mathrm{P_i}$$

where P_i is the P-value associated to the score of the i^{th} test.

Following this idea, Grossman et al. implemented a Composite Multiple Score (CMS; Grossman et al. 2010) which multiplies P-values of five individual tests based on long haplotypes—XPEHH, ΔiHH, and iHS—and genetic differentiation—F_{ST} and ΔDAF. The main improvement from a rather simplistic combination score is that they computed P-values from simulations using the demographic model calibrated by Schaffner et al. (2005) under a neutral scenario and with a selective event. Then, the CMS is obtained as the following:

$$\mathrm{CMS} = \frac{\prod_{i=1}^{5} P(s_i \mid \text{selected}) \times \pi}{P(s_i \mid \text{selected}) \times \pi + P(s_i \mid \text{unselected}) \times (1 - \pi)}$$

where s_i is the score of the i^{th} method, the P-values are obtained from reference distributions from simulations under either neutral (*unselected*) or selective scenarios and π is the uniform prior probability of selection.

CMS and other combination tests (e.g., Fisher's combination) cannot use any kind of individual tests since they rely on the assumption of the independence among tests. Moreover, they attribute equally to the combined score. In Pybus et al. (submitted), an alternative framework, *Boosting*, incorporates the information from different methods. Based on *Boosting* functions (Lin et al. 2011), this framework allows detection and classification of selective events. *Boosting* is a Support Vector Machine (SVM; Schapire 1990) which is trained on simulated data to estimate the best regression function of scores from different individual methods to distinguish between two scenarios. The algorithm begins with a neutral demographic model (Schaffner et al. 2005) to which a selective sweep scenario can be incorporated (Grossman et al. 2010); thousands of genomic regions have been simulated under a selectively neutral scenario and 45 selective ones. Then, two *Boosting* functions have been trained to distinguish among the scenarios, (1) evolution under either pure genetic drift or with a partial selective sweep (where the selected mutation reaches a final allele frequency (FAF) of 0.2 or 0.4); (2) evolution with an incomplete selective sweep (FAF = 0.6 or 0.8); and (3) evolution with a complete sweep (FAF = 1). Two further boosting functions have been built to classify regions evolving under a complete or incomplete sweep into recent or ancient sweep categories (Fig. 2.4). Those functions are included in a classification tree as shown in Fig. 2.5. This framework uses the combination of different, although relatively correlated, tests to classify the mode of positive selection for the detected selective events. As seen in Fig. 2.6, the standardized coefficients for each test give valuable insight into the methods that contribute the most when distinguishing between two given scenarios, and thus, on their ability to detect a given selective event. Moreover, the boosting coefficients are quite similar for the three populations analyzed (African (AFR), European (EUR), and Asian (ASN)), and thus seems quite robust to demography.

2.2.5 Selection Not Only by Hard Sweep

On the one hand, clear evidence of morphological and physiological adaptations in modern human populations exists, such as pigmentation for solar radiation, body size for thermal condition, and blood flow and oxygen delivery for high altitude. On

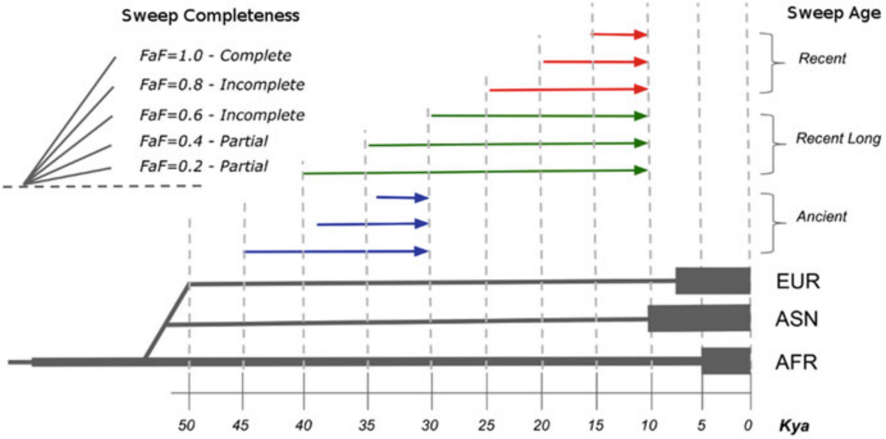


Fig. 2.4 Simulation scenarios. Simulations were run following a calibrated human demography that resembles population genetic data from three reference continental populations (European (EUR), Asian (ASN), and African (AFR), from left to right) [45]. Nine different time-spanning selective sweeps were simulated (grouped as Neutral, Recent, Recent Long, and Ancient) allowing for five different final allele frequencies (FaF = 0.2, 0.4, 0.6, 0.8, and 1.0)

the other hand, there are few examples of fixed, or almost fixed, genetic differences among populations and/or validated cases of adaptive mutations (see Sect. 3 for an overview). Moreover, Hernandez et al. (2011) showed that hard sweeps may have been rare during human evolution (but see Sect. 3.3.3). This striking inconsistency between the number of known phenotypic and genotypic adaptive examples may be explained by the simplistic way that positive selection has been researched. Indeed, until now, most studies of natural selection relied on the hard sweep model making use of methods designed to detect molecular patterns expected to remain in the genome under this model. In order to have a complete picture of adaptation and its genomic processes, it is important to consider other modes of positive selection. The other types of positive selection do not leave the same molecular footprints as a hard sweep. These alternate modes of positive selection require theoretical development but are beginning to be studied after being overlooked for many decades (Pritchard et al. 2010).

2.2.5.1 Soft Sweep

Recently, empirical (Colosimo et al. 2005; Hamblin and Di Rienzo 2000; Jeong et al. 2008; Scheinfeldt et al. 2009; Tishkoff et al. 2007) and theoretical (Hermisson and Pennings 2005; Innan and Kim 2004; Orr and Betancourt 2001; Pennings and Hermisson 2006; Przeworski et al. 2005) studies indicate the importance of soft sweeps which can occur through two different modes of adaptation:

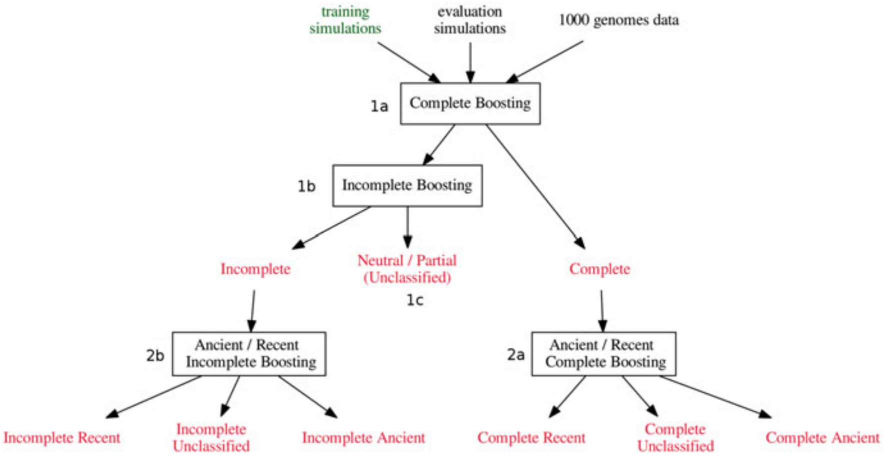


Fig. 2.5 Implemented classification tree. The implemented classification tree was organized in two levels: an unknown genomic region is firstly classified according to the completeness of the sweep, as being Complete, Incomplete, or Unclassified. In the second step, it is then classified according to the age of the sweep, being Ancient, Recent, or Unclassified. The algorithm can be described as following: (1a) If the Complete Boosting score is above the 99th percentile of the distribution of the Complete Boosting scores for the training simulations under the Neutral, Partial, and Incomplete scenarios, the region is classified as Complete and go to step 2a, otherwise go to step 1b. (1b) If the Incomplete Boosting score is above the 99th percentile of the distribution of the Incomplete Boosting scores for the training simulations under the Neutral and Partial scenarios, the region is classified as Incomplete and go to step 2b, otherwise go to step 1c. (1c) If not classified at iteration 1a or 1b, the genomic region is left unclassified and the algorithm stops. (2a) If the Ancient/Recent Complete Boosting score is above the 99th percentile of the distribution of the Ancient/Recent Complete Boosting scores for the training simulations under the Complete Recent scenario the region is classified as Complete Ancient, while if it is below the 1st percentile of the distribution of the Ancient/Recent Complete Boosting scores for the training simulations under the Complete Ancient scenario the region is classified as Complete Recent, otherwise the region remains only classified as Complete. (2b) If the Ancient/Recent Incomplete Boosting score is above the 99th percentile of the distribution of the Ancient/Recent Incomplete Boosting scores for the training simulations under the Incomplete Recent scenario the region is classified as Incomplete Ancient, while if it is below the 1st percentile of the distribution of the Ancient/Recent Incomplete Boosting scores for the training simulations under the Incomplete Ancient scenario the region is classified as Incomplete Recent, otherwise the region remains only classified as Incomplete

- *Selection on a standing variant.* In opposition to a hard sweep, selection on a standing variant does not rely on the appearance of an advantageous mutation to arise in the population, but rather targets a variant already segregating at a relatively important frequency when a change of environment occurs.
- *Selection on recurrent mutation.* For selection on recurrent mutations to occur, the derived and advantageous allele arises in the population several times independently, as a result of recurrent mutations or gene flow from another population. All copies of the derived allele increase in frequency until the allele reaches fixation. However, if all copies of the derived allele have similar selective

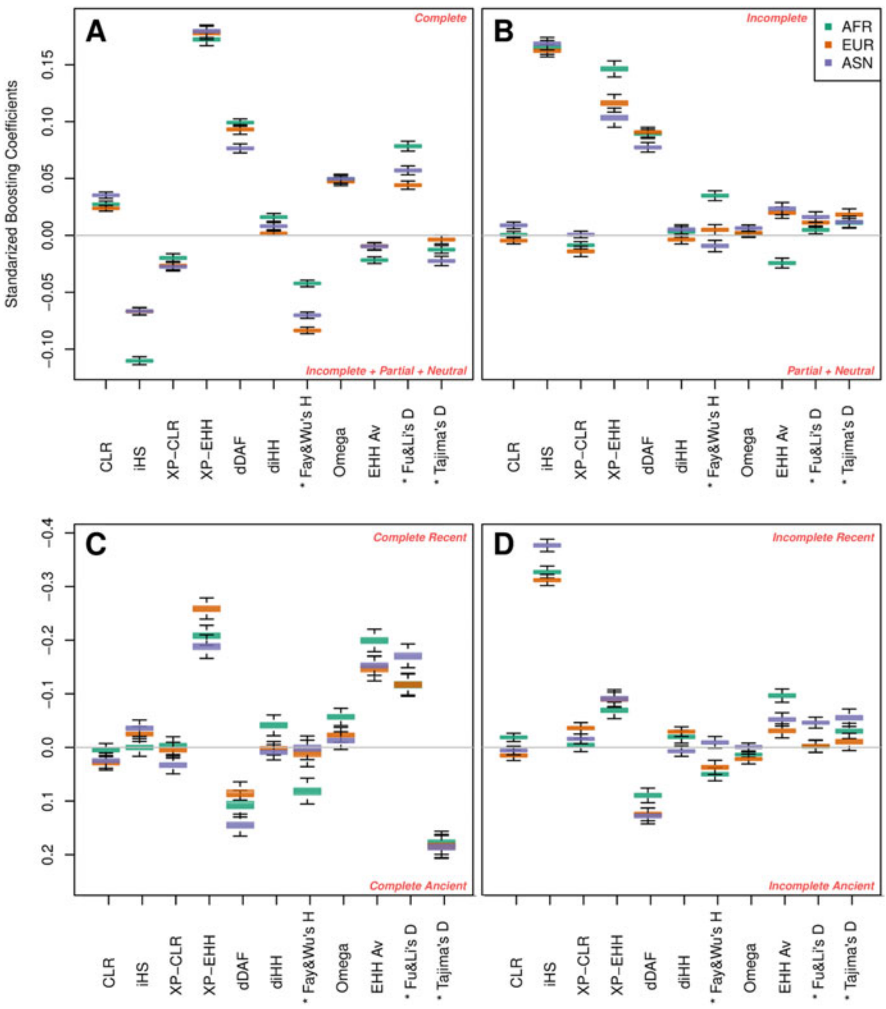


Fig. 2.6 Standardized coefficients for the European (EUR), Asian (ASN), and African (AFR) populations and for each implemented boosting function. Estimated coefficients for each population in the four boosting functions used in the classification tree: Complete (**a**), Incomplete (**b**), Complete Recent/Ancient (**c**), and Incomplete Recent/Ancient (**d**). The relevance of the positive selection tests to classify the different scenarios is given by the strength of its standardized coefficient

coefficients (because the genetic background has no affect through, for example, intragenic epistasis), none of the haplotype carrying one of these copies will fix during the selective event (Hermisson and Pennings 2005; Pennings and Hermisson 2006). Actually, diferent haplotypes, each carrying one copie of the advatageous allele, will increase in frequency until the allele has fixed in the population.

In both cases different copies of the selected allele may belong to different haplotypes: in the case of standing variation it was already segregating on different haplotypes before the selective event, while in the recurrent mutation case, it arose on different haplotypes. In both cases tests based on long haplotypes are not suited to detect this mode of adaptation. However, if the selective pressure is population specific, methods based on genetic differentiation may be able to detect it. In addition, other haplotype patterns, beyond the EHH, can be informative (see below).

2.2.5.2 Polygenic Adaptation

Recent genome-wide association studies (GWASs) confirm the view of classic quantitative genetics that many phenotypes are encoded by several dozens, hundreds, or even thousands of genes, rather than a unique one (Fu et al. 2013). This drastically contrasts with the idea that positive selection acts on a single advantageous mutation to drive phenotypic adaptation. Therefore, more focus on polygenic adaptation is required. Such a mode of adaptation would simultaneously cause a limited shift in allele frequencies at several variants located in different genomic regions and have small effects on fitness. This pattern is extremely difficult to distinguish from pure genetic drift.

2.2.5.3 Recent Methodological Advances in Detecting Alternative Sweep Scenarios

The molecular patterns expected to be left by soft sweeps and polygenic adaptations are not as evident as those left by hard sweeps (Fig. 2.1). Therefore, a lack of methods designed to detect such selective events at the genetic level exists. However, ongoing methodological development is in progress. Some already existing methods can be used to detect soft sweeps. Indeed, as mentioned above, if the selective pressure is population specific, a locus-based statistic of genetic differentiation (e.g., F_{ST}) may be powerful provided the variant is segregating at low frequency in the reference populations. iHS shows sensitivity when positive selection acts on a standing variant that was segregating at low frequency before the selective event (Ferrer-Admetlla et al. 2014). Two other methods relying on specific haplotype patterns have been recently developed (Ferrer-Admetlla et al. 2014; Garud et al. 2014). First, nS_L (Ferrer-Admetlla et al. 2014) is based on the comparison between EHH for derived and ancestral alleles, as in iHS, but also takes into account the length of the segment of haplotype homozygosity between a pair of haplotypes. Besides showing greater power than iHS for scenarios where the advantageous allele was already present in the population at frequency $> 3\%$, it does not need any genetic map and is robust to recombination rate and mutation rate. Second, the H12 and H2/H1 statistics (Garud et al. 2014) also rely on homozygosity of multiple haplotypes. H12 use the combined frequency of the first and second most frequent haplotypes observed in a genomic region as the following:

$$\mathrm{H12} = (p_1 + p_2)^2 + \sum_{i>2} p_{i^2}$$

where p_i is the frequency of the i^{th} most common haplotype in the sample.

The H12 statistic has power to detect hard sweeps and—not so—soft sweeps, i.e. when the starting frequency is below 0.1%. In order to distinguish between those two scenarios Garud et al. further developed the H2/H1 statistics (Garud et al. 2014):

$$\mathrm{H2/H1} = \frac{\sum_{\mathrm{i} \geqslant 2} p_{1^2}}{\sum_{\mathrm{i} \geqslant 1} p_{1^2}}$$

where p_i is the frequency of the i^{th} most common haplotype in the sample.

While H1 is expected to be higher under the hard sweep model, H2 is expected to be higher under the soft sweep scenario. Therefore H2/H1 increases with the softness of the sweep, i.e. the number of haplotypes on which the advantageous mutation is segregating prior to the selective event.

Those two recent methods demonstrate that accurate theoretical implementation allows detection of soft sweeps despite difficulty in recognizing the molecular patterns. Further theoretical work is required to increase the power to detect even softer sweeps. Despite the fact that the reduced shift in allele frequency expected under polygenic adaptation leaves very weak footprints in the genome, it could be argued that increasing the sample size would increase the power, and implementing methods using only genetic information seems a losing battle. For this reason, the few methods that have been proposed include other kinds of information. First, the BayENV (Coop et al. 2010; Günther and Coop 2013) method uses environmental variables. It is based on the correlation between allele frequency and an environmental variable observed in many populations. For each locus, it provides a Bayes Factor which is the ratio between two Bayesian posterior probabilities:

- Under the null (neutral) model, the correlation we observe in allele frequencies between different populations is just explained by demographic factors (genetic drift, migration, and population size changes).
- Under the model where a specific environmental variable has caused a selective pressure in (a) population(s) it may have caused an imbalance in the allele frequency spectrum across the populations.

Therefore, this method detects variants that shifted similarly in allele frequency in populations facing the same environmental pressures compared to their neighboring populations. Parallel selection, recently theoretically analyzed by Ralph and Coop (2010), is more likely to occur on ancient variants that are shared among worldwide populations. Note that the signal of selection is driven by the shift in allele frequency across populations rather than by its amplitude. This method corrects for population structure and therefore is less sensitive to demography than a simple correlation analysis. Indeed, the genetic differentiation among populations is directly related to their geographic distance (Gutenkunst et al. 2009) due to the isolation by distance phenomena. However, retrieving environmental variables from many populations may be challenging, especially because it relies on representative geo-localization.

An approach suggested by Mendizabal et al. (2012) includes phenotypic information rather than selective pressure. More precisely, the authors have analyzed the covariance between allele frequencies and height measurements to detect genetic variants allowing Pygmy adaptation to the rainforest climate by better thermo regulating with size reduction which is known as Bergmann's rule. This approach would require extensive phenotypic measurements but the authors implemented a permutation procedure that only requires the average and variance of the phenotype found in literature. This method can detect advantageous variants only if the phenotype is hypothesized to be the result of an adaptive process.

H. Allen Orr suggested a sign test (Orr 1998), to determine whether the observed number of plus (or minus) alleles at Quantitative Trait Loci (QTLs) is different in two groups of individuals with different phenotypes, instead of being similar as expected under genetic drift. Orr's sign test has recently been used for expression QTLs (eQTLs) where polygenic adaptation can be indicated by the accumulation of many eQTL even if each eQTL has low effect on the phenotype (Fraser et al. 2011). Similarly, an alternative is to use a set of SNPs associated with a given phenotype, e.g. height in European populations (Turchin et al. 2012), and show systematic allele frequency differences between populations with different phenotypic values that better fit a model of adaptive evolution than genetic drift. Finally, Berg and Coop (2014) have implemented a test using the mean additive genetic value, Q_X, estimated from the additive effect size of loci associated with a given phenotype (GWAS loci). The test is an extension of the BayENV method and determines whether the genetic value (instead of the allele frequency) covaries with a given environmental variable. They further developed a generalization of the Q_{ST}/F_{ST} comparison (Leinonen et al. 2013). The Q_{ST}/F_{ST} test of neutrality contrasts whether there is an excess of quantitative trait differentiation (as measured by the Q_{ST} index) to the genetic differentiation among populations (as measured in a large set of loci by the F_{ST} index), to identity traits that have evolved adaptively. In their implementation Berg and Coop (2014) use the estimated Q_X instead of Q_{ST}.

The theoretical development to identify variants with a small effect on fitness but the basis of phenotypic adaptation through polygenic adaptation is progressing. However, most of the methods rely on GWAS loci, and as a consequence, are still limited. First they assume that the associated loci act in a strictly additive manner, ignoring the putative dominance or epistasis among them. Second, GWAS loci are unlikely to be the causal ones, but rather tag the true positives; since the LD patterns are variable among populations, the GWAS loci may not be a good proxy of the causal variant in all the studied populations. Third, the genetic values are relatively accurate when calculated in a population where the association studies were performed, but the GWAS loci may not be portable to all genetic backgrounds.

2.2.6 From Putative Advantageous Mutation to Increased Fitness

Most studies attempt to identify advantageous mutations. This goal may be reached if, at least, the four following steps are completed.

1. *Identify candidate adaptive loci*. The main issue is to disentangle whether a strong statistical signal for detecting positive selection is truly due to positive selection or alternative processes aforementioned.
2. *Identify the underlying functional variant*. Strong LD within a genomic region with hitchhiking must be removed in order to pinpoint the variant targeted by positive selection.
3. *Quantify the phenotypic consequences of the candidate adaptive allele* by performing experiments in vivo with model organism (mouse, zebrafish, etc...), in vitro using cell cultures, or genotype–phenotype association studies. An alternative is to use the wealth of functional public databases to retrieve information from the literature.
4. *Clarify the relationship between phenotype and reproductive fitness* in the population and environment where the allele has increased in frequency. This is a complicated task because one must infer the relevant environment which selected the variant in the ancestors of the studied population, and whether the phenotypic change encoded by the functional variant is fitter than the ancestral one.

Few studies present conclusive results from the four steps together. Particularly, the fourth step may result in story-telling and it is impossible to formally test such relationship in humans. Therefore, it is important not to dismiss the possibility that a locus is adaptive despite the inability to determine the past selective pressures and to demonstrate that the phenotypic change resulted from an increase in fitness in past populations.

In the future, the recent wealth in *omics* data will most probably allow partially to bridge the gap between genotype and phenotype when studying adaptive evolution. Indeed, thanks to NGS data, functional data has been produced in the past few years in epigenomics, metabolomics, transcriptomics, and interactomics, among others. For example, the Encyclopedia of DNA Elements (ENCODE) project (Dunham et al. 2012) has identified functional elements across the genome, in coding and non-coding regions. In order to identify the underlying functional variant, one may use this emerging functional data, for example, through an integrative genomics approach, along with results from population genetics of positive selection (Barrett and Hoekstra 2011; Scheinfeldt and Tishkoff 2013).

2.3 Current Knowledge on Positive Selection in the Human Genome

The previous sections emphasize the practical challenges in (1) detecting positive selection in the genome, (2) confirming the adaptive loci, and (3) linking the genotype to the phenotype. Although research into human adaptation has many challenges there have been several striking success stories since the beginning of the genomic era one decade ago (Table 2.3). Studies of the impact of positive selection can be divided between candidate gene studies and genome-wide scans.

Table 2.3 Examples of positively selected genes supported by functional evidence. Caution: a unique article is cited while for many genes, several studies were required to conclude about the impact of positive selection and on the function of the putative selective allele

Gene	Selected function(s)	Adapted population	Approach	Reference
ABCC11	Ear wax secretion	Asian	Genome-wide scan	Xue et al. (2009)
CASP12	Sepsis resistance	Worldwide	Candidate gene	Xue et al. (2006)
CCR5	Bubonic plague or smallpox resistance	European	Candidate gene	Sabeti et al. (2005)
CD5	Pathogen recognition	East Asian	Candidate gene	Carnero-Montoro et al. (2012)
DARC	Malaria resistance	African	Candidate gene	Hamblin and Di Rienzo (2000)
EDAR	Hair/teeth/sweat gland development	Asian	Genome-wide scan	Sabeti et al. (2007)
EGLN1	Response to hypoxia	Tibetan and Sherpa	Genome-wide scan	Jeong et al. (2014) and Simonson et al. (2010)
EPAS1	Response to hypoxia	Tibetan and Sherpa	Genome-wide scan	Beall et al. (2010) and Jeong et al. (2014)
G6PD	Malaria resistance	African	Candidate gene	Tishkoff et al. (2001)
HBB	Malaria resistance	African	Candidate gene	Ayodo et al. (2007)
HERC2	Eye pigmentation	European	Candidate gene	Wilde et al. (2014)
LCT	Lactase persistence	European and African	Candidate gene	Bersaglieri et al. (2004) and Tishkoff et al. (2007)
SLC24A5	Skin pigmentation	European	Candidate gene	Lamason et al. (2005)
SLC45A2	Skin pigmentation	European	Genome-wide scan	Sabeti et al. (2007)
TLR5	Bacterial flagellin	African	Genome-wide scan	Grossman et al. (2013)
TNFSF5	Malaria resistance	African	Candidate gene	Sabeti et al. (2002a, b)
ZIP4	Zinc uptake	West Africa	Candidate gene	Engelken et al. (2014)

2.3.1 Candidate Gene Studies of Positive Selection

Candidate gene studies are driven by an a priori hypothesis about the implication of a gene in a putatively adaptive phenotype. Before the recent wealth of genomic data, this approach was most commonly used to detect positive selection. These studies show the impact of positive selection on specific genomic regions, identify candidate adaptive loci, and provide informative insights into the molecular basis of phenotypic adaptation across human populations. For example, several genes have been identified as targets of positive selection with supporting functional evidence for and a link to a phenotypic change conferring a fitness increase (Table 2.3): *G6PD*, *DARC*, *TNFS5*, and *HBB* which provide malaria resistance in Africa (Ayodo et al. 2007; Hamblin and Di Rienzo 2000; Sabeti et al. 2002b; Tishkoff et al. 2007); *LCT* which proffers lactose metabolism in populations with herder ancestors in Europe (Bersaglieri et al. 2004) and Africa (Tishkoff et al. 2007); *CASP12* which increases resistance to sepsis (Xue et al. 2006); and *CD5* which allows better pathogen recognition (Carnero-Montoro et al. 2012).

Although the aforementioned successes in detecting variants that have been selected, the candidate gene approach suffers from the three following main drawbacks:

1. An a priori hypothesis is required about which genes have been under positive selection, as well as knowledge of the relationship between genotype and phenotype. A candidate gene approach aims to pinpoint the functional variant, but the goal is rarely reached. Furthermore, when the function of the adaptive allele is established, it is difficult to determine how it confers a selective advantage to its carriers.
2. The adaptive variant can be located far from the region spanning the gene either within the coding or flanking region. In that case, if no previous knowledge on the gene regulatory regions exist, it would be impossible to detect the adaptive locus within a candidate gene framework.
3. In general, no sufficient biological knowledge on the molecular basis of adaptive phenotypes (or even diseases) across most of the genome can make a good a priori hypothesis of the underlying molecular bases of traits. Thus, a candidate gene approach is reduced to the study of annotated genes encoding relatively simple phenotypes.

For those reasons, and with the recent wealth of polymorphism data, an alternative approach has been developed: the genome-wide scan approach.

2.3.2 Genome-Wide Scans for Positive Selection

During the last decade, impressive technological progress in genotyping has been made, from high-throughput genotyping arrays to NGS, resulting in the bulk of genotype data needed to perform population genetics analyses. Now, large catalogs of genetic variability in worldwide human populations are publicly available allowing the study of the impact of natural selection on our genome. A large number of genome-wide scans of positive selection in different populations have been published recently (reviewed in Akey 2009; Fu and Akey 2013; Scheinfeldt and Tishkoff 2013). A top-down approach, with no a priori hypothesis on the adaptive phenotype, avoids the limitations of candidate-gene studies. The first genome-wide scan for positive selection in human populations was performed by Akey et al. (2002) and was rapidly followed by more than 20 others (Akey 2009). Since 2002, the number of individuals and markers available increases constantly and there has been theoretical development and implementation of several new methods for hard sweeps and alternative modes of positive selection. The boom of data and statistical methods to detect positive selection has revealed many more genomic regions that have putatively evolved in at least one population. In 2009, more than 5000 regions in the genome spanning a total of 400 Mb and encompassing more than 4000 protein-coding genes were reported in a review of 21 genome-wide scans published at that time (Akey 2009). Those 21 scans used methods designed to detect the molecular patterns left by a hard sweep. They also relied on the outlier approach and, therefore, established an a priori proportion of the genome under positive selection in the studied populations, likely leading to a high FPR. In his review (Akey 2009), Joshua Akey looked at the overlap of the genomic regions reported by 10 studies using the same data, but different statistics. Strikingly, only 14.1%, 5.3%, and 2.5% of the overall regions were reported in two, three, or four studies, respectively. Besides the FPR issue, it is clear that those genome-wide scans can also miss real events of selection as suggested by the fact that neither *G6PD* nor *DARC* has been reported by such studies.

Although the overlap among individual scans is low, more than 700 regions have been identified encompassing previous candidate adaptive loci and new well-supported ones (Table 2.3). Moreover, it appears that most signals of putative positive selection are not shared among populations from different geographic regions (for example, see Pickrell et al. 2009; Voight et al. 2006). This is expected when considering that the scans mostly relied on the hard sweep model, and therefore detected advantageous mutations that appear in the population just before being selected for. Indeed, geographically distant populations present different genetic backgrounds and have to adapt to very heterogeneous environmental conditions.

Genome-wide scans can map the signals of putative positive selection and will give great insights into how natural selection has shaped the human genome. They will also continue to aid in the discovery of functional elements. However, it remains challenging to extract the relevant information in the bulk of signals of positive selection from genome-wide scans in order to understand how the human population really evolved and what is at the molecular basis of phenotypic adaptation. Indeed, although the genome-wide approach circumvents some limitations of the candidate gene approach, it presents its own ones.

1. Large scale studies do not allow the extensive control for many layers of complexity. Indeed, in opposition to candidate gene approach, in a genome-wide scan it is extremely difficult to build an accurate model including both demographic and genomic processes that describe the evolution of a specific genomic region or to investigate in depth the molecular mechanisms affecting the genetic variability. Therefore, most scans rely on the outlier approach, and as already mentioned, only detect the most extreme cases of positive selection as well as suffering a likely high FPR (Teshima et al. 2006). As described before, one solution to reduce the FPR is to cross the results from different scans performed with different methods and/or on different populations.
2. Regions reported by genome-wide scans are usually large, spanning hundreds of kilobases and containing several contiguous genes and regulatory regions. Sometimes signals can be located in intergenic regions where no function has been reported yet. Therefore, it is often difficult to follow-up on the signals to identify whether the selected variant and the phenotype putatively increase the fitness.
3. For most genes, a certain amount of speculative discussion (story-telling) is necessary to determine which could be the adaptive phenotype.

For those reasons, most genome-wide scans focus on a reduced set of signals of putative selection based on biological information for a follow-up analysis. This practice is often referred as cherry picking. Hence, most of the signals already reported remain unexplained.

The recent scan performed by Grossman et al. (2013) developed new standards to overcome the aforementioned limitations and represents an important step toward the identification of putative adaptive variants as well as the underlying phenotypes increasing the fitness. This study made progress in several areas: (1) they used CMS which pinpoints more accurately the selected variant (Grossman et al. 2010); (2) they performed their analysis on the 1000 Genomes Project Pilot 1 re-sequencing data (The 1000 Genomes Project Consortium 2010); and (3) they analyzed the putative phenotypic implications of the selective variants by interrogating the ENCODE database and the GWAS catalog (Hindorff et al. 2009).

2.3.3 *Insights from Published Studies of Positive Selection in Humans*

All the studies aforementioned allowed the identification of putative adaptive loci, but also provide interesting insights in the nature of the genomic regions that have been preferentially targeted by positive selection in human populations. Allowing exploration of the phenotypic differences among populations and species that are induced from adaptation to new environments and which were the underlying biological functions at play.

2.3.3.1 Functional Categories for the Selected Protein-Coding Genes

A functional enrichment analysis is almost always performed after a genome-wide scan for positive selection. Such analysis tests whether the set of variants located within the regions of a signal for positive selection enrichment is in a biological process or functional pathway by contrasting whether more variants belong to a given functional class or pathway than expected by chance. To perform a functional enrichment analysis, these following databases are available:

1. *Gene Ontology* (GO; The Gene Ontology Consortium 2000) groups genes according to the features of the gene product. There are three main domains: (1) cellular component, i.e. the parts of the cell or its extracellular environment where the gene product is active; (2) molecular function, i.e. the elemental activities of the gene product at the molecular level (e.g. binding, catalysis, etc...); and (3) biological process, i.e. operations and sets of molecular events with a defined beginning and end that is pertinent to the functioning of integrated living units.
2. *PANTHER* (Protein Analysis Through Evolutionary Relationships; Mi et al. 2013) relies on annotation from GO among others and classifies proteins (and the encoding genes) as one of the following: (1) family, i.e. groups of evolutionarily related proteins and subfamilies (related proteins that also have the same function); (2) molecular function of the protein by itself or with directly interacting proteins at a biochemical level; (3) biological process, i.e. the function of the protein in the context of a larger network of proteins that interact to accomplish a process at the level of the cell or organism (e.g., mitosis); or (4) pathway which explicitly specifies the relationships between the interacting molecules.
3. *KEGG* (Kyoto Encyclopedia of Genes and Genomes; Kanehisa and Goto 2000) is a collection of manually curated databases integrating genomes, biological pathways, diseases, drugs, and chemical substances.
4. *Reactome Pathway Database* (Croft et al. 2011) contains curated functional pathway annotations that cover a diverse set of topics in molecular and cellular biology.

Genome-wide scans of positive selection using polymorphism data in human populations pointed to different categories enriched for genes that have evolved under a selective scenario: skin pigmentation, immunity, hair density, sweat gland, etc. (Kelley et al. 2006). Scans based on comparative genomics have revealed categories such as immunity and pathogen defense or sensory perception (Kosiol et al. 2008; Marques-Bonet et al. 2009).

However, functional enrichment analyses using such databases are biased toward protein-coding genes. In addition, an assumption of these databases is that all genes are independent and that all genes have the same level of importance within a pathway or a functional category. Although functional enrichment analysis has shed light on important functions and pathways that are preferentially targeted by positive selection, it does not provide a formal test for selection on a function. The current approach commonly used for large genome-wide analysis of positive selection is to detect signals at individual genes or regions. However, selected loci are just at the molecular basis of positive selection acting on the phenotypic level. Thus, single mutations rarely act in isolation to improve a function or to contribute to the acquisition of new ones. To overcome those limitations, Serra et al. created a new method called the Gene Set Selection Analysis (GSSA) to detect significant differences in scores of natural selection over functionally related genes (Serra et al. 2011). The method was applied genome-wide to coding regions of five mammals. But it still has never been used to interrogate non-coding elements or for polymorphism data.

2.3.3.2 Complex Adaptive Traits

The studies listed above describe the first attempts to move from individual genes to the biological modules they belong to. These studies start from individual genes or loci to then integrate the information on functional systems. The idea is that many loci will be involved in phenotypic adaptation, excluding Mendelian traits. This implies that polygenic adaptation is likely to be the main adaptive force acting on the human genome. First, Daub et al. used a gene-set enrichment test based on the F_{ST} statistic (*SUMSTAT*) to test for functional pathways or gene sets enriched in differentiated loci among populations (Daub et al. 2013). They found pathway enrichment in immune response confirming the general idea that response to pathogens has been a major selective force for human populations (for reviews, see Barreiro and Quintana-Murci 2010; Quintana-murci and Clark 2013). They also observed evidence of epistatic interactions between members of the same pathway. Specifically, a genome-wide scan detected several signals of selection for genes involved in the hypoxia-inducible factor 1 (HIF1) pathway which is involved in physiological response to hypoxic conditions (Simonson et al. 2010).

In order to examine polygenic adaptation and soft sweeps, several studies used methods better suited to study small shifts in allele frequency (Fumagalli et al. 2011; Hancock et al. 2010, 2011). When looking at covariation of diet, subsistence, or ecoregion, Hancock et al. found that pathways involved in starch and sucrose metabolism are enriched with signals of polygenic adaptation to a diet rich in roots and tubers, as well as an over-representation of signals associated to polar climate in genes involved in energy metabolism pathways (Hancock et al. 2010). Applying the same method with other environmental variables, they also described an enrichment of signals in gene sets related to UV radiation, infection, and immunity (Hancock et al. 2011). Conversely, Fumagalli et al. (2011), using a similar method, showed that local adaptation has been driven by the diversity of the local pathogenic environment while climate played a relatively minor role.

Berg and Coop, using the mean additive genetic value Q_X described several complex traits likely to have evolved through the action of polygenic adaptation (Berg and Coop 2014): height, pigmentation, and body mass index.

2.3.3.3 The Importance of Regulatory Elements

Although the method proposed by Berg and Coop (2014) is limited by relying on GWAS loci and the problem of portability among populations, it represents a major shift in the field. It is becoming clear that focusing only on protein-coding elements is not enough to understand adaptive evolution in humans. Although protein-coding sequences are very well annotated, they only represent around 1.2% of the human genome. Furthermore, the similarity between humans and chimpanzees in their protein-coding gene sequences cannot explain the observed phenotypic differences. In 1975, King and Wilson (King and Wilson 1975) suggested that differences in gene regulation may largely account for those phenotypic differences among species and populations. Since 1975, the relative contribution of variants located within protein-coding genes and regulatory regions has been debated. Evidence of the functionality of non protein-coding regions is the amount of conservation among species across the genome. For instance, 5% of the genome has been estimated to be largely conserved since the MRCA of mouse and human through the action of purifying selection. Hence, the conserved proportion of the genome is likely to be functional (Siepel et al. 2005). Since the proportion of conservation is higher than the proportion of protein-coding sequences in the genome, a large fraction of the elements with relevant biological function is non-coding.

Until recently, technical limitations have barred the exploration of the adaptive role of non-coding elements. Annotation outside gene regions has been lacking making it difficult to distinguish functional evidence of putative adaptation. This makes comparative genomic studies difficult as they rely on the comparison of the rate of substitution on functional versus non-functional elements and struggle to find an equivalent to the non-synonymous and synonymous changes in these badly annotated regions. In recent years, evidence indicates the role of regulatory elements in adaptive evolution. Using putatively neutral elements as a reference, Haygood et al. found that variants located in promoter regions had signatures of positive selection in the human and chimpanzee lineages (Haygood et al. 2007). Strikingly, they found an enrichment of signals of selection in nervous-system functions. Recent population genetics studies also indicate similar findings. First, Kadaravalli et al. using a genome-wide set of eQTLs and the statistic iHS they found that SNPs showing signals of selection are more likely than random to be associated with gene expression levels in *cis* (Kudaravalli et al. 2009). Second, with a similar study design but taking advantage of the recent wealth in eQTL databases and the ENCODE project, Hunter B. Fraser uses BayENV scores for polygenic adaptation to perform the first genome-scale study on the hypothesis that changes in gene expression have driven human adaptation (Fraser 2013). Third, Enard et al. observed a greater correlation in the observed signatures of positive selection (as inferred by iHS, XPEHH, and CLR) with the presence of regulatory sequences from ENCODE than with amino acid substitutions (Enard et al. 2014). Fourth, Arbiza et al. (2014) found a substantial amount of adaptive changes during human evolution affecting transcription binding sites.

All those studies suggest the functional importance of regulatory regions, their implication in adaptive evolution, and thus a substantial proportion of adaptive changes responsible for biological diversity are both inter and intra specific in regulatory regions. The aforementioned observation that hard sweeps were rare during human evolution (Hernandez et al. 2011) was based on a study design focusing on protein-coding regions. Thus, although the relative scarcity in hard sweeps pointed in this study is usually mentioned as a genomic trend and used against the hard sweep model, we think that generalizing those results to the whole genome is groundless.

2.4 Concluding Remarks

Identifying the molecular basis of phenotypic adaptation is a major challenge in evolutionary biology. The insights from population genetics are paramount to understanding human evolution through adaptive changes. However, most remain to be discovered. An exhaustive detection of selected variants will only be possible with tests for positive selection and in particular beyond the hard sweep model. We have discussed in this chapter that other scenarios further than positive selection must be considered. This is particularly true for the genes from the immune system which demonstrates that balancing selection has been impacting genome variability. Moreover, the examples of recently discovered regulatory adaptations and their importance in human adaptive evolution strongly suggest that only considering variants located within protein-coding regions is outdated. As genome annotation is getting more precise every day we are able to discover more targets of natural selection in non-coding regions. Moreover, while in this chapter we mostly focused on studies of point mutation (i.e., SNPs), other kinds of mutations segregate in the genome in large parts (i.e., structural variants) and have been overlooked by population geneticists.

The identification of variants encoding phenotypic selective changes relies on the downstream implementation of accurate models of neutral evolution that account for the complex human demography which have affected the genetic variability within and between populations. Those models must also integrate genomic mechanisms influencing the molecular patterns across the genome (e.g., mutation, recombination, and gene conversion).

Furthermore, the recent advances in many biological areas with the advent of the *omics* (e.g., transcriptomics, metabolomics, epigenomics, proteomics, and genomics) promise future groundbreaking discoveries. Although the rate at which data is currently generated may seem overwhelming, it allows many layers of complexity to come together which reduces the gap between the genotype and the phenotype. Therefore, population genetics must now work within a multidisciplinary framework in order to achieve its final goal of understanding the fitness consequences of selective variants.

Acknowledgements This work is a part of a research program developed in Jaume Bertranpetit's group with funding by grant PID2019-110933GB-I00/AEI/10.13039/501100011033 awarded by the Agencia Estatal de Investigación (AEI), Ministerio de Ciencia, Innovación y Universidades (MCIU, Spain) and with the support of Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya (GRC 2017 SGR 702). Part of the "Unidad de Excelencia María de Maeztu", funded by the AEI (CEX2018-000792-M). We are very grateful to Jessica Nye for her comprehensive proofreading of the manuscript. We want also to thank Elena Carnero-Montero, Brandon Invergo, and Ludovica Montanucci for sharing valuable material.

References

Akey JM (2009) Constructing genomic maps of positive selection in humans: where do we go from here? Genome Res 19(5):711–722

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002) Interrogating a high-density SNP map for signatures of natural selection. Genome Res 12:1805–1814

Arbiza L, Gronau I, Aksoy BA, Hubisz MJ, Gulko B, Keinan A, Siepel A (2014) Genome-wide inference of natural selection on human transcription factor binding sites. Nat Genet 45 (7):723–729

Ayodo G, Price AL, Keinan A, Ajwang A, Otieno MF, Orago ASS, Reich D (2007) Combining evidence of natural selection with association analysis increases power to detect malaria-resistance variants. Am J Hum Genet 81(2):234–242

Barreiro LB, Quintana-Murci L (2010) From evolutionary genetics to human immunology: how selection shapes host defence genes. Nat Rev Genet 11(1):17–30

Barrett RDH, Hoekstra HE (2011) Molecular spandrels: tests of adaptation at the genetic level. Nat Rev Genet 12(11):767–780

Barton N (1998) The geometry of adaptation. Nature 395(6704):751–752

Beall CM, Cavalleri GL, Deng L, Elston RC, Gao Y, Knight J, Li C (2010) Natural selection on EPAS1 (HIF2 α) associated with low hemoglobin concentration in Tibetan highlanders. Proc Natl Acad Sci U S A 107(25):11459–11464

Berg JJ, Coop G (2014) A population genetic signal of polygenic adaptation. PLoS Genet 10(8): e1004412. https://doi.org/10.1371/journal.pgen.1004412

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Hirschhorn JN (2004) Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet 74:1111–1120

Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W (1995) The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics 140(2):783–796

Cann H, de Toma C, Cazes L, Legrand M, Morel V, Piouffre L, Cavalli-Sforza L (2002) A human genome diversity cell line panel. Science 12(296(5566)):261–262

Carnero-Montoro E, Bonet L, Engelken J, Bielig T, Martínez-Florensa M, Lozano F, Bosch E (2012) Evolutionary and functional evidence for positive selection at the human CD5 immune receptor gene. Mol Biol Evol 29(2):811–823

Casals F, Bertranpetit J (2012) Human genetic variation, shared and private. Science (New York, NY) 337(6090):39–40

Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. Genetics 134(4):1289–1303

Chen H, Patterson N, Reich D (2010) Population differentiation as a test for selective sweeps. Genome Res 20(3):393–402

Colosimo PF, Hosemann KE, Balabhadra S, Villarreal G, Dickson M, Grimwood J, Kingsley DM (2005) Widespread parallel evolution in sticklebacks by repeated fixation of Ectodysplasin alleles. Science (New York, NY) 307(5717):1928–1933. https://doi.org/10.1126/science.1107239

Conrad DF, Jakobsson M, Coop G, Wen X, Wall JD, Rosenberg N a, Pritchard JK (2006) A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nat Genet 38(11):1251–1260

Coop G, Witonsky D, Di Rienzo A, Pritchard JK (2010) Using environmental correlations to identify loci underlying local adaptation. Genetics 1423(August):1411–1423

Croft D, Kelly GO, Wu G, Haw R, Gillespie M, Matthews L, Stein L (2011) Reactome : a database of reactions, pathways and biological processes. Nucleic Acids Res 39(Database issue):691–697

Darwin CR (1859) On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. John Murray, London

Darwin CR, Wallace AR (1858) On the tendency of species to form varieties; and on the perpetuation of varieties and species by natural means of selection. J Proc Linnean Soc London 3(9):46–50

Daub JT, Hofer T, Cutivet E, Dupanloup I, Quintana-murci L, Robinson-rechavi M, Excoffier L (2013) Evidence for polygenic adaptation to pathogens in the human genome article fast track. Mol Biol Evol 30(7):1544–1558. https://doi.org/10.1093/molbev/mst080

Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Lochovsky L (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489(7414):57–74

Enard D, Messer PW, Petrov D a (2014) Genome-wide signals of positive selection in human evolution. Genome Res 24(6):885–895

Engelken J, Carnero-Montoro E, Pybus M, Andrews GK, Lalueza-Fox C, Comas D, Bosch E (2014) Extreme population differences in the human zinc transporter ZIP4 (SLC39A4) are explained by positive selection in sub-Saharan Africa. PLoS Genet 10(2):e1004128

Ewing G, Hermisson J (2010) MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics 26(16):2064–2065

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M (2013) Robust demographic inference from genomic and SNP data. PLoS Genet 9(10):e1003905

Fagny M, Patin E, Enard D, Barreiro LB, Quintana-Murci L, Laval G (2014) Exploring the occurrence of classic selective sweeps in humans using whole-genome sequencing data sets. Mol Biol Evol 31(7):1850–1868

Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. Genetics 155(3):1405–1413

Ferrer-Admetlla A, Liang M, Korneliussen T, Nielsen R (2014) On detecting incomplete soft or hard selective sweeps using haplotype structure. Mol Biol Evol 31(5):1275–1291

Fraser HB (2013) Gene expression drives local adaptation in humans. Genome Res 23 (7):1089–1096

Fraser HB, Babak T, Tsang J, Zhou Y, Zhang B, Mehrabian M, Schadt EE (2011) Systematic detection of polygenic cis-regulatory evolution. PLoS Genet 7(3)

Fu YX (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics 147(2):915–925

Fu W, Akey JM (2013) Selection and adaptation in the human genome. Annu Rev Genomics Hum Genet 14:467–489

Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. Genetics 133(3):693–709

Fu W, O'Connor TD, Akey JM (2013) Genetic architecture of quantitative traits and complex diseases. Curr Opin Genet Dev 23(6):678–683

Fumagalli M, Sironi M, Pozzoli U, Ferrer-Admetlla A, Ferrer-Admettla A, Pattini L, Nielsen R (2011) Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. PLoS Genet 7(11):e1002355

Garud NR, Messer PW, Buzbas EO, Petrov DA (2014) Recent selective sweeps in Drosophila were abundant and primarily soft. *arXiv*

Gillespie JH (2000) Genetic drift in an infinite population : the Pseudohitchhiking model. Genetics 155:909–919

González-Neira A, Ke X, Lao O, Calafell F, Navarro A, Comas D, Cann H, Bumpstead S, Ghori J, Hunt S, Deloukas P, Dunham I, Cardon LR, Bertranpetit J (2006) The portability of tagSNPs across populations: a worldwide survey. Genome Res 16(3):323–330

Granka JM, Henn BM, Gignoux CR, Kidd JM, Bustamante CD, Feldman MW (2012) Limited evidence for classic selective sweeps in African populations. Genetics 192(3):1049–1064

Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Bustamante CD (2011) Demographic history and rare allele sharing among human populations. Proc Natl Acad Sci U S A 108(29):11983–11988

Grossman SR, Shylakhter I, Karlsson EK, Byrne EH, Morales S, Frieden G, Sabeti PC (2010) A composite of multiple signals distinguishes causal variants in regions of positive selection. Science 327(5967):883–886

Grossman SR, Andersen KG, Shlyakhter I, Tabrizi S, Winnicki S, Yen A, Sabeti PC (2013) Identifying recent adaptations in large-scale genomic data. Cell 152(4):703–713

Günther T, Coop G (2013) Robust identification of local adaptation from allele frequencies. Genetics 195(1):205–220

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet 5(10):e1000695

Hallatschek O, Nelson DR (2008) Gene surfing in expanding populations. Theor Popul Biol 73 (1):158–170

Hamblin MT, Di Rienzo A (2000) Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. Am J Hum Genet 66(5):1669–1679

Hancock AM, Alkorta-Aranburu G, Witonsky DB, Di Rienzo A (2010) Adaptations to new environments in humans: the role of subtle allele frequency shifts. Philos Trans R Soc Lond Ser B Biol Sci 365(1552):2459–2468

Hancock AM, Witonsky DB, Alkorta-aranburu G, Beall CM, Sukernik R, Utermann G, Di Rienzo A (2011) Adaptations to climate-mediated selective pressures in humans. PLoS Genet 7(4): e1001375

Haygood R, Fedrigo O, Hanson B, Yokoyama K-D, Wray GA (2007) Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. Nat Genet 39(9):1140–1144

Hermisson J, Pennings PS (2005) Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics 169(4):2335–2352

Hernandez RD (2008) A flexible forward simulator for populations subject to selection and demography. Bioinformatics 24(23):2786–2787

Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, Przeworski M (2011) Classic selective sweeps were rare in recent human evolution. Science 331:920–924

Hindorff LA, Sethupathy P, Junkins H a, Ramos EM, Mehta JP, Collins FS, Manolio TA (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106(23):9362–9367

Hofer T, Ray N, Wegmann D, Excoffier L (2009) Large allele frequency differences between human continental groups are more likely to have occurred by drift during range expansions than by selection. Ann Hum Genet 73(1):95–108

Holsinger KE, Weir BS (2009) Genetics in geographically structured populations: defining, estimating and interpreting FST. Nat Rev Genet 10(9):639–650

Hudson RR (1991) Oxford surveys in evolutionary biology. Oxford University Press, Oxford, pp 1–44

Hudson RR (2002) Bioinformatics applications note. Bioinformatics 18(2):337–338

Innan H, Kim Y (2004) Pattern of polymorphism after strong artificial selection in a domestication event. Proc Natl Acad Sci U S A 101(29):10667–10672

Jeong S, Rebeiz M, Andolfatto P, Werner T, True J, Carroll SB (2008) The evolution of gene regulation underlies a morphological difference between two drosophila sister species. Cell 132 (5):783–793

Jeong C, Alkorta-Aranburu G, Basnyat B, Neupane M, Witonsky DB, Pritchard JK, Di Rienzo A (2014) Admixture facilitates genetic adaptations to high altitude in Tibet. Nat Commun 5:3281

Kanehisa M, Goto S (2000) KEGG : Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28(1):27–30

Kaplan NL, Hudsont RR, Langle CH (1989) The "hitchhiking effect" revisited. Genetics 899:887–899

Keinan A, Clark AG (2012) Recent explosive human population growth has resulted in an excess of rare genetic variants. Science (New York, NY) 336(6082):740–743

Kelley JL, Madeoy J, Calhoun JC, Swanson W, Akey JM (2006) Genomic signatures of positive selection in humans and the limits of outlier approaches. Genome Res 16(8):980–989

Kelly JK (1997) A test of neutrality based on Interlocu associations. Genetics 1206:1197–1206

Kim Y (2006) Allele frequency distribution under recurrent selective sweeps. Genetics 1978:1967–1978

King M, Wilson AC (1975) Humans and Chimpanze es. Science 188(4184):107–116

Kingman JFC (1982) The coalescent. Stoch Process Appl 13:235–248

Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A (2008) Patterns of positive selection in six Mammalian genomes. PLoS Genet 4(8):e1000144

Kudaravalli S, Veyrieras J-B, Stranger BE, Dermitzakis ET, Pritchard JK (2009) Gene expression levels are a target of recent natural selection in the human genome. Mol Biol Evol 26 (3):649–658

Lamason RL, Mohideen M-APK, Mest JR, Wong AC, Norton HL, Aros MC, Cheng KC (2005) SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. Science (New York, NY) 310(5755):1782–1786

Leinonen T, McCairns RJS, O'Hara RB, Merilä J (2013) Q(ST)-F(ST) comparisons: evolutionary and ecological insights from genomic heterogeneity. Nat Rev Genet 14(3):179–190

Lin K, Li H, Schlötterer C, Futschik A (2011) Distinguishing positive selection from neutral evolution: boosting the performance of summary statistics. Genetics 187(1):229–244

Luisi P, Alvarez-Ponce D, Pybus M, Fares MA, Bertranpetit J, Laayouni H (2015) Recent positive selection has acted on genes encoding proteins with more interactions within the whole human interactome. Genome Biol Evol 7(4):1141–1154. https://doi.org/10.1093/gbe/evv055

Mailund T, Schierup MH, Pedersen CNS, Mechlenborg PJM, Madsen JN, Schauser L (2005) CoaSim: a flexible environment for simulating genetic data under coalescent models. BMC Bioinformatics 6:252

Marjoram P, Wall JD (2006) Fast "coalescent" simulation. BMC Genet 7:16

Marques-Bonet T, Ryder O a, Eichler EE (2009) Sequencing primate genomes: what have we learned? Annu Rev Genomics Hum Genet 10:355–386

Maynard-Smith J, Haigh J (1974) The hitch-hiking effect of a favourable gene. Genet Res 23 (1):23–35

McEvedy C (1988) The bubonic plague. Sci Am 258(2):117–123

Mendizabal I, Marigorta UM, Lao O, Comas D (2012) Adaptive evolution of loci covarying with the human African pygmy phenotype. Hum Genet 131(8):1305–1317

Mi H, Muruganujan A, Thomas PD (2013) PANTHER in 2013 : modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Res 41 (Database issue):377–386

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York, NY

Nei M, Li WH (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci U S A 76(10):5269–5273

Nielsen R, Hubisz MJ, Clark AG (2004) Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. Genetics 168:2373–2382

Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C (2005) Genomic scans for selective sweeps using SNP data. Genome Res 15:1566–1575

Orr HA (1998) Testing natural selection vs. genetic drift in phenotypic evolution using quantitative trait locus data. Genetics 149:2099–2104

Orr HA, Betancourt AJ (2001) Haldane ' s sieve and adaptation from the standing genetic variation. Genetics 157:875–884

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Reich D (2012) Ancient admixture in human history. Genetics 192(3):1065–1093

Pennings PS, Hermisson J (2006) Soft sweeps II--molecular population genetics of adaptation from recurrent mutation or migration. Mol Biol Evol 23(5):1076–1084

Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, Pritchard JK (2009) Signals of recent positive selection in a worldwide sample of human populations. Genome Res 19 (5):826–837

Pickrell JK, Patterson N, Barbieri C, Berthold F, Gerlach L, Güldemann T, Pakendorf B (2012) The genetic prehistory of southern Africa. Nat Commun 3:1143

Pritchard JK, Pickrell JK, Coop G (2010) The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr Biol 20(4):R208–R215

Przeworski M (2002) The signature of positive selection at randomly chosen loci. Genetics 1189:1179–1189

Przeworski M, Coop G, Wall JD (2005) The signature of positive selection on standing genetic variation. Evol Int J Organic Evol 59(11):2312–2323

Pybus M, Dall'Olio GM, Luisi P, Uzkudun M, Carreño-Torres A, Pavlidis P, Engelken J (2014) 1000 genomes selection browser 1.0: a genome browser dedicated to signatures of natural selection in modern humans. Nucleic Acids Res 42(Database issue):D903–D909

Quintana-murci L, Clark AG (2013) Population genetic tools for dissecting innate immunity in humans. Nat Rev Immunol 13(4):280–293

Ralph PL, Coop G (2010) Parallel adaptation: one or many waves of advance of an advantageous allele? Genetics 668(October):647–668

Ramos-Onsins SE, Rozas J (2002) Statistical properties of new neutrality tests against population growth. Mol Biol Evol 19(12):2092–2100

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lander ES (2001) Linkage disequilibrium in the human genome. Nature 411(6834):199–204

Rozas J, Gullaud M, Blandin G, Aguadé M (2001) DNA variation at the rp49 gene region of Drosophila simulans: evolutionary inferences from an unusual haplotype structure. Genetics 158(3):1147–1155

Sabeti P, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, Lander ES (2002a) Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832–837

Sabeti P, Usen S, Farhadian S, Jallow M, Doherty T, Newport M, Pinder M, Ward R, Kwiatkowski D (2002b) CD40L association with protection from severe malaria. Genes & Immun 3(5):286–291

Sabeti PC, Walsh E, Schaffner SF, Varilly P, Fry B, Hutcheson HB, Lander ES (2005) The case for selection at CCR5-Delta32. PLoS Biol 3(11):e378

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Stewart J (2007) Genome-wide detection and characterization of positive selection in human populations. Nature 449(7164):913–918

Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D (2005) Calibrating a coalescent simulation of human genome sequence variation. Genome Res 15(11):1576–1183

Schapire RE (1990) The strength of weak learnability. Mach Learn 5:197–227

Scheinfeldt LB, Tishkoff SA (2013) Recent human adaptation: genomic approaches, interpretation and insights. Nat Rev Genetics 14(10):692–702

Scheinfeldt LB, Biswas S, Madeoy J, Connelly CF, Schadt EE, Akey JM (2009) Population genomic analysis of ALMS1 in humans reveals a surprisingly complex evolutionary history. Mol Biol Evol 26(6):1357–1367

Serra F, Arbiza L, Dopazo J, Dopazo H (2011) Natural selection on functional modules, a genome-wide analysis. PLoS Comput Biol 7(3):e10001093

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Haussler D (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15(8):1034–1050

Simonson TS, Yang Y, Huff CD, Yun H, Qin G, Witherspoon DJ, Ge R (2010) Genetic evidence for high-altitude adaptation in Tibet. Science 329(5987):72–75

Spencer CCA, Coop G (2004) SelSim: a program to simulate population genetic data with natural selection and recombination. Bioinformatics 20(18):3673–3675

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123(3):585–595

Teshima KM, Coop G, Przeworski M (2006) How reliable are empirical genomic scans for selective sweeps? Genome Res 16(6):702–712

The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467(7319):1061–1073

The 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491(7422):56–65

The Gene Ontology Consortium (2000) Gene ontology : tool for the. Nat Genet 25:25–29

Thornton KR, Jensen JD (2007) Controlling the false-positive rate in multilocus genome scans for selection. Genetics 175(2):737–750

Tishkoff SA, Varkonyi R, Cahinhinan N, Abbes S, Argyropoulos G, Destro-Bisol G, Clark AG (2001) Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. Science 293(5529):455–462

Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Deloukas P (2007) Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet 39 (1):31–40

Turchin M, Chiang CWK, Palmer CD, Sankararaman SRD, Hirschhorn JN, GIANT consortium (2012) Evidence of widespread selection on standing variation in Europe at height-associated SNPs. Nat Genet 44(9):1015–1019

Uricchio LH, Hernandez RD (2014) Robust forward simulations of recurrent hitchhiking. Genetics 197:221–236, 1–33

Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS Biol 4(3):e72

Wakeley J (2008) Coalescent theory: an introduction. Roberts & Company Publishers, Greenwood Village

Wall JD (1999) Recombination and the power of statistical tests of neutrality. Genet Res 74:65–79

Wall JD (2000) A comparison of estimators of the population recombination rate. Mol Biol Evol 17 (1):156–163

Watterson GA (1978) The homozygosity test of neutrality. Genetics 88(2):405–417

Weir BS, Cockerham C (1984) Estimating F-statistics for the analysis of population structure. Evolution 38:1358–1370

Wilde S, Timpson A, Kirsanow K, Kaiser E, Kayser M, Unterländer M, Burger J (2014) Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y. Proc Natl Acad Sci U S A 111(13):4832–4837

Wollstein A, Lao O, Becker C, Brauer S, Trent RJ, Nürnberg P, Kayser M (2010) Demographic history of Oceania inferred from genome-wide data. Curr Biol 20(22):1983–1992

Xue Y, Daly A, Yngvadottir B, Liu M, Coop G, Kim Y, Tyler-smith C (2006) Spread of an inactive form of Caspase-12 in humans is due to recent positive selection. Am J Hum Genet 78:659–670

Xue Y, Zhang X, Huang N, Daly A, Gillson CJ, Macarthur DG, Tyler-Smith C (2009) Population differentiation as an indicator of recent positive selection in humans: an empirical evaluation. Genetics 183(3):1065–1077

Zeng K, Fu Y, Shi S, Wu C (2006) Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics 174:1431–1439

Zeng K, Shi S, Wu C-I (2007) Compound tests for the detection of hitchhiking under positive selection. Mol Biol Evol 24(8):1898–1908

Chapter 3
Population Genomics of High-Altitude Adaptation

Lian Deng and Shuhua Xu

Abstract The natural hypoxic experiments performed on native human populations residing the highlands provide an excellent opportunity to learn how environmental challenges reform human genetic architecture. In this chapter, we give a broad overview of current evidence for physiological and genetic adaptations based on three renowned highland groups from the Qinghai-Tibet Plateau, the Andes Altiplano, and the Ethiopian Plateau. We summarize several well-recognized adaptive signals strongly suggested by early studies and highlight recent findings accumulating rapidly and broadly with whole-genome sequencing and multi-omics approaches. These studies offer a glimpse into the complex driving forces and mechanisms of adaptive evolution and imply the genetic predisposition of relevant diseases and possible therapeutic strategies.

Keywords High altitude · Tibetan · Andean · Ethiopian · Hypoxia · Natural selection · Adaptive introgression · Convergent adaptation

L. Deng
Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China
e-mail: denglian@picb.ac.cn

S. Xu (✉)
Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China

School of Life Science and Technology, ShanghaiTech University, Shanghai, China

Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming, China

Henan Institute of Medical and Pharmaceutical Sciences, Zhengzhou University, Zhengzhou, China

Collaborative Innovation Center of Genetics and Development, School of Life Sciences, Fudan University, Shanghai, China
e-mail: xushua@picb.ac.cn

N. Saitou (ed.), *Evolution of the Human Genome II*, Evolutionary Studies,
https://doi.org/10.1007/978-4-431-56904-6_3

3.1 Background

Modern human populations have been undergoing rapid adaptive evolution in response to the changing environmental conditions during their spread out of Africa and settlement in new challenging habitats in the last 100,000 years (Hawks et al. 2007; Fan et al. 2016). Recent development in genomic technologies and methodologies and expanded population samples have enabled us to understand the genetic determinants of phenotypic variations in response to diverse environmental pressures. Well-known examples of human genetic adaptations include innate immune response, height, skin pigmentation, lactose tolerance, fatty acid metabolic efficiency, and hemoglobin (Hb) concentration, all of which are highlighted in detail in Fan et al. (2016)

Populations that have adapted to environments that are extremely challenging to survival provide well-characterized cases for local adaptation. One such harsh environment is the high-altitude hypoxia. The barometric pressure falls with ascent of hilly altitude, and the oxygen concentration at elevation >5000 meters (m) is only <60% of that at the sea level. Human adaptation to high altitudes has a long history, and is considered as a fascinating micro-evolutionary mechanism to combat environmental stress. In this chapter, we review the genetic evidence of human adaptation to high-altitude environments, with a focus on three major indigenous highland populations in the Tibetan, Andean, and Ethiopian plateaus. This will be better understood in the context of population history at high altitudes and physiological characteristics in the hypoxic environment. Earlier findings largely focus on a limited number of genes and traits, while rapid growth of multi-omics data in recent years highlights the complexity of high-altitude adaptation.

3.2 High-Altitude Human Populations

According to a recent survey of geospatial data, the high-altitude areas at elevation over 3000 m in the world are populated by over 370 million people in 80 countries (Table 3.1) (Center for International Earth Science Information Network, C. C. U 2012). Only <1% of these people are permanently living and spending healthy lives at above 5000 m, and they are grouped in Asia and Americas. Most of the highlanders depend on agriculture and pastoralism for subsistence in the rural locations. High-altitude areas are thought to be natural laboratories to test how human populations have adapted to extreme environmental challenges, and most of our current knowledge is based on three renowned populations inhabiting the Qinghai-Tibet Plateau, the Andes Altiplano, and the Ethiopian Plateau.

Table 3.1 Global population estimates at elevation above 3000 m

Country	2010 population estimate	Country	2010 population estimate
Africa	**108,620,694**		
Burundi	5,041,897	Angola	3,890,468
Comoros	17,597	Cameroon	934,290
Eritrea	2,401,614	Congo	5,945,225
Ethiopia	62,923,716	Algeria	123,043
Kenya	21,031,864	Morocco	1,949,564
Madagascar	1,889,171	Lesotho	1,975,589
Mozambique	29,090	Namibia	308,709
Malawi	158,857		
Americas	**101,895,587**		
Dominica	216,449	*Argentina	596,223
Haiti	124,944	*Bolivia	6,514,125
Costa Rica	900,704	Brazil	189,350
Guatemala	5,620,071	*Chile	2,895,044
Honduras	271,362	*Colombia	20,254,055
*Mexico	58,634,788	*Ecuador	5,614,842
Canada	63,630		
Asia	**161,830,272**		
*Kazakhstan	494,851	*India	20,970,922
*Kyrgyzstan	2,085,851	*Iran	26,250,181
*China	88,197,931	Sri Lanka	324,160
Japan	551,606	Armenia	1,091,531
Mongolia	890,292	Azerbaijan	583,336
Indonesia	2,096,971	Georgia	424,478
Laos	47,004	Iraq	275,155
*Myanmar	581,541	*Bhutan	425,462
Malaysia	69,735	Jordan	10,737
*Afghanistan	16,111,231	Lebanon	347,297
Europe	**2,388,728**		
Bulgaria	115,334	Italy	630,297
Albania	113,229	Macedonia	54,619
Andorra	62,192	Austria	488,643
Bosnia-Herzegovina	72,726	Switzerland	303,941
Spain	251,273	Germany	32,073
Greece	41,834	France	222,567
Total	**374,735,281**		

Countries with high-altitude population estimate over 10,000 are listed. A country is labeled with an asterisk "*" if some of its population live at elevation above 5000 m. Data are acquired from SEDAC—Socioeconomic Data and Applications Center (available at: https://sedac.ciesin.columbia.edu/data/set/nagdc-population-landscape-climate-estimates-v3/data-download)

3.2.1 Tibet

Residing at "the Third Pole of the Earth" with an average elevation over 4000 m, Tibetan highlanders are the most extensively studied high-altitude population. Long-term human occupation of the Qinghai-Tibet Plateau has been proved by massive evidence. Most of the recent archeological and genetic studies documented a human presence at the Tibetan Plateau as early as 30,000–40,000 years ago (Lu et al. 2016; Zhang et al. 2018; Qi et al. 2013; Aldenderfer 2011); while evidence for sustained agricultural and artisanal activity suggested a permanent occupation of the Tibetan Plateau during Neolithic era (Chen et al. 2015; Li et al. 2019). However, a chronological analysis of the Chusang site at an elevation of ~4270 m confirmed the permanent preagricultural occupation of the central plateau (Meyer et al. 2017). Recent discovery of a Denisovan mandible from Baishiya Karst Cave (3280 m above sea level) indicated a late Middle Pleistocene occupation on Tibetan Plateau around 100,000–60,000 years ago (Chen et al. 2019; Zhang et al. 2020). This supports the hypothesis that archaic hominins are involved in the prehistory evolution of Tibetan highlanders and have contributed to their high-altitude adaptation (Lu et al. 2016; Huerta-Sanchez et al. 2014). Estimates of divergence time between Tibetans and Han Chinese vary largely across genetic studies, which is 2750 years before present (BP) based on exome sequencing data (Yi et al. 2010), 6650–6796 years BP using genome-wide single nucleotide polymorphisms (SNPs) (Qi et al. 2013), and as early as 15,000–9000 years BP according to a whole-genome sequencing study (Lu et al. 2016). Extensive gene flow between the highlanders and people from outside area may lead to biased estimation of the population genetic divergence (Qi et al. 2013; Zhang et al. 2017a; Qin et al. 2010).

3.2.2 Andes

The Andean Altiplano is the largest and highest plateau in the world outside of Tibet. It has an average elevation of 3750 m above sea level and even reaches 6000 m in some places. Radiocarbon dating of human skeletons and cultural remains suggested that the Altiplano has been occupied by human for 12,000 years (Lynch and Kennedy 1970; Rothhammer and Santoro 2001; Rothhammer and Silva 1989; Rademaker et al. 2014). The Cuncaicha rock shelter at 4480 m is the oldest known site in high Andes, and it is dated to ~12,400 years BP (Rademaker et al. 2014). The peopling of the Andes in late Pleistocene is also evidenced by genetic estimates on ancient and modern human DNA (Fuselli et al. 2003; Moreno-Mayar et al. 2018; Posth et al. 2018). Similar dates of different archeological sites along the coast suggested rapid human dispersal in Andes despite the harsh environment (Rademaker et al. 2014; Dillehay and Collins 1988; Sandweiss et al. 1998). Moreover, long-standing continuity between ancient and modern Andeans and great homogeneity of geographical modern Andeans populations were detected by genetic investigations (Fuselli et al. 2003; Posth et al. 2018; Fehren-Schmitz et al. 2011; Tarazona-Santos et al. 2001; Lindo et al. 2018; Llamas et al. 2016). According to a recent study of ancient human, the highland population further split into northern and southern populations around 5800 BP (Nakatsuka et al. 2020). Uros are thought to be the first settlers of the Andean Altiplano, yet their origin has been subjected to considerable academic debate (Sandoval et al. 2013). The arrival of Aymara pushed the Uro tribes to local habitats. Aymara and Quechua are two major linguistically defined ethnic groups currently inhabiting the Andes highland, and they are thought to be precursors and direct descendants of the Inca Empire (Twelfth to fifteenth century), respectively. European admixture in the seventeenth to twentieth centuries contributes to the linguistic distinction, physiological differences, and genomic diversity of present-day Andeans populations (Brutsaert et al. 2003, 2005; Eichstaedt et al. 2014; Julian et al. 2009; Rupert and Hochachka 2001).

3.2.3 Ethiopia

The Ethiopian Plateau is the most extensive high-altitude feature in Africa, with an average elevation of 3000 m above sea level, which is lower than that of the Qinghai-Tibet Plateau and the Andes Altiplano. Two well-known Ethiopian highlander groups living at altitude >2500 m are Amhara and Oromo, constituting 73% of the total Ethiopian population and have been targeted in most of the recent genetic studies on high-altitude adaptation (Alkorta-Aranburu et al. 2012; Huerta-Sanchez et al. 2013; Scheinfeldt et al. 2012; Udpa et al. 2014). The Amhara (Semitic-speaking) are Arabic in origin, while Oromo (Cushitic-speaking) are known as Ethiopian Somalians (Cavalli-Sforza et al. 1995). Archeological studies showed that Amhara have been settled at >2500 m elevation since 5000 years ago (Lewis 1966; Pleurdeau 2005), while their inhabitation at 2300–2400 m could be dated to as early as 70,000 years BP (Aldenderfer 2003). The Oromo, however, have adapted living at altitude >2500 m recently, i.e. 500 years ago (Hassen 1990). Excessive gene flow has been identified between lowland and highland communities in this area (Huerta-Sanchez et al. 2013; Pagani et al. 2012a).

3.3 Physiological Clues of High-Altitude Adaptation

The long-term residence of the indigenous people under extreme environmental stresses at high altitudes indicates existence of genetic adaptation. Although genetic adaptation is restricted to physiological changes affecting reproductive success by evolutionary biologists or geneticists, it is indeed a composite outcome of systemic responses, many of which are not necessarily direct determinants of reproduction. A comprehensive characterization of the physiological changes may facilitate our understanding of genetic mechanisms underlying high-altitude adaptation.

Hypobaric hypoxia is considered to be the most severe challenge to human survival at high altitudes, although low temperature, high ultraviolet (UV) radiation, and pathogens are also potential deleterious factors. The physiological changes under hypoxia have been extensively accessed via two sources of information.

One source of information is the adverse effects of acute and chronic exposure to hypoxia on human normal body functioning. Over the past decades there have been substantial studies revealing the hypoxia-induced disorders in respiratory system, cardiovascular system, nervous system, immune system, blood, reproduction, skeletal muscle tissue, endocrine function, nutrition, and metabolism (Schumacker et al. 2014). Based on the effects of altitude and acclimation on individual performance and well-being, Bartsch and Saltin (2008) proposed a classification of the altitude levels. At lower altitude (<2000 m), the decrease of partial pressure of oxygen has no effect on well-being but relevant impairment of aerobic performance. This can be overcome completely by acclimatization, such as higher ventilation and heart rate to enlarge the oxygen intake, as well as reduced plasma volume in the first 24–48 h and an enhanced erythropoiesis and larger Hb mass during a prolonged exposure in hypoxia to allow for larger oxygen-carrying capacity of the blood. When ascending to moderate altitude (2000–3000 m), acclimatization is not adequate and 10–30% of subjects may have acute mountain sickness (AMS) featured as headache, vomiting, tiredness, and trouble sleeping. At higher altitudes (>3000 m), serious consequences of high-altitude cerebral edema (HACE, at >3000 m) and high-altitude pulmonary edema (HAPE, at >4000 m) would occur. In addition to the respiratory and cardiac outcomes, sustained low barometric oxygen pressure adversely affects body weight, muscle structure, exercise capacity, energy metabolism, mental functioning, and olfactory sensation in low-altitude accustomed people when they are certainly exposed to hypoxia (Leon-Velarde et al. 2005; Naeije 2010; Sargent et al. 2013; Altundag et al. 2014; Ferezou et al. 1993; Ou and Leiter 2004). It has the heaviest impairments during perinatal life, increases the risk of intrauterine growth restriction and pre-eclampsia, and influences neonatal survival (Crocker et al. 2020; Niermeyer et al. 2009; Moore et al. 2004; Julian and Moore 2019).

The other source of information involves distinct physiological patterns of long-term evolutionary adaptation and short-term biological acclimatization, as well as those of sea level. This is the most convincing evidence of genetic adaptation to high altitudes. When native lowlanders ascend to the highlands, there would be immediate increases of resting ventilation and basic metabolic rate, as well as reduction in plasma volume to maintain the arterial oxygen content (Martin and Windsor 2008; Virues-Ortega et al. 2006; West 1982). The initial cardiovascular response to altitude is characterized by an increase in heart rate and blood pressure, but no change of stroke volume (Naeije 2010). The native highlanders generally resemble acclimatized newcomers with higher levels of resting ventilation, Hb concentration, and cardiac output than the sea-level ranges. However, different continental highland populations have acquired distinct physiological traits to offset hypoxia and hence cluing towards independent arising of these adaptive phenotypes among geographically distinct populations (Scheinfeldt and Tishkoff 2010). Compared to sea level, Andeans show increased Hb concentration, 4–5% lower oxygen saturation (SaO_2), and 16% higher arterial oxygen content; Tibetans also have slightly increased Hb concentration and >5% lower SaO_2, but 10% lower arterial oxygen content (Beall 2006). Hb concentration in Tibetans is lower than that in Andeans at the same altitude, but it reaches a turning point at 4500 m with a sharp increase in Tibetans (Beall et al. 1998; Beall and Goldstein 1987; Beall and Reichsman 1984; Zhang et al. 2017b). Previous studies identified no distinct adaptive pattern for the Ethiopian highlanders compared with sea level (Beall et al. 2002), except one reported increased Hb concentration and decreased SaO_2 (Alkorta-Aranburu et al. 2012). Nevertheless, the increase of Hb concentration in the indigenous highlanders is to a much lower extent than that exhibited by lowlanders ascending to the highlands (Beall 2006; Beall et al. 2002; Wu et al. 2005). Both Hb concentration and SaO_2 are heritable, according to twins and family data analyses. Heritability (h^2) of Hb concentration was estimated to be 0.64 and 0.89 in Tibetans and Andeans, respectively (Beall et al. 1998). Likewise, SaO_2 showed moderate heritability in Tibetans (i.e., $h^2 = 0.4$) but no heritability in Andeans (Beall 2006). These values are not trivial compared with the estimated heritability of ~0.8 for height and 0.4 for body mass index (Elks et al. 2012; Silventoinen et al. 2003).

At 3658 m altitude, Tibetans ventilate as much as the newcomers acclimatized to high altitude and have larger resting ventilation and greater hypoxic ventilatory responses (HVRs) than do the non-native long-term residents (Zhuang et al. 1993; Huang et al. 1984). The Tibetan resting ventilation and HVR are higher than those of Andean Aymara at a comparable altitude (3800–4065 m), and it has been proved to be ancestry-associated in both populations (Brutsaert et al. 2005; Zhuang et al. 1993). The total lung volumes are reported to be larger in the Andeans and Tibetans relative to their lowland counterparts born and raised at high altitude (Brutsaert et al. 1999; Weitz et al. 2016), but those of Ethiopians have not been studied. The accelerated lung growth begins in childhood in Tibetans and mid- to late-adolescence in Andeans (Weitz et al. 2016). The elevated nitric oxide (NO) to dilate vessels and increase blood flow has also been hypothesized as a unique adaptive physiological trait in native highlanders (Beall et al. 1997; Groves et al. 1993; Erzurum et al. 2007; Levett et al. n.d.). However, a recent study contradicts this hypothesis as they found blunted NO regulation in Tibetans than in lowlander immigrants at high altitude (He et al. 2018). Another study even reported lower partial pressure of exhaled NO in Tibetans and Andeans, compared to sea level (Ghosh et al. 2019). These traits described are among many that involve in response to high-altitude hypoxia. They constitute a complex phenotypic network, in which they act as primary components and secondary effects of each other to reach an optimal state of body function. Although some of the physiological responses are still under debate and additional studies are in continuing demands, current results showing cross-ancestry and cross-altitude differences of physiological traits provide clues for further investigation of genetic basis of high-altitude adaptation.

3.4 Genetic Basis of High-Altitude Adaptation

The recent decade witnessed exponential growth of genomic studies on high-altitude adaptation, largely attributing to the development of genotyping and sequencing technologies that make it possible to perform genome-wide scans at relatively low cost. Here we review evidence that natural selection is acting or has acted on indigenous high-altitude populations, mostly Andeans, Tibetans, and Ethiopians. Earlier findings largely focus on limited number of genes and traits, while rapid growth of multi-omics data in recent years highlights the complexity of high-altitude adaptation.

*3.4.1 Two Well-Recognized Genes (***EPAS1** *and* **EGLN1***) and a Core Pathway (HIF)*

The years of 2010 and 2011 witnessed a major boost of genome-wide scan for natural selection in high-altitude populations. Most of the papers published during that period reported adaptive signals in Tibetans (Yi et al. 2010; Wang et al. 2011; Simonson et al. 2010; Beall et al. 2010; Xu et al. 2010; Peng et al. 2011; Bigham et al. 2010), and only a few focused on Andeans (Bigham et al. 2010, 2009). These studies are based on two complementary strategies. One is a de novo genome-wide screening for strong evidence of positive selection, which requires no prior knowledge of the candidate loci or phenotypes; the other depends on a priori list of candidate genes reported in similar studies or involved in hypoxia-related pathways. Signals of natural selection are indicated by two sources of evidence in these studies: one is significantly large allele frequency differences between highland and lowland populations that can be measured by *F*ST (Weir and Hill 2002); the other is the unexpected long stretches of haplotypes compared with those under neutrality, such as the integrated haplotype score (iHS) (Voight et al. 2006), and the cross-population extended haplotype homozygosity (XP-EHH) (Sabeti et al. 2007). The haplotype-based methods do not detect signals of positive selection with time scale as early as those based on allele frequencies (<30,000 years BP vs. 50,000–75,000 years BP) (Sabeti et al. 2006), because accumulated recombination would break down the haplotypes of populations that have long been living at high altitudes. Therefore, different methods would possibly provide inconsistent results, which should be carefully interpreted. Intriguingly, *EPAS1* (also known as *HIF2A*) encoding the hypoxia-inducible factor 2-alpha (HIF2α) and *EGLN1* coding for the prolyl hydroxylase domain-containing protein 2 (PHD2) are repeatedly highlighted in these studies as strong candidates (Fig. 3.1).

The EPAS1 protein is a transcription factor involved in the induction of genes when oxygen levels fall and is essential for embryonic development (Tian et al. 1998, 1997). Specific haplotypes and genotypes in *EPAS1* were reported to be associated with lower Hb concentrations and lower pulmonary vasoconstriction

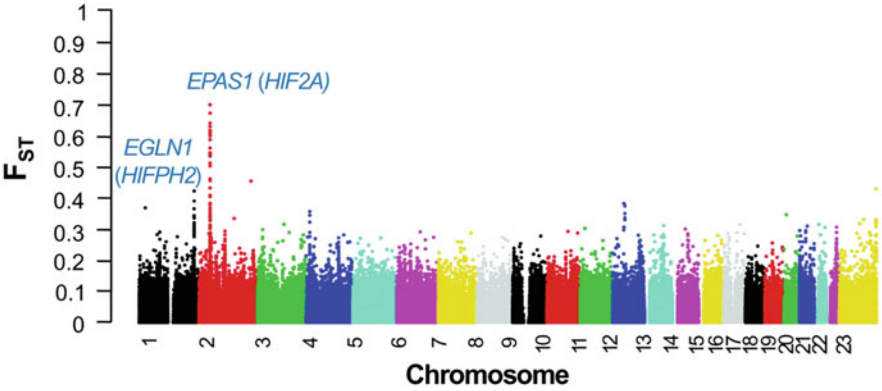


Fig. 3.1 *EPAS1* and *EGLN1* show significant high F_{ST} between Tibetan and Han Chinese populations in the genome-wide scan. From Xu et al. (2010)

response in Tibetans, as a protectant of polycythemia and pulmonary hypertension (Beall et al. 2010; Peng et al. 2017). Some genome-wide association studies (GWASs) linked the *EPAS1* variants to the birth weights of Tibetan newborns (Horikoshi et al. 2016; Xu et al. 2014). The heterozygous *EPAS1* knockout mice showed blunted physiological responses to chronic hypoxia, including less weight loss, lower Hb concentration, lower right ventricle pressure, and less right ventricle hypertrophy (Peng et al. 2017). Ho et al. (2012) found that two NO synthases (iNOS and eNOS) producing NO are the targets of *EPAS1* in mice, but it was contradicted by Peng et al. (2017) that no difference of blood NO concentration was observed between the *EPAS1*$^{+/-}$ and the wild-type mice under normoxia or hypoxia. Gene regulation is one possible mediator of the phenotypic effects of *EPAS1*. *EPAS1* is highly expressed in lungs and placentas, as shown in the Human Protein Atlas database (https://www.proteinatlas.org). The adaptive *EPAS1* haplotype was found to down-regulate expression in human umbilical endothelial cells and placentas in Tibetans (Peng et al. 2017). In Argentinian highlanders from the Andes, a novel missense variant mapped to *EPAS1* (rs570553380:A>G) was reported as a target of selection (Eichstaedt et al. 2017), but little is known about the variant effect on protein stability and phenotypic variation.

EGLN1 exhibits a signature of positive selection in both Tibetan and Andean populations (Bigham et al. 2010; Foll et al. 2014). It has a broad tissue expression. The EGLN1 protein encoded is an oxygen sensor and regulates the activation of HIF-α by its degradation, which has been proved in the knock-down experiments in many transformed cell lines (Berra et al. 2003a, b; Schofield and Ratcliffe 2004; Appelhoff et al. 2004). There are two key non-synonymous variants identified in this gene, rs12097901 (c.12C>G; Cys127Ser) and rs186996510 (c.380G>C; Asp4Glu) (Lorenzo et al. 2014; Xiang et al. 2013). Adaptive alleles at these loci are on the same adaptive haplotype specifically enriched in Tibetan highlanders, reaching a frequency of ~80%. They showed strong evidence of positive selection with an estimated onset of ~8000 years ago. Significant associations were detected between the two variants and Hb levels in Tibetans, indicating a possible mechanism for the abrogation of polycythemia in Tibetan highlanders. The rs12097901 variant also showed a strong association with percutaneous arterial SaO_2 responses to acute hypobaric hypoxia in a Japanese cohort (Yasukochi et al. 2018). Moreover, the key roles of *EGLN1* variants are supported by both in cellulo and in vivo experiments. The erythroid progenitor cells from Tibetan homozygous for rs12097901 and rs186996510 have decreased proliferation, erythropoietin sensitivity, colony size, and hemoglobinization at low oxygen level (Lorenzo et al. 2014; Xiang et al. 2013). Heterozygous knockout of *EGLN1* mutations leads to erythrocytosis and enhanced ventilatory responses to in mice (Arsenault et al. 2013; Bishop et al. 2013). However, the *EGLN1* adaptive alleles found in Tibetans are absent or at low frequency in Andean highlanders (Heinrich et al. 2019). Direct evidence linking *EGLN1* genotypes to phenotypes in Andeans is limited. A recent association study on Peruvian Quechua highlighted rs1769793 as it distinguishes the aerobic capacity between highlanders and lowlanders (Brutsaert et al. 2019).

The strong adaptive signals at *EPAS1* and *EGLN1* have been confirmed in various Himalayan populations from Nepal, Bhutan, North India, or Tibet (Zhang et al. 2017a; Arciero et al. 2018). *EGLN1* even stands out in North Caucasian highlanders living at a mild elevation (~2000 m above sea level) (Pagani et al. 2012b). However, no evidence was found that *EPAS1* or *EGLN1* variants associated with Hb concentration or SaO_2 in Ethiopian Amhara highlanders. Instead, an intergenic variant, rs10803083, was detected, and the effect size of this variant on Hb concentration (0.83 g/dL) is half that of the *EGLN1* SNPs (~1.7 g/dL) but comparable to that of the *EPAS1* SNPs in Tibetans (0.8 g/dL) (Alkorta-Aranburu et al. 2012; Simonson et al. 2010; Beall et al. 2010). Another study comparing the high- and low-altitude populations living in Ethiopia reported a set of candidate genes, including *CBARA1*, *VAV3*, *ARNT2,* and *THRB*. Several of these genes also show suggestive associations with Hb levels (Scheinfeldt et al. 2012). These studies suggest a different way of high-altitude adaptation in Ethiopian populations.

EPAS1 and *EGLN1* are central to the HIF pathway, which is the most important pathway for oxygen sensing and adaptability in almost all animals (Semenza 1999, 2004; Loenarz et al. 2011). The discovery of HIF pathway has been recognized as a groundbreaking work by the 2019 Nobel Prize in Physiology and Medicine (Zhang et al. 2019; Prabhakar 2020). HIF is a family of transcriptional complex, and each member is a heterodimer composed of an alpha and a beta subunits (Lisy and Peet 2008). In human, there are three paralogs of the HIF-α subunit (HIF-1α, HIF-2α, and HIF-3α, coded by *HIF1A*, *EPAS1*, and *HIF3A*, respectively) and two paralogs of the HIF-β subunit (HIF-1β and HIF-2β, coded by *ARNT* and *ARNT2*, respectively). The α subunits essentially determine the activity of HIF (Ivan et al. 2001; Jaakkola et al. 2001). Under normal oxygen level, HIF-α binds to PHDs (PHD1, PHD2, and PHD3, coded by *EGLN2*, *EGLN1*, and *EGLN3*, respectively) and is further targeted for degradation mediated by von Hippel-Lindau tumor suppressor (VHL) proteins; while under hypoxia, HIF-α is stabilized as the PHD activity decreases, and the HIF heterodimer further binds to the hypoxia response elements (HREs) of the targeted genes to initiate downstream reactions (Wenger et al. 2005).

Genes related to the HIF pathway have usually been the focus of the candidate gene analysis. A bunch of HIF-related genes have been investigated in studies conducted on gene-edited mice or human patients with erythrocytosis, as reviewed in Bigham and Lee (2014). These studies showed that HIF genes (e.g., *EPAS1*, *EGLN1*, and *HIF1A*) and HIF targeted genes (e.g., *ET-1*, *PDK1*, *EPO*, *VEGFA*, and *NOS2*) function in various organs in response to the high-altitude hypoxia. For example, HIFs induce angiogenesis by activating the vascular endothelial growth factor A (VEGFA) (Keith et al. 2011); PHD2 plays a critical role in controlling the *EPO* (glycoprotein hormone erythropoietin) transcription in the kidney, and thus regulates red blood cell mass (Lee and Percy 2011). HIF-related genes have also attracted great attention of genomic studies. Specific sequence polymorphisms of *HIF1A* were proposed to be associated with maximal oxygen consumption during exercise (Prior et al. 2003), and with hypoxia adaptation in Sherpas in comparison with lowland Han Chinese and Japanese (Liu et al. 2007; Suzuki et al. 2003). A non-synonymous variant within *HIF1A* was discovered as an adaptive locus in a highland population called Laks from Russia (Pagani et al. 2012b). However, the *HIF1A* sequences are conserved across altitudinal Andean populations (Hochachka and Rupert 2003). Genome-wide scans for positive selection did not highlight *HIF1A* as a strong candidate as *EPAS1*. One possible explanation is that *HIF1A* is expressed in all cell types and controls core responses that are intolerant to variations, whereas *EPAS1* is specifically expressed in certain cell types and thus would be more adaptable (MacInnis and Rupert 2011).

Candidate genes targeted by HIF include those involved in oxygen delivery and oxidative metabolism. For instance, *HMOX2* (Heme Oxygenase 2) harbors potentially adaptive variants in Tibetans (Simonson et al. 2010). An intronic *HMOX2* variant (rs4786504:T>C) showed a male-specific association with Hb levels. In vitro experiments showed that the derived allele C at rs4786504 could increase the expression of *HMOX2* and presumably lead to a more efficient breakdown of heme (Yang et al. 2016). *HMOX2* was identified to be a shared target of natural selection in two distinct Tibetan populations living at 4200 m and 4500 m, respectively (Wuren et al. 2014), but showed neither elevated allele frequency nor significant association with hematocrit, arterial pressure, SaO_2, or hypoxia ventilatory response in Quechua Andean highlanders living at comparable altitudes (4350 m) (Yip et al. 2018). A case of candidate gene contributed to energy metabolism under hypoxia is *PPARA*. The putatively advantageous alleles in *PPARA* enriched in Sherpas appeared to enhance efficiency of oxygen utilization and protect against oxidative stress (Horscroft et al. 2017). A similar case is *PRKAA1* identified in Andeans. The *PRKAA1*-rs1345778 genotypes differentiated the transcriptional activity of metabolic pathways (Bigham et al. 2014), but no phenotypic association has been reported.

Early genetic evidence strongly supported the essential roles of *EPAS1*, *EGLN1*, and other HIF-related genes in the high-altitude adaptation. These studies rely largely on custom SNP genotyping arrays or sequencing of targeted genomic regions scattered throughout the genome without a full capture of genome variations. It should be noted that only a few adaptive variants identified in highland populations located in the coding regions. Protein-coding DNA sequences and non-coding sequences have been proposed to play distinct roles in the adaptive evolution of different gene ontologies (Haygood et al. 2010). Whether the importance of coding changes in human high-altitude adaptation should be underscored needs to be assessed based on comprehensive surveys of genome-wide adaptive variants. Furthermore, these genomic findings could not fully explain the adaptive phenotype variations and different adaptive mechanisms of various highland populations. Whether other types of variants, such as structural variants (SVs), or other genes also underpin the genetic adaptation of high altitudes remains a concern.

3.4.2 A Broader Perspective from Multi-omics Resources

Recent employment of whole-genome sequencing (WGS) tools and multi-omics evaluation have brought vast new data and have afforded greater resolution for multiple layers of genetic adaptation to high altitudes. Here we describe some new insights from the perspectives of genomics, transcriptomics, and epigenomes. Advances in knowledge of Tibetan's genetic adaptation far exceed those of Andeans (Zhou et al. 2013) and Ethiopians (Udpa et al. 2014).

Genomic variant analysis has been greatly facilitated by the integration of large-scale omics evaluation. One successful example is a genome-wide study of ~3000 Tibetans and ~7000 non-Tibetan individuals of East Asian ancestry, in which the phenotypic effects of the adaptive variants were assessed using 91 quantitative traits, including morphological, blood biochemistry, and optometric measures (Yang et al. 2017). Another outstanding one compared Tibetan highlanders and Han Chinese lowlanders by compiling deep-sequenced genomes (30×), RNA-Seq transcriptomes of placentas, and quantitative traits (Deng et al. 2019). Although different omics data in this study were not generated from identical individuals, it is fair enough to investigate the differentiation on the population level. The growing wealth of genomic data presents the challenge of extracting signals of positive selection as different statistics vary in their power and may return inconsistent results. A common approach to identify high-altitude adaptive signals in previous studies is to overlap the outliers in several independent tests, which only captured genetic loci under strong selection (Lotterhos and Whitlock 2015). Two recent investigations of genome sequences of Tibetan highlanders have applied a composite measure combining *p*-values obtained from multiple tests (Deng et al. 2019; Hu et al. 2017). This method, composed of multiple signals (CMS), was originally developed by Grossman et al. (2010) and was modified when performed by Hu et al. (2017) and Deng et al. (2019). It is possible that more statistics calculated for multi-dimensional data, such as variant effects or conservation (Deng et al. 2019; Szpak et al. 2018), can be integrated to improve the prioritization of candidate genes.

These efforts presented new findings in the causative factors of high-altitude adaptation. Seven of the nine adaptive loci detected in Yang et al. (2017) are novel, and the most outstanding one is a folate-increasing allele of rs1801133 at *MTHFR*, which is probably adaptive to high UV radiation. Hu et al. (2017) reported *PTGIS*,

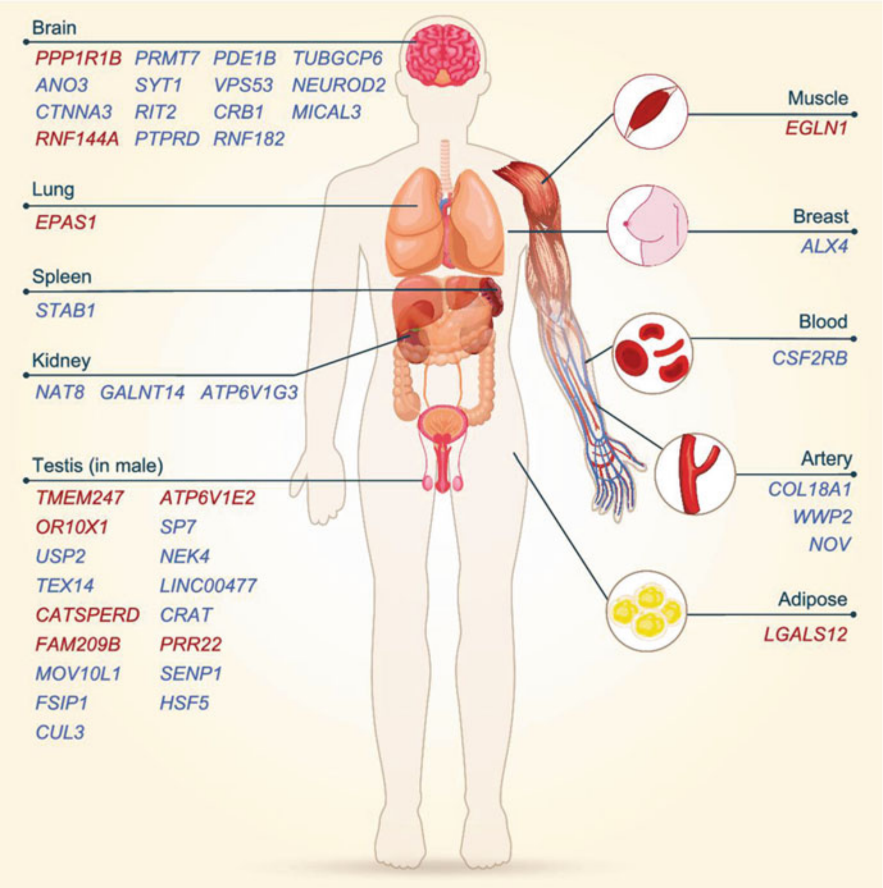


Fig. 3.2 Tissue-specific expression of the candidate genes of high altitude adaptation. The expression profiles for these genes were obtained from the genotype-tissue expression (GTEx) database. Genes showing significant associations with Tibetan phenotypes are highlighted in red. From Deng et al. (2019)

VDR, and *KCTD12* as new candidate genes. In particular, Deng et al. (2019) constructed a map of adaptive genetic variants of Tibetan highlanders involving multiple tissue-specific gene expression and phenotypes (Fig. 3.2). Intriguingly, rs116983452 (c.248C>T; p.Ala83Val) in *TMEM247*, which is ~40 kb downstream to *EPAS1*, presented as the top signal of selection in the Tibetan genome. It is so far the most highly differentiated missense variant between Tibetans and the lowlanders and is the only functional variant identified around *EPAS1* with transcriptional and phenotypic effects. Statistical modeling showed that rs116983452 has greater effect size and better predicts the phenotypic outcome than any *EPAS1* variants in the association with adaptive traits in Tibetans (Deng et al. 2019). These findings pinpointed a series of adaptation events that have been neglected so far and indicated that multiple variants may jointly deliver the fitness of Tibetans on the Plateau.

SVs including many different types of chromosomal rearrangement and encompassing millions of bases have appreciable impacts on human genome diversity and phenotype evolution (Hurles et al. 2008; Chiang et al. 2017). Their impact on high-altitude adaptation was first realized with a 3.4-kb Tibetan-enriched deletion (TED) downstream to *EPAS1* identified based on microarray-based analyses (Lou et al. 2015) and has been much better quantified using WGS strategies. Taking advantage of the long-read genome sequencing platforms and optical mapping (Munroe and Harris 2010; Yuan et al. 2020), Ouzhuluobu et al. (2020) built a de novo reference genome assembly of Tibetan and further explored a large number of Tibetan-specific SVs that have been underinvestigated so far. These novel SVs provide exciting prospects in understanding high-altitude adaptation. One notable example is a 163-bp deletion within an intron of *MKL1*, which encodes megakaryoblastic leukemia 1 protein and has been associated with systolic pulmonary arterial pressure, red blood cell count, and platelets in Tibetans (Ouzhuluobu et al. 2020). A similar effort was made by Quan et al. (2020) that generated a comprehensive catalog of SVs in Tibetans using nanopore sequencing technology, alternative to the single-molecule real-time DNA sequencing method applied in Ouzhuluobu et al. (2020). Although much remains to be learned about the evolutionary role of SVs in high-altitude adaptation, it is becoming increasingly clear that they can no longer be simply ignored.

Previous studies provided clues of human blunted responses to hypoxia at transcriptomic level (Peng et al. 2017; Leon-Velarde and Mejia 2008; Gaur et al. 2020; Piperno et al. 2011). In particular, most of the so far reported variants under positive selection are located in non-coding regions, suggesting the need for understanding the role of gene regulation in high-altitude adaptation. Hypoxia adaptation is a complicated phenotype involving multiple tissues. Elucidating the dynamics and mechanisms of genomic action on different tissues and the eventually constructed large-scale regulatory network would be crucial to establish the genotype–phenotype relationship. Using cultured human umbilical vein endothelial cells (HUVECs), Xin et al. (2020) generated regulatory element maps across time series under both hypoxic and normoxic conditions, and then constructed a hypoxia regulatory network by integrating expression and chromatin accessibility data (Fig. 3.3). They discovered causal non-coding SNPs in *EPAS1* and many other interesting genes, i.e., *GCH1*, *NRP2*, *NQO1*, *NOTCH1*, *NOS3*, *HYOU1*, *BNIP3*, and *BCL6*, etc., and further illustrated the profound role of *EPAS1* in hypoxic responses and angiogenesis. In addition, the transcriptomic landscape has also been revealed in high-altitude animals and plants (Hao et al. 2019; Qi et al. 2019; Gurung et al. 2019; Liu et al. 2020; Long et al. 2019). All these studies emphasize the transcriptional route of high-altitude adaptation. In this regard, systematic transcriptional regulation shaping the high-altitude phenotypes needs to be modeled in different highland human populations.

In addition to the nucleotide changes that may alter genome functions, epigenetic modifications, including DNA methylation, histone modification, and non-coding

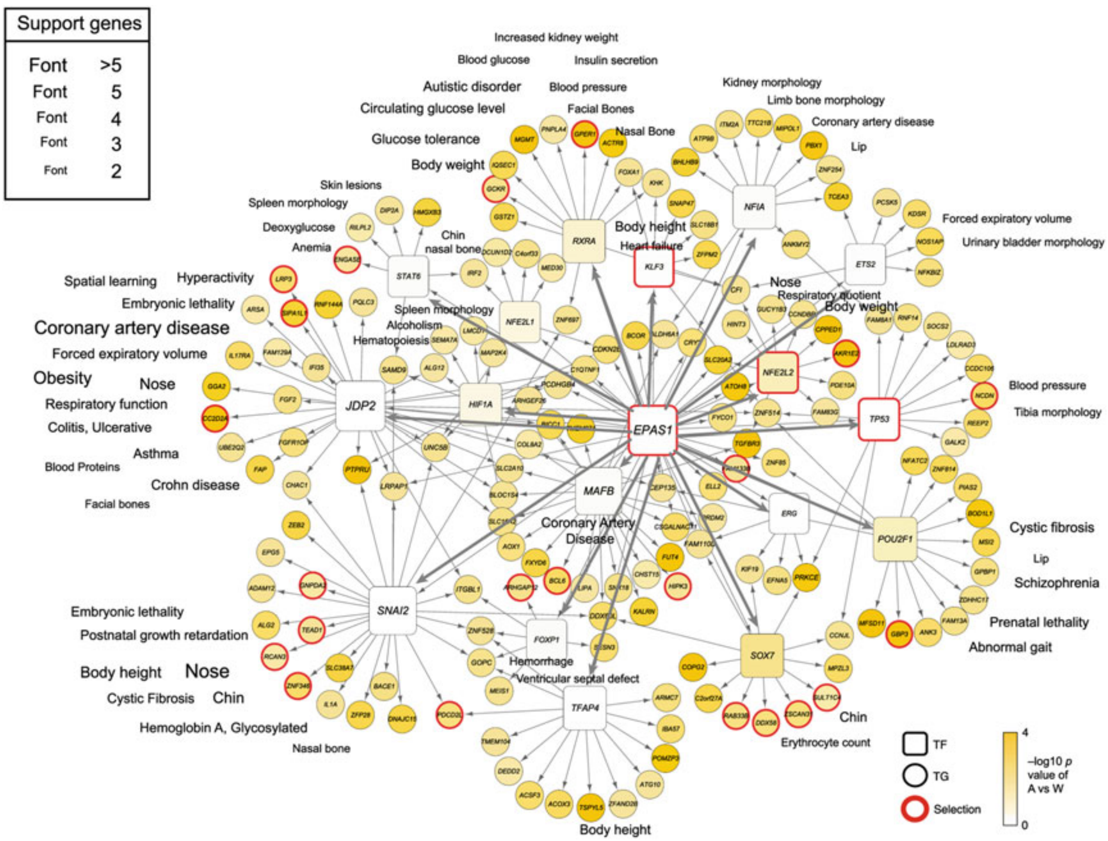


Fig. 3.3 The *EPAS1*-oriented regulatory network associates various phenotypes. Rectangles refer to transcription factors (TFs) and circles represent target genes (TGs). Colors denote the $-\log_{10}$(p-value) of *t*-test of gene expression between adaptive (A) and wildtype (W). From Xin et al. (2020)

RNA (ncRNA)-associated gene silencing (Egger et al. 2004), could also affect gene expression, and thus contribute to phenotypic variations. Some epigenetic signals are inheritable across generations and may produce adaptive phenotypes (Furrow 2014; Kronholm and Collins 2016), while others are sensitive to environmental exposure and could change over lifetime, resulting in phenotypic plasticity (Kelly et al. 2012; Richards 2006; Forsman 2015). Both types are expected to form the basis of high-altitude adaptation (Cheviron et al. 2012; Storz et al. 2010). Under tumor hypoxia, HIF-related genes, such as *HIF1A*, *EPAS1,* and *EPO*, are transcriptionally regulated by epigenetic processes (Rawluszko-Wieczorek et al. 2014; Lachance et al. 2014; Steinmann et al. 2011). Epigenomic annotation was included in constructing the regulatory network of Tibetan in Xin et al. (2020). A recent discovery indicated that exposure to high-altitude hypoxia has a decreasing effect on *EPAS1* methylation and an increasing effect on LINE-1 methylation in Andean Quechua (Childebayeva et al. 2019). Using the Comprehensive High-Throughput Relative Methylation method (Irizarry et al. 2008), Julian (2017) conducted a genome-wide methylation study in the peripheral blood mononuclear cells of Andean males living at 3600 to 4100 m. They identified several differentially methylated regions at base-pair resolution in samples with excessive erythrocytosis, and the most notable signal located around rs12097901, which is a key variant showing evidence of high-altitude selection in Tibetans (Lorenzo et al. 2014). Yet, the epigenetic components in Ethiopian adaptive responses to high altitudes are not known. Although confronted with challenges, which may derive from the nature of epigenome, difficulties in genome–epigenome integration, and technical limitations of site-specific methylation editing (Julian and Moore 2019), large-scale epigenomic studies will definitely expand our knowledge of non-sequence-based genomic contribution to the high-altitude physiological features.

3.4.3 “Borrowed Fitness” from Archaic Hominins

The archaic hominins, although now extinct, overlapped temporally and spatially with modern humans and are evidenced to influence modern humans genetically. Previous analyses attributed 1–3% of most non-African genomes to Neanderthal ancestry (Prufer et al. 2017, 2014; Green et al. 2010), <1% of that to Denisovan ancestry (Meyer et al. 2012), and perhaps some to other unknown archaic lineages (Hubisz et al. 2020; Mondal et al. 2016; Tucci et al. 2018). Archaic admixture has also been detected in Africans according to recent studies (Chen et al. 2020; Durvasula and Sankararaman 2020; Hammer et al. 2011). A large number of archaic alleles appear to have deleterious effects and are subject to purifying selection in modern humans (Petr et al. 2019; Sankararaman et al. 2014), while the surviving archaic sequences affect a variety of modern human phenotypes and are possibly under positive selection. Well-documented examples of the so-called adaptive introgression include immune function, pigmentation, height, diet, and metabolism

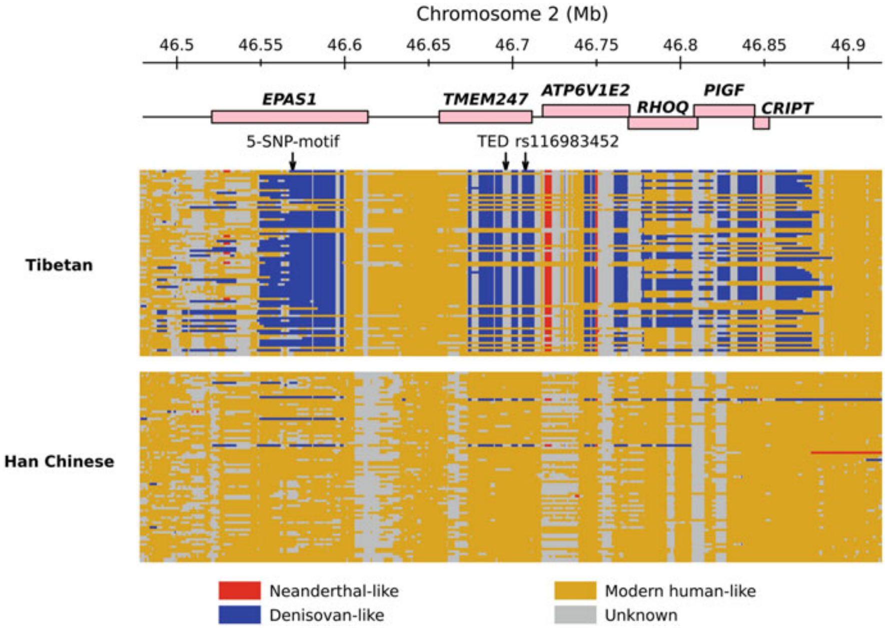


Fig. 3.4 Source of ancestry inferred for the EPAS1 region. Several important loci reported in previous studies are indicated with vertical arrows, including three Tibetan-specific variant tags, i.e., the 5-SNP-motif (Huerta-Sánchez et al. 2014), TED (Locke et al. 2015), and rs116983452 (Deng et al. 2019). Each row represents a haplotype of Tibetan or Han Chinese, with ancestry inferred to be derived from Neanderthal, Denisovan, modern human, or uncertain lineages. Adapted from Deng et al. (2019)

(Sankararaman et al. 2014; Mendez et al. 2012; Ding et al. 2014; Enard and Petrov 2018; Vernot and Akey 2014; Khrameeva et al. 2014; Dannemann and Kelso 2017).

A 5-SNP-motif in a 32-kb region in *EPAS1* is the first candidate of adaptive archaic introgression in the high-altitude populations (Huerta-Sanchez et al. 2014). This motif is located in the intronic region, suggesting possible effects on transcriptional regulation of *EPAS1*. It presents at high frequency (~80%) in Tibetans and as homozygotes in the Denisovan genome, but is rare in Han Chinese and almost missing in other worldwide populations. The Denisovan sequence at this region closely matches with that of Tibetans on the haplotype network. It is not likely to be caused by incomplete lineage sorting considering the long haplotype length and the unusually high frequency, nor is it expected from mutation, genetic drift, or directional selection alone. Therefore, it appears to be the consequence of Denisovan introgression. A fine-scale inference of the local ancestry at the *EPAS1* region surprisingly identified a complicated archaic ancestry pattern, showing a mix of Denisovan, Neanderthal, and unknown archaic ancestry elevated specifically in Tibetans (Fig. 3.4) (Lu et al. 2016). The Denisovan-like beneficial sequences stretch ~300 kb and cover 6 protein-coding genes. Further estimation of haplotype age, ~62,000–38,000 years ago, supports the ancient origin of it. One notable variant constituting the haplotype is *TMEM247*-rs116983452-T, a key functional variant identified around *EPAS1* (Deng et al. 2019).

Detection of adaptive introgression at *EPAS1* is a great breakthrough in this field and provides special insights into the understanding of high-altitude adaptation, but many questions remain to be answered. An intuitive one is whether there are other adaptive archaic sequences in Tibetans. Based on the WGS data, the overall proportion of archaic sequences in Tibetans (6.17%) was estimated to be significantly higher than that in Han Chinese (5.86%) (Lu et al. 2016). This observation largely depends on the local genomic ancestry inferred using the commonly used *S** statistic (Vernot and Akey 2014) and a later method called *ArchaicSeeker (*Lu et al. 2016*)*. Both methods assume that no archaic introgression occurred in sub-Saharan Africans, and alleles absent in sub-Saharan Africans are potential archaic-like alleles. The *S** statistic examines the linkage disequilibrium (LD) of archaic-like alleles; *ArchaicSeeker* estimates the probability of archaic introgression resulting in the observed rate of archaic-like alleles in a given segment. The larger proportion of archaic sequences in Tibetans than in Han Chinese was confirmed using the long-read sequencing data (Ouzhuluobu et al. 2020). A systematic assessment of the sequences absent from the human reference genome showed much more sharing between archaic hominid (Altai Neanderthal and Denisovan) and Tibetan (1.32–1.53%) compared to other high-quality East Asian genomes, such as a Korean genome (Seo et al. 2016) (0.70–0.82%) and a Han Chinese genome (Shi et al. 2016) (0.85–0.98%). One notable example is a Tibetan-specific 662-bp intronic insertion in *SCUBE2*, which is enriched and associated with better lung function in Tibetans (Ouzhuluobu et al. 2020). In addition, Hu et al. (2017) reported several genomic regions with higher Denisovan ancestry in Tibetans than in Han Chinese. These signals are putative candidates of adaptive introgression and deserve further investigations.

Another key question is how Tibetans acquired those archaic sequences from Denisovans. One possible scenario is that Denisovan introgression occurred at lowlands, prior to the split of Han Chinese and Tibetans, and positive selection leads to the subsequent accumulation of the adaptive archaic sequences in Tibetans. Does it suggest that Denisovan or Denisovan-related ancient groups also adapted to high altitude? It is possible, but the scenario could be very much more complicated than our current understanding. For instance, the adaptive effect of the archaic sequences could be a consequence of adaptation after a long period of exposure to the extreme environment on the Tibet Plateau since they were induced to Tibetans' ancestors. In this case, it is not necessarily that Denisovan adapted to high altitude, but rather the group (or ancestors of Tibetans) who received the Denisovan sequence reserved it after experienced natural selection. Alternatively, the adaptive archaic sequences in Tibetans were likely inherited from Paleolithic settlers at highlands and were subsequently introduced to Han Chinese via recent gene flow. Inspired by the mixed ancestry architecture at *EPAS1*, Lu et al. (2016) proposed a two-wave "Admixture of Admixture" model, which characterizes the Paleolithic ancestry at Tibetan Plateau as a mixture of ancient Siberian modern human and several known and unknown archaic populations. This hypothesis is supported by a late middle

Pleistocene Denisovan mandible discovered from the Tibetan Plateau (Chen et al. 2019). The driving forces of the ancient settlement at the environmentally inhospitable highlands are not known yet. Whether the Tibetan Plateau acted as a shelter for human during the Last Glacial Maximum is still controversial.

The evidence of adaptive introgression in Tibetan highlanders facilitates our understanding of high-altitude adaptation as a process of selection on standing variations, although it brings about challenges in dating the onsets of selection as most of the existing methods are based on a model of selection on de novo mutations (Peng et al. 2011; Jeong et al. 2014). More work needs to be done to demonstrate the post-introgression dynamics of these archaic haplotypes, for instance, whether they reached high frequencies immediately after the introgression event occurred due to selection, or they initially segregated neutrally and subsequently underwent positive selection (Jagoda et al. 2018). In addition, it needs to be addressed whether adaptive archaic introgression is a common strategy adopted by different geographical highlanders, as single cases in Tibetans could not represent the full picture of the "borrowed fitness" in highland human populations.

3.4.4 Convergent Adaptation

High-altitude adaptation is an ideal genotype–phenotype model to understand convergent evolution as it incorporates distantly related human populations and non-human species that have attained fitness for the long residence under extreme hypoxia. It can be strictly defined as a convergent reproductive success across lineages through shared and distinct physiological changes. Underlying each of the physiological traits, there could be convergent or divergent genetic mechanisms. Genetic convergence can be assessed at various levels, ranging from strictly identical mutations accumulated independently in different lineages, to different mutations in the same genes, and an even loosened criteria only requiring evolutionary changes in the same pathways. Disentangling the roles of convergent and non-convergent evolution, and further distinguishing different patterns of convergent evolution on de novo mutations or standing variations is a challenging task. So far, the detection of similar genetic adaptation to high altitudes has largely relied on genome-wide selection or association scans. Here we summarize the state of knowledge reviewed by Witt and Huerta-Sanchez (2019) last year and focus on current understanding of human populations. These insightful studies provide evidence for shared and distinct genetic solutions to survive under chronic hypoxia.

A map of convergent evolution among human and non-human species on three continental highlands is shown in Fig. 3.5. Overlapping the genome-wide selection signals across populations located *EGLN1* as a candidate of convergent evolution in Andeans and Tibetans, while the dominant haplotype is unique to each population (Bigham et al. 2010). Further studies found that the Hb-associated variants in Tibetans, including rs186996510, the only functional variant with experimental evidence, were rare or had no phenotypic effect in Andeans (Heinrich et al. 2019;

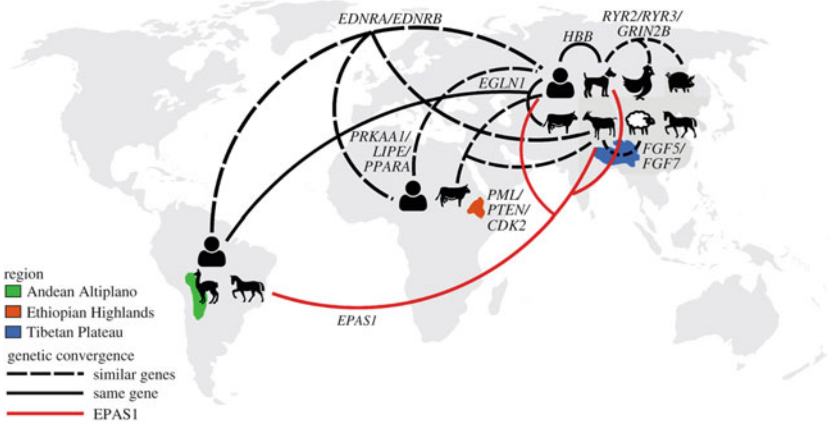


Fig. 3.5 Sharing of adaptive genes among high-altitude populations. Genomic studies conducted on highland species inhabiting the Andean Altiplano, the Ethiopian highlands, and the Tibetan Plateau are highlighted on the map. Genetic convergence identified across multiple populations are indicated with solid lines (adaptation on the same gene, with *EPAS1* highlighted in red) and dashed lines (adaptation on different genes). From Witt and Huerta-Sanchez (2019)

Bigham et al. 2013; Lawrence et al. 2018). Five significant SNPs in the non-coding regions of *EGLN1* or outside *EGLN1* were identified in the Andean Quechua in a recent GWAS and covariance study (Brutsaert et al. 2019). Further studies are needed to check whether these variants also function in Tibetans. Another target of convergent adaptation is *EDNRA*. It is originally identified through neutrality tests in Andeans (Bigham et al. 2009) and shows significant association with high-altitude polycythemia risk in Tibetans (Xu et al. 2015). A closely related endothelin receptor, *EDNRB*, was identified to be a candidate for high-altitude tolerance in Ethiopians in a whole-genome sequencing study (Udpa et al. 2014). *EDNRA* and *EDNRB* are known to play important but opposite roles in cardiovascular function, *EDNRA* in vasoconstriction and *EDNRB* in vasodilatation (Schneider et al. 2007). A mouse model suggested low expression of *EDNRB* increases fitness to hypoxia (Stobdan et al. 2015). Despite that only a few genes were found to be subject to altitudinal convergent adaptation in numerous studies, Foll et al. (2014) applied a hierarchical Bayesian approach and showed widespread signals (31% of all adaptive SNPs at a false-positive rate of 1%) of convergent evolution between Tibetans and Andeans. Their models not only confirmed the signal at *EGLN1* in both populations, but also suggested possible cross-population effects of some well-known genes, e.g., *EPAS1*. Functional enrichment analysis further suggested adaptive response to hypoxia-induced toxic effects on fatty acids degradation and neuron system.

Some genes are shared targets of selection between human and non-human species, suggesting an extraordinary landscape of high-altitude adaptation, although the specific variants differ across lineages. *EGLN1* was identified in human populations from Andes and Tibet, as well as Tibetan cattle and ducks (Graham and McCracken 2019; Wu et al. 2018); *HBB* shows signals of positive selection in humans and dogs of the same environment on Tibetan Plateau (Wang et al. 2014). *EPAS1* is the most commonly reported adaptation-associated gene in a variety of species, including Tibetan human and domesticated animals like Andean horse, Tibetan dog, Tibetan Chicken, and Cashmere goat (Li et al. 2014, 2017; Liu et al. 2019; Song et al. 2016; Gou et al. 2014), as well as wild animals on Tibetan plateau, such as wolf, hot-spring snake, frog, snow leopard, and plateau zokor (Li et al. 2018; Zhang et al. 2014, 2016; Yun Sung Cho et al. 2013; Shao et al. 2015). Key variants in *EPAS1* have been identified in these species, but only a few have been functionally investigated in cellular experiments to connect to biological effects in hypoxia (Pamenter et al. 2020). For instance, the N203H variant significantly affects HIF2 transcriptional activity and protein stability in the AC16 human cardiomyocyte cell line and has been linked to nonsyndromic congenital heart disease in Tibetans (Pan et al. 2018). None of the *EPAS1* variants has been validated *in vivo*.

Convergent evolution has also been observed at the pathway level. Annotation of adaptive genes using Gene Ontology (GO) suggested that many high-altitude populations share similar GO terms (Witt and Huerta-Sanchez 2019). There are some notable examples in human populations. For example, *PPARA* and *PTEN* in Tibetans and *TBX5* in Andeans are involved in regulation of cardiac muscle tissue development (GO:0055024, annotated using clusterProfiler version 3.10.1 (Yu et al. 2012)), which is related to the cardiovascular functions in hypoxia (Velotta et al. 2018); *CYP2E1* and *PTGIS* in Tibetans and *LIPE* in Ethiopians are related to fatty acid metabolism (GO:0001676), which regulates energy utilization at high altitudes (Ge et al. 2012); *THRB* in Ethiopians and *TNC* in Andeans function in respiratory system development (GO:0060541); *HBB* in Tibetans and *RORA* in Ethiopians play roles in NO biosynthesis and metabolism (GO:0006809 and GO:0046209).

It is more likely that convergent adaptation with a shared genetic basis often acts on existing variations (Sackton and Clark 2019). Genetic admixture with an already adapted population is a more efficient way to obtain fitness than random processes. *EPAS1* is amongst the most important genes showing repeated evidence of adaptation via gene flow, from Denisovan to Tibetan human populations (Huerta-Sanchez et al. 2014), and from Tibetan wolves to Tibetan Mastiff (Miao et al. 2017). It seems that *EPAS1* is more easily to get evolve under selective pressure and to be spread across species, while the underlying mechanism has never been figured out yet. Other instances include the *EGLN1* loci introgressed from yak to Tibetan cattle (Wu et al. 2018), and an adaptive *BHLHE41* variant introduced to Ethiopian Oromo by admixture with Ethiopian Amhara (Huerta-Sanchez et al. 2013). All these evidences indicate adaptive introgression on single locus. Polygenic basis of adaptive introgression has been suggested by immunity and metabolic functions, and in most cases it would remain below the detection threshold of most methods (Gouy and Excoffier 2020). To what extent it would affect chronic hypoxia responses needs to be examined in future studies.

Coexistence of convergent and independent genetic responses to high-altitude subsistence informs the complexity of adaptive evolution. It’s hard to expect that a common genetic route has been acquired by people with different genetic background to settle different continental highlands with various altitudes and environmental niches. Differentiated adaptive genetic variants were identified even between Sherpa and Tibetan, two closely related groups at Tibetan Plateau (Zhang et al. 2017a). Meanwhile, the extremely hostile environments on the plateaus do not allow for easy evolvement by large random feasible ways, leading to some key genes and pathways repeatedly emphasized in different high-altitude species. Potential confounders and unforeseen variable should be taken with caution when interpreting the apparent convergence or non-convergence inferred from genomic studies, such as epistasis effect, cross-trait interactions, and methodological limitations. Understanding the genetic basis of convergent adaptation and quantifying the degree of genetic convergence underlying phenotypic convergence may facilitate our understanding of the general rules of evolution.

3.5 Current Limitations and Future Directions

Natural experiments conducted in the high-elevation environments offer great opportunities to assess how humans react when facing long-term environmental challenges. As reviewed above, human inhabitation at high altitudes depends on tremendous physiological and underlying genetic capabilities. However, limitations exist in current studies and our knowledge of the genetic predisposition of high-altitude adaptation, especially at the variant level, remains incomplete.

Firstly, the adaptive phenotypes at chronic hypoxia have not been fully measured. Limitations of techniques and sampling conditions make it difficult to ascertain the phenotypes of interest systematically and accurately. Most of the previous studies concentrated on a small number of phenotypes related to the oxygen transport traits, such as ventilation volume, SaO_2, Hb, and blood flow. Some efforts have also been made to investigate how the metabolic system responses to maximize the oxygen utility efficiency. However, it is still uncertain whether the values observed are adaptive values, what physiological mechanisms might be involved, how these traits collectively affect the reproductive fitness, and which is the key trait of adaptive evolution. Moreover, most of the adaptive phenotypes were characterized in Tibetans and Andeans, and would be inappropriate to apply to Ethiopians and other minor high-altitude groups.

Secondly, only a few genetic variants have been linked to adaptive phenotypes. WGS is becoming low-cost and routine in screening for adaptive variants of phenotypic evolution, along with advancement in genomic statistical methodologies. Compared to the conventional array technologies, WGS provides dense variants information and brings a breakthrough in detecting variants that had been acted upon by selection at high altitudes. However, there is a remaining gap between the genomic significance of variants and the final assurances with high confidence of phenotypic effects. Especially, most of the reported adaptive variants are located in non-coding regions, and thus would possibly lead to indirect consequences that are difficult to measure. One of the major tasks of subsequent studies is to consolidate and expand the genotype–phenotype interactions, taking use of multi-omics resources. A complex model is needed to elucidate the adaptive evolution at highlands. For example, an omnigenic model has been suggested for blood cell traits and relevant diseases (Vuckovic et al. 2020), and to which extent it applies to the high-altitude populations needs to be further tested.

Thirdly, most of the current findings are based on SNPs, and uncovering the roles of different types of variants is necessary in future work to address the missing heritability in high-altitude adaptation. Though SNPs are considered as a main source of human genetic and phenotypic variations, the contribution of SVs has also been well-appreciated (Hurles et al. 2008; Chiang et al. 2017). The de novo genome assembly of Tibetan demonstrated that a large number of SVs that have never been surveyed in other studies have great potential in determining Tibetan phenotypes. More follow-up studies would be expected to give additional insights into the adaptive basis from a genomic perspective.

Lastly, despite the preponderance of the three major high-altitude groups, i.e., Andeans, Tibetans, and Ethiopians, in so far conducted studies, investigations of other minor highland tribes with diverge genetic background and living conditions are crucial to construct a full picture of high-altitude adaptations. The repeated high-altitude experiments have been carried out independently in different species, and thus might have led to large uncertainties and complexities in the results. For example, different species in the same environment could have genetic convergence, while geographical populations of the same species show distinct genetic patterns. This emphasizes the necessity to comprehensively survey diverse samples for answering the key question that how many genetic solutions are available for success settlements at high altitudes.

Studies on high-altitude adaptation are ideally situated for decoding the genetic factors that determine the phenotypes under extreme environmental pressures. They may also imply genetic susceptibility to high-altitude-related diseases and possible therapeutic strategies. The currently unaccounted heritability of high-altitude adaptive traits would be explained with the help of the advanced sequencing technologies, genome engineering tools and models, and statistical methodologies (Moore 2017; Hall et al. 2020). Moreover, the importance of linguistic, ethnographical, and archeological data of the high-altitude indigenous people should never be overlooked when we interpret the results of genome analyses. It is not easy to reveal the mystery of human genomic adaptations to high altitudes. Research is underway, and discovery takes time.

Acknowledgements This study was supported by the National Science Fund for Distinguished Young Scholars (31525014), the National Natural Science Foundation of China (NSFC) (31771388, 91731303, 32030020, 31961130380, and 32041008), the Strategic Priority Research Program (XDB38000000) and Key Research Program of Frontier Sciences (QYZDJ-SSW-SYS009) of the Chinese Academy of Sciences (CAS), the UK Royal Society-Newton Advanced Fellowship (NAF\R1\191094), the National Key Research and Development Program (2016YFC0906403), and the Shanghai Municipal Science and Technology Major Project (2017SHZDZX01). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests The authors declare that they have no competing interests.

References

Aldenderfer M (2003) Moving up in the world: archaeologists seek to understand how and when people came to occupy the Andean and Tibetan plateaus. Am Sci 91:542–549

Aldenderfer M (2011) Peopling the Tibetan plateau: insights from archaeology. High Alt Med Biol 12:141–147. https://doi.org/10.1089/ham.2010.1094

Alkorta-Aranburu G et al (2012) The genetic architecture of adaptations to high altitude in Ethiopia. PLoS Genet 8:e1003110. https://doi.org/10.1371/journal.pgen.1003110

Altundag A et al (2014) The effect of high altitude on olfactory functions. Eur Arch Otorhinolaryngol 271:615–618. https://doi.org/10.1007/s00405-013-2823-3

Appelhoff RJ et al (2004) Differential function of the prolyl hydroxylases PHD1, PHD2, and PHD3 in the regulation of hypoxia-inducible factor. J Biol Chem 279:38458–38465. https://doi.org/10.1074/jbc.M406026200

Arciero E et al (2018) Demographic history and genetic adaptation in the Himalayan region inferred from genome-wide SNP genotypes of 49 populations. Mol Biol Evol 35:1916–1933. https://doi.org/10.1093/molbev/msy094

Arsenault PR et al (2013) A knock-in mouse model of human PHD2 gene-associated erythrocytosis establishes a haploinsufficiency mechanism. J Biol Chem 288:33571–33584. https://doi.org/10.1074/jbc.M113.482364

Bartsch P, Saltin B (2008) General introduction to altitude adaptation and mountain sickness. Scand J Med Sci Sports 18(Suppl 1):1–10. https://doi.org/10.1111/j.1600-0838.2008.00827.x

Beall CM (2006) Andean, Tibetan, and Ethiopian patterns of adaptation to high-altitude hypoxia. Integr Comp Biol 46:18–24. https://doi.org/10.1093/icb/icj004

Beall CM, 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. https://doi.org/10.1002/ajpa.1330730404

Beall CM, Reichsman AB (1984) Hemoglobin levels in a Himalayan high altitude population. Am J Phys Anthropol 63:301–306. https://doi.org/10.1002/ajpa.1330630306

Beall CM et al (1997) Ventilation and hypoxic ventilatory response of Tibetan and Aymara high altitude natives. Am J Phys Anthropol 104:427–447. https://doi.org/10.1002/(SICI)1096-8644(199712)104:4<427::AID-AJPA1>3.0.CO;2-P

Beall CM et al (1998) Hemoglobin concentration of high-altitude Tibetans and Bolivian Aymara. Am J Phys Anthropol 106:385–400. https://doi.org/10.1002/(SICI)1096-8644(199807)106:3<385::AID-AJPA10>3.0.CO;2-X

Beall CM et al (2002) An Ethiopian pattern of human adaptation to high-altitude hypoxia. Proc Natl Acad Sci U S A 99:17215–17218. https://doi.org/10.1073/pnas.252649199

Beall CM et al (2010) Natural selection on EPAS1 (HIF2alpha) associated with low hemoglobin concentration in Tibetan highlanders. Proc Natl Acad Sci U S A 107:11459–11464. https://doi.org/10.1073/pnas.1002443107

Berra E et al (2003a) HIF prolyl-hydroxylase 2 is the key oxygen sensor setting low steady-state levels of HIF-1alpha in normoxia. EMBO J 22:4082–4090. https://doi.org/10.1093/emboj/cdg392

Berra E et al (2003b) HIF prolyl-hydroxylase 2 is the key oxygen sensor setting low steady-state levels of HIF-1alpha in normoxia. EMBO J 22:4082–4090. https://doi.org/10.1093/emboj/cdg392

Bigham AW, Lee FS (2014) Human high-altitude adaptation: forward genetics meets the HIF pathway. Genes Dev 28:2189–2204. https://doi.org/10.1101/gad.250167.114

Bigham AW et al (2009) Identifying positive selection candidate loci for high-altitude adaptation in Andean populations. Hum Genomics 4:79–90. https://doi.org/10.1186/1479-7364-4-2-79

Bigham A et al (2010) Identifying signatures of natural selection in Tibetan and Andean populations using dense genome scan data. PLoS Genet 6:e1001116. https://doi.org/10.1371/journal.pgen.1001116

Bigham AW et al (2013) Andean and Tibetan patterns of adaptation to high altitude. Am J Hum Biol 25:190–197. https://doi.org/10.1002/ajhb.22358

Bigham AW et al (2014) Maternal PRKAA1 and EDNRA genotypes are associated with birth weight, and PRKAA1 with uterine artery diameter and metabolic homeostasis at high altitude. Physiol Genomics 46:687–697. https://doi.org/10.1152/physiolgenomics.00063.2014

Bishop T et al (2013) Carotid body hyperplasia and enhanced ventilatory responses to hypoxia in mice with heterozygous deficiency of PHD2. J Physiol 591:3565–3577. https://doi.org/10.1113/jphysiol.2012.247254

Brutsaert TD, Soria R, Caceres E, Spielvogel H, Haas JD (1999) Effect of developmental and ancestral high altitude exposure on chest morphology and pulmonary function in Andean and European/North American natives. Am J Hum Biol 11:383–395. https://doi.org/10.1002/(SICI)1520-6300(1999)11:3<383::AID-AJHB9>3.0.CO;2-X

Brutsaert TD et al (2003) Spanish genetic admixture is associated with larger V(O2) max decrement from sea level to 4338 m in Peruvian Quechua. J Appl Physiol 95:519–528

Brutsaert TD et al (2005) Ancestry explains the blunted ventilatory response to sustained hypoxia and lower exercise ventilation of Quechua altitude natives. Am J Physiol Regul Integr Comp Physiol 289:R225–R234. https://doi.org/10.1152/ajpregu.00105.2005

Brutsaert TD et al (2019) Association of EGLN1 gene with high aerobic capacity of Peruvian Quechua at high altitude. Proc Natl Acad Sci 116:24006. https://doi.org/10.1073/pnas.1906171116

Cavalli-Sforza L, Menozzi P, Piazza A (1995) Genetics and the origin of human “races”. Russ J Genet 54:853–867

Center for International Earth Science Information Network, C. C. U (2012) NASA socioeconomic data and applications center. SEDAC, Palisades

Chen FH et al (2015) Agriculture facilitated permanent human occupation of the Tibetan Plateau after 3600 BP. Science 347:248–250. https://doi.org/10.1126/science.1259172

Chen F et al (2019) A late middle pleistocene denisovan mandible from the Tibetan plateau. Nature 569:409–412. https://doi.org/10.1038/s41586-019-1139-x

Chen L, Wolf AB, Fu W, Li L, Akey JM (2020) Identifying and interpreting apparent Neanderthal ancestry in African individuals. Cell 180:677–687. https://doi.org/10.1016/j.cell.2020.01.012

Cheviron ZA, Bachman GC, Connaty AD, McClelland GB, Storz JF (2012) Regulatory changes contribute to the adaptive enhancement of thermogenic capacity in high-altitude deer mice. Proc Natl Acad Sci U S A 109:8635–8640. https://doi.org/10.1073/pnas.1120523109

Chiang C et al (2017) The impact of structural variation on human gene expression. Nat Genet 49:692–699. https://doi.org/10.1038/ng.3834

Childebayeva A et al (2019) LINE-1 and EPAS1 DNA methylation associations with high-altitude exposure. Epigenetics 14:1–15. https://doi.org/10.1080/15592294.2018.1561117

Crocker ME et al (2020) Effects of high altitude on respiratory rate and oxygen saturation reference values in healthy infants and children younger than 2 years in four countries: a cross-sectional study. Lancet Glob Health 8:e362–e373. https://doi.org/10.1016/S2214-109X(19)30543-1

Dannemann M, Kelso J (2017) The contribution of Neanderthals to phenotypic variation in modern humans. Am J Hum Genet 101:578–589. https://doi.org/10.1016/j.ajhg.2017.09.010

Deng L et al (2019) Prioritizing natural-selection signals from the deep-sequencing genomic data suggests multi-variant adaptation in Tibetan highlanders. Natl Sci Rev 6:1201–1222. https://doi.org/10.1093/nsr/nwz108

Dillehay TD, Collins MB (1988) Early cultural evidence from Monte-Verde in Chile. Nature 332:150–152. https://doi.org/10.1038/332150a0

Ding Q, Hu Y, Xu S, Wang J, Jin L (2014) Neanderthal introgression at chromosome 3p21.31 was under positive natural selection in East Asians. Mol Biol Evol 31:683–695. https://doi.org/10.1093/molbev/mst260

Durvasula A, Sankararaman S (2020) Recovering signals of ghost archaic introgression in African populations. Sci Adv 6:eaax5097. https://doi.org/10.1126/sciadv.aax5097

Egger G, Liang G, Aparicio A, Jones PA (2004) Epigenetics in human disease and prospects for epigenetic therapy. Nature 429:457–463. https://doi.org/10.1038/nature02625

Eichstaedt CA et al (2014) The Andean adaptive toolkit to counteract high altitude maladaptation: genome-wide and phenotypic analysis of the Collas. PLoS One 9:e93314. https://doi.org/10.1371/journal.pone.0093314

Eichstaedt CA et al (2017) Evidence of early-stage selection on EPAS1 and GPR126 genes in Andean high altitude populations. Sci Rep 7:13042. https://doi.org/10.1038/s41598-017-13382-4

Elks CE et al (2012) Variability in the heritability of body mass index: a systematic review and meta-regression. Front Endocrinol 3:29. https://doi.org/10.3389/fendo.2012.00029

Enard D, Petrov D (2018) Evidence that RNA viruses drove adaptive introgression between Neanderthals and modern humans. Cell 175:360–371. https://doi.org/10.1016/j.cell.2018.08.034

Erzurum SC et al (2007) Higher blood flow and circulating NO products offset high-altitude hypoxia among Tibetans. Proc Natl Acad Sci 104:17593. https://doi.org/10.1073/pnas.0707462104

Fan S, Hansen ME, Lo Y, Tishkoff SA (2016) Going global by adapting local: A review of recent human adaptation. Science 354:54–59. https://doi.org/10.1126/science.aaf5098

Fehren-Schmitz L et al (2011) Diachronic investigations of mitochondrial and Y-chromosomal genetic markers in pre-Columbian Andean highlanders from South Peru. Ann Hum Genet 75:266–283. https://doi.org/10.1111/j.1469-1809.2010.00620.x

Ferezou J et al (1993) Reduction of postprandial lipemia after acute exposure to high altitude hypoxia. Int J Sports Med 14:78–85. https://doi.org/10.1055/s-2007-1021150

Foll M, Gaggiotti OE, Daub JT, Vatsiou A, Excoffier L (2014) Widespread signals of convergent adaptation to high altitude in Asia and america. Am J Hum Genet 95:394–407. https://doi.org/10.1016/j.ajhg.2014.09.002

Forsman A (2015) Rethinking phenotypic plasticity and its consequences for individuals, populations and species. Heredity 115:276–284. https://doi.org/10.1038/hdy.2014.92

Furrow RE (2014) Epigenetic inheritance, epimutation, and the response to selection. PLoS One 9:e101559. https://doi.org/10.1371/journal.pone.0101559

Fuselli S et al (2003) Mitochondrial DNA diversity in south America and the genetic history of Andean highlanders. Mol Biol Evol 20:1682–1691. https://doi.org/10.1093/molbev/msg188

Gaur P et al (2020) Comparative analysis of high altitude hypoxia induced erythropoiesis and iron homeostasis in Indian and Kyrgyz lowlander males. Curr Res Biotechnol 2:120–130. https://doi.org/10.1016/j.crbiot.2020.10.001

Ge RL et al (2012) Metabolic insight into mechanisms of high-altitude adaptation in Tibetans. Mol Genet Metab 106:244–247

Ghosh S et al (2019) Exhaled nitric oxide in ethnically diverse high-altitude native populations: a comparative study. Am J Phys Anthropol 170:451–458. https://doi.org/10.1002/ajpa.23915

Gou X et al (2014) Whole-genome sequencing of six dog breeds from continuous altitudes reveals adaptation to high-altitude hypoxia. Genome Res 24(8):1308–1315

Gouy A, Excoffier L (2020) Polygenic patterns of adaptive introgression in modern humans are mainly shaped by response to pathogens. Mol Biol Evol 37:1420–1433. https://doi.org/10.1093/molbev/msz306

Graham AM, McCracken KG (2019) Convergent evolution on the hypoxia-inducible factor (HIF) pathway genes EGLN1 and EPAS1 in high-altitude ducks. Heredity 122:819–832. https://doi.org/10.1038/s41437-018-0173-z

Green RE et al (2010) A draft sequence of the Neandertal genome. Science 328:710–722. https://doi.org/10.1126/science.1188021

Grossman SR et al (2010) A composite of multiple signals distinguishes causal variants in regions of positive selection. Science 327:883–886. https://doi.org/10.1126/science.1183863

Groves BM et al (1993) Minimal hypoxic pulmonary hypertension in normal Tibetans at 3,658 m. J Appl Physiol 74:312–318. https://doi.org/10.1152/jappl.1993.74.1.312

Gurung PD, Upadhyay AK, Bhardwaj PK, Sowdhamini R, Ramakrishnan U (2019) Transcriptome analysis reveals plasticity in gene regulation due to environmental cues in Primula sikkimensis, a high altitude plant species. BMC Genomics 20:989. https://doi.org/10.1186/s12864-019-6354-1

Hall JE, Lawrence ES, Simonson TS, Fox K (2020) Seqing higher ground: functional investigation of adaptive variation associated with high-altitude adaptation. Front Genet 11:471. https://doi.org/10.3389/fgene.2020.00471

Hammer MF, Woerner AE, Mendez FL, Watkins JC, Wall JD (2011) Genetic evidence for archaic admixture in Africa. Proc Natl Acad Sci U S A 108:15123–15128. https://doi.org/10.1073/pnas.1109300108

Hao Y et al (2019) Comparative transcriptomics of 3 high-altitude passerine birds and their low-altitude relatives. Proc Natl Acad Sci 116:11851. https://doi.org/10.1073/pnas.1819657116

Hassen M (1990) The oromo of Ethiopia: a history, 1570–1860. Cambridge University Press, Cambridge

Hawks J, Wang ET, Cochran GM, Harpending HC, Moyzis RK (2007) Recent acceleration of human adaptive evolution. Proc Natl Acad Sci U S A 104:20753–20758. https://doi.org/10.1073/pnas.0707650104

Haygood R, Babbitt CC, Fedrigo O, Wray GA (2010) Contrasts between adaptive coding and noncoding changes during human evolution. Proc Natl Acad Sci 107:7853. https://doi.org/10.1073/pnas.0911249107

He Y et al (2018) Blunted nitric oxide regulation in Tibetans under high-altitude hypoxia. Natl Sci Rev 5:516–529. https://doi.org/10.1093/nsr/nwy037

Heinrich EC et al (2019) Genetic variants at the EGLN1 locus associated with high-altitude adaptation in Tibetans are absent or found at low frequency in highland Andeans. Ann Hum Genet 83:171–176. https://doi.org/10.1111/ahg.12299

Ho JJ, Man HS, Marsden PA (2012) Nitric oxide signaling in hypoxia. J Mol Med 90:217–231. https://doi.org/10.1007/s00109-012-0880-5

Hochachka PW, Rupert JL (2003) Fine tuning the HIF-1 'global' O2 sensor for hypobaric hypoxia in Andean high-altitude natives. Bioessays 25:515–519. https://doi.org/10.1002/bies.10261

Horikoshi M et al (2016) Genome-wide associations for birth weight and correlations with adult disease. Nature 538:248–252. https://doi.org/10.1038/nature19806

Horscroft JA et al (2017) Metabolic basis to Sherpa altitude adaptation. Proc Natl Acad Sci U S A 114:6382–6387. https://doi.org/10.1073/pnas.1700527114

Hu H et al (2017) Evolutionary history of Tibetans inferred from whole-genome sequencing. PLoS Genet 13:e1006675. https://doi.org/10.1371/journal.pgen.1006675

Huang SY, Ning XH, Zhou ZN, Gu ZZ, Hu ST (1984) Ventilatory function in adaptation to high altitude: studies in Tibet. Springer, New York

Hubisz MJ, Williams AL, Siepel A (2020) Mapping gene flow between ancient hominins through demography-aware inference of the ancestral recombination graph. PLoS Genet 16:e1008895. https://doi.org/10.1371/journal.pgen.1008895

Huerta-Sanchez E et al (2013) Genetic signatures reveal high-altitude adaptation in a set of ethiopian populations. Mol Biol Evol 30:1877–1888. https://doi.org/10.1093/molbev/mst089

Huerta-Sanchez E et al (2014) Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature 512:194–197. https://doi.org/10.1038/nature13408

Huerta-Sánchez E et al (2014) Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature 512:194–197. https://doi.org/10.1038/nature13408

Hurles ME, Dermitzakis ET, Tyler-Smith C (2008) The functional impact of structural variation in humans. Trends Genet 24:238–245. https://doi.org/10.1016/j.tig.2008.03.001

Irizarry RA et al (2008) Comprehensive high-throughput arrays for relative methylation (CHARM). Genome Res 18:780–790. https://doi.org/10.1101/gr.7301508

Ivan M et al (2001) HIFalpha targeted for VHL-mediated destruction by proline hydroxylation: implications for O2 sensing. Science 292:464–468. https://doi.org/10.1126/science.1059817

Jaakkola P et al (2001) Targeting of HIF-alpha to the von Hippel-Lindau ubiquitylation complex by O2-regulated prolyl hydroxylation. Science 292:468–472. https://doi.org/10.1126/science.1059796

Jagoda E et al (2018) Disentangling immediate adaptive introgression from selection on standing introgressed variation in humans. Mol Biol Evol 35:623–630. https://doi.org/10.1093/molbev/msx314

Jeong C et al (2014) Admixture facilitates genetic adaptations to high altitude in Tibet. Nat Commun 5:3281. https://doi.org/10.1038/ncomms4281

Julian C (2017) Epigenomics and human adaptation to high altitude. J Appl Physiol 123:1362–1370. https://doi.org/10.1152/japplphysiol.00351.2017

Julian CG, Moore LG (2019) Human genetic adaptation to high altitude: evidence from the Andes. Gene 10:10020150. https://doi.org/10.3390/genes10020150

Julian CG et al (2009) Augmented uterine artery blood flow and oxygen delivery protect Andeans from altitude-associated reductions in fetal growth. Am J Physiol Regul Integr Comp Physiol 296:R1564–R1575. https://doi.org/10.1152/ajpregu.90945.2008

Keith B, Johnson RS, Simon MC (2011) HIF1alpha and HIF2alpha: sibling rivalry in hypoxic tumour growth and progression. Nat Rev Cancer 12:9–22. https://doi.org/10.1038/nrc3183

Kelly SA, Panhuis TM, Stoehr AM (2012) Phenotypic plasticity: molecular mechanisms and adaptive significance. Compr Physiol 2:1417–1439. https://doi.org/10.1002/cphy.c110008

Khrameeva EE et al (2014) Neanderthal ancestry drives evolution of lipid catabolism in contemporary Europeans. Nat Commun 5:3584. https://doi.org/10.1038/ncomms4584

Kronholm I, Collins S (2016) Epigenetic mutations can both help and hinder adaptive evolution. Mol Ecol 25:1856–1868. https://doi.org/10.1111/mec.13296

Lachance G et al (2014) DNMT3a epigenetic program regulates the HIF-2alpha oxygen-sensing pathway and the cellular response to hypoxia. Proc Natl Acad Sci U S A 111:7783–7788. https://doi.org/10.1073/pnas.1322909111

Lawrence ES, Heinrich EC, Wu L, Villafuerte FC, Simonson TS (2018) Genetic missense variants at the EGLN1 Locus associated with high-altitude adaptation in Tibetans are rare in Andeans. FASEB J 32:lb478. https://doi.org/10.1096/fasebj.2018.32.1_supplement.lb478

Lee FS, Percy MJ (2011) The HIF pathway and erythrocytosis. Annu Rev Pathol 6:165–192. https://doi.org/10.1146/annurev-pathol-011110-130321

Leon-Velarde F, Mejia O (2008) Gene expression in chronic high altitude diseases. High Alt Med Biol 9:130–139. https://doi.org/10.1089/ham.2007.1077

Leon-Velarde F et al (2005) Consensus statement on chronic and subacute high altitude diseases. High Alt Med Biol 6:147–157. https://doi.org/10.1089/ham.2005.6.147

Levett DZ et al The role of nitrogen oxides in human adaptation to hypoxia. Sci Rep 1(109):109

Lewis H (1966) The origins of the Galla and Somali. J Afr Hist 7:27–46. https://doi.org/10.1017/S0021853700006058

Li Y et al (2014) Population variation revealed high-altitude adaptation of Tibetan Mastiffs. Mol Biol Evol 31:1200–1205

Li S, Li D, Zhao X, Wang Y, Zhu Q (2017) A non-synonymous SNP with the allele frequency correlated with the altitude may contribute to the hypoxia adaptation of Tibetan chicken. PLoS One 12:e0172211

Li JT et al (2018) Comparative genomic investigation of high-elevation adaptation in ectothermic snakes. Proc Natl Acad Sci U S A 26:21

Li YC et al (2019) Neolithic millet farmers contributed to the permanent settlement of the Tibetan Plateau by adopting barley agriculture. Natl Sci Rev 6:1005–1013. https://doi.org/10.1093/nsr/nwz080

Lindo J et al (2018) The genetic prehistory of the Andean highlands 7000 years BP though European contact. Sci Adv 4:eaau4921. https://doi.org/10.1126/sciadv.aau4921

Lisy K, Peet DJ (2008) Turn me on: regulating HIF transcriptional activity. Cell Death Differ 15:642–649. https://doi.org/10.1038/sj.cdd.4402315

Liu KX, Sun XC, Wang SW, Hu B (2007) Association of polymorphisms of 1772 (C-->T) and 1790 (G-->A) in HIF1A gene with hypoxia adaptation in high altitude in Sherpas. Zhonghua Yi Xue Yi Chuan Xue Za Zhi 24:230–232

Liu X et al (2019) EPAS1 gain-of-function mutation contributes to high-altitude adaptation in Tibetan Horses. Mol Biol Evol 36:2591–2603. https://doi.org/10.1093/molbev/msz158

Liu Y et al (2020) De novo transcriptomic and metabolomic analyses reveal the ecological adaptation of high-altitude Bombus pyrosoma. Insects 11:090631. https://doi.org/10.3390/insects11090631

Llamas B et al (2016) Ancient mitochondrial DNA provides high-resolution time scale of the peopling of the Americas. Sci Adv 2:e1501385. https://doi.org/10.1126/sciadv.1501385

Locke AE et al (2015) Genetic studies of body mass index yield new insights for obesity biology. Nature 518:197–206

Loenarz C et al (2011) The hypoxia-inducible transcription factor pathway regulates oxygen sensing in the simplest animal, Trichoplax adhaerens. EMBO Rep 12:63–70. https://doi.org/10.1038/embor.2010.170

Long K et al (2019) Small non-coding RNA transcriptome of four high-altitude vertebrates and their low-altitude relatives. Sci Data 6:192. https://doi.org/10.1038/s41597-019-0204-5

Lorenzo FR et al (2014) A genetic mechanism for Tibetan high-altitude adaptation. Nat Genet 46:951–956. https://doi.org/10.1038/ng.3067

Lotterhos KE, Whitlock MC (2015) The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. Mol Ecol 24:1031–1046. https://doi.org/10.1111/mec.13100

Lou H et al (2015) A 3.4-kb copy-number deletion near EPAS1 is significantly enriched in high-altitude Tibetans but absent from the Denisovan sequence. Am J Hum Genet 97:54–66. https://doi.org/10.1016/j.ajhg.2015.05.005

Lu D et al (2016) Ancestral origins and genetic history of Tibetan highlanders. Am J Hum Genet 99:580–594. https://doi.org/10.1016/j.ajhg.2016.07.002

Lynch TF, Kennedy KAR (1970) Early human cultural and skeletal remains from Guitarrero Cave, Northern Peru. Science 169:1307. https://doi.org/10.1126/science.169.3952.1307

MacInnis MJ, Rupert JL (2011) Ome on the Range: altitude adaptation, positive selection, and Himalayan genomics. High Alt Med Biol 12:133–139. https://doi.org/10.1089/ham.2010.1090

Martin D, Windsor J (2008) From mountain to bedside: understanding the clinical relevance of human acclimatisation to high-altitude hypoxia. Postgrad Med J 84:622–627. https://doi.org/10.1136/pgmj.2008.068296

Mendez FL, Watkins JC, Hammer MF (2012) A haplotype at STAT2 Introgressed from neanderthals and serves as a candidate of positive selection in Papua New Guinea. Am J Hum Genet 91:265–274. https://doi.org/10.1016/j.ajhg.2012.06.015

Meyer M et al (2012) A high-coverage genome sequence from an archaic denisovan individual. Science 338:222–226. https://doi.org/10.1126/science.1224344

Meyer MC et al (2017) Permanent human occupation of the central Tibetan Plateau in the early Holocene. Science 355:64–67. https://doi.org/10.1126/science.aag0357

Miao B, Wang Z, Li Y (2017) Genomic analysis reveals hypoxia adaptation in the Tibetan Mastiff by introgression of the gray wolf from the Tibetan Plateau. Mol Biol Evol 34:734–743. https://doi.org/10.1093/molbev/msw274

Mondal M et al (2016) Genomic analysis of Andamanese provides insights into ancient human migration into Asia and adaptation. Nat Genet 48:1066–1070. https://doi.org/10.1038/ng.3621

Moore LG (2017) Human genetic adaptation to high altitudes: Current status and future prospects. Quat Int 461:4–13. https://doi.org/10.1016/j.quaint.2016.09.045

Moore LG et al (2004) Maternal adaptation to high-altitude pregnancy: an experiment of nature--a review. Placenta 25(Suppl A):60–71. https://doi.org/10.1016/j.placenta.2004.01.008

Moreno-Mayar JV et al (2018) Early human dispersals within the Americas. Science 362:eaav2621. https://doi.org/10.1126/science.aav2621

Munroe DJ, Harris TJR (2010) Third-generation sequencing fireworks at Marco Island. Nat Biotechnol 28:426–428. https://doi.org/10.1038/nbt0510-426

Naeije R (2010) Physiological adaptation of the cardiovascular system to high altitude. Prog Cardiovasc Dis 52:456–466. https://doi.org/10.1016/j.pcad.2010.03.004

Nakatsuka N et al (2020) A paleogenomic reconstruction of the deep population history of the Andes. Cell 181:1131–1145. https://doi.org/10.1016/j.cell.2020.04.015

Niermeyer S, Andrade Mollinedo P, Huicho L (2009) Child health and living at high altitude. Arch Dis Child 94:806–811. https://doi.org/10.1136/adc.2008.141838

Ou LC, Leiter JC (2004) Effects of exposure to a simulated altitude of 5500 m on energy metabolic pathways in rats. Respir Physiol Neurobiol 141:59–71. https://doi.org/10.1016/j.resp.2004.04.001

Ouzhuluobu et al (2020) De novo assembly of a Tibetan genome and identification of novel structural variants associated with high-altitude adaptation. Natl Sci Rev 7:391–402. https://doi.org/10.1093/nsr/nwz160

Pagani L et al (2012a) Ethiopian genetic diversity reveals linguistic stratification and complex influences on the Ethiopian gene pool. Am J Hum Genet 91:83–96. https://doi.org/10.1016/j.ajhg.2012.05.015

Pagani L et al (2012b) High altitude adaptation in Daghestani populations from the Caucasus. Hum Genet 131:423–433. https://doi.org/10.1007/s00439-011-1084-8

Pamenter ME, Hall JE, Tanabe Y, Simonson TS (2020) Cross-species insights into genomic adaptations to hypoxia. Front Genet 11:00743. https://doi.org/10.3389/fgene.2020.00743

Pan H et al (2018) Mutations in EPAS1 in congenital heart disease in Tibetans. Biosci Rep 38:1389. https://doi.org/10.1042/bsr20181389

Peng Y et al (2011) Genetic variations in Tibetan populations and high-altitude adaptation at the Himalayas. Mol Biol Evol 28:1075–1081. https://doi.org/10.1093/molbev/msq290

Peng Y et al (2017) Down-regulation of EPAS1 transcription and genetic adaptation of Tibetans to high-altitude hypoxia. Mol Biol Evol 34:818–830. https://doi.org/10.1093/molbev/msw280

Petr M, Paabo S, Kelso J, Vernot B (2019) Limits of long-term selection against Neandertal introgression. Proc Natl Acad Sci U S A 116:1639–1644. https://doi.org/10.1073/pnas.1814338116

Piperno A et al (2011) Modulation of hepcidin production during hypoxia-induced erythropoiesis in humans in vivo: data from the HIGHCARE project. Blood 117:2953–2959. https://doi.org/10.1182/blood-2010-08-299859

Pleurdeau D (2005) Human technical behavior in the African middle stone age: the lithic assemblage of Porc-Epic Cave (Dire Dawa, Ethiopia). Afr Archaeol Rev 22:177–197. https://doi.org/10.1007/s10437-006-9000-7

Posth C et al (2018) Reconstructing the deep population history of Central and South America. Cell 175:1185–1197. https://doi.org/10.1016/j.cell.2018.10.027

Prabhakar NR (2020) 2019 nobel prize in physiology or medicine. Physiology 35:81–83. https://doi.org/10.1152/physiol.00001.2020

Prior SJ et al (2003) Sequence variation in hypoxia-inducible factor 1alpha (HIF1A): association with maximal oxygen consumption. Physiol Genomics 15:20–26. https://doi.org/10.1152/physiolgenomics.00061.2003

Prufer K et al (2014) The complete genome sequence of a Neanderthal from the Altai mountains. Nature 505:43–49. https://doi.org/10.1038/nature12886

Prufer K et al (2017) A high-coverage Neandertal genome from Vindija Cave in Croatia. Science 358:655–658. https://doi.org/10.1126/science.aao1887

Qi X et al (2013) Genetic evidence of paleolithic colonization and neolithic expansion of modern humans on the tibetan plateau. Mol Biol Evol 30:1761–1778. https://doi.org/10.1093/molbev/mst093

Qi X et al (2019) The transcriptomic landscape of Yaks reveals molecular pathways for high altitude adaptation. Genome Biol Evol 11:72–85. https://doi.org/10.1093/gbe/evy264

Qin ZD et al (2010) A mitochondrial revelation of early human migrations to the Tibetan Plateau before and after the last glacial maximum. Am J Phys Anthropol 143:555–569. https://doi.org/10.1002/ajpa.21350

Quan, C. et al. (2020). Characterization of structural variation in Tibetans reveals new evidence of high-altitude adaptation and introgression. bioRxiv, 2020.2012.2001.401174. https://doi.org/10.1101/2020.12.01.401174

Rademaker K et al (2014) Paleoindian settlement of the high-altitude Peruvian Andes. Science 346:466–469. https://doi.org/10.1126/science.1258260

Rawluszko-Wieczorek AA, Horbacka K, Krokowicz P, Misztal M, Jagodzinski PP (2014) Prognostic potential of DNA methylation and transcript levels of HIF1A and EPAS1 in colorectal cancer. Mol Cancer Res 12:1112–1127. https://doi.org/10.1158/1541-7786.MCR-14-0054

Richards EJ (2006) Inherited epigenetic variation--revisiting soft inheritance. Nat Rev Genet 7:395–401. https://doi.org/10.1038/nrg1834

Rothhammer F, Santoro CM (2001) El Desarrollo Cultural En El Valle De Azapa, Extremo Norte De Chile Y Su Vinculación Con Los Desplazamientos Poblacionales Altiplánicos. Lat Am Antiq 12:59–66. https://doi.org/10.2307/971757

Rothhammer F, Silva C (1989) Peopling of Andean South America. Am J Phys Anthropol 78:403–410. https://doi.org/10.1002/ajpa.1330780308

Rupert JL, Hochachka PW (2001) Genetic approaches to understanding human adaptation to altitude in the Andes. J Exp Biol 204:3151–3160

Sabeti PC et al (2006) Positive natural selection in the human lineage. Science 312:1614–1620. https://doi.org/10.1126/science.1124309

Sabeti PC et al (2007) Genome-wide detection and characterization of positive selection in human populations. Nature 449:913–918. https://doi.org/10.1038/nature06250

Sackton TB, Clark N (2019) Convergent evolution in the genomics era: new insights and directions. Philos Trans R Soc B 374:20190102. https://doi.org/10.1098/rstb.2019.0102

Sandoval JR et al (2013) The genetic history of indigenous populations of the Peruvian and Bolivian Altiplano: the legacy of the Uros. PLoS One 8:e73006. https://doi.org/10.1371/journal.pone.0073006

Sandweiss DH et al (1998) Quebrada jaguay: early south American maritime adaptations. Science 281:1830–1832. https://doi.org/10.1126/science.281.5384.1830

Sankararaman S et al (2014) The genomic landscape of Neanderthal ancestry in present-day humans. Nature 507:354–357. https://doi.org/10.1038/nature12961

Sargent C et al (2013) The impact of altitude on the sleep of young elite soccer players (ISA3600). Br J Sports Med 47(Suppl 1):i86–i92. https://doi.org/10.1136/bjsports-2013-092829

Scheinfeldt LB, Tishkoff SA (2010) Living the high life: high-altitude adaptation. Genome Biol 11:133. https://doi.org/10.1186/gb-2010-11-9-133

Scheinfeldt LB et al (2012) Genetic adaptation to high altitude in the Ethiopian highlands. Genome Biol 13:R1. https://doi.org/10.1186/gb-2012-13-1-r1

Schneider MP, Boesen EI, Pollock DM (2007) Contrasting actions of endothelin ET(A) and ET(B) receptors in cardiovascular disease. Annu Rev Pharmacol Toxicol 47:731–759. https://doi.org/10.1146/annurev.pharmtox.47.120505.105134

Schofield CJ, Ratcliffe PJ (2004) Oxygen sensing by HIF hydroxylases. Nat Rev Mol Cell Biol 5:343–354. https://doi.org/10.1038/nrm1366

Schumacker PT et al (2014) High altitude: human adaptation to hypoxia. Springer, New York

Semenza GL (1999) Regulation of mammalian O2 homeostasis by hypoxia-inducible factor 1. Annu Rev Cell Dev Biol 15:551–578. https://doi.org/10.1146/annurev.cellbio.15.1.551

Semenza GL (2004) Hydroxylation of HIF-1: oxygen sensing at the molecular level. Physiology 19:176–182. https://doi.org/10.1152/physiol.00001.2004

Seo JS et al (2016) De novo assembly and phasing of a Korean human genome. Nature 538:243–247. https://doi.org/10.1038/nature20098

Shao Y et al (2015) Genetic adaptations of the plateau zokor in high-elevation burrows. Sci Rep 5:17262. https://doi.org/10.1038/srep17262

Shi L et al (2016) Long-read sequencing and de novo assembly of a Chinese genome. Nat Commun 7:12065. https://doi.org/10.1038/ncomms12065

Silventoinen K et al (2003) Heritability of adult body height: a comparative study of twin cohorts in eight countries. Twin Res 6:399–408. https://doi.org/10.1375/136905203770326402

Simonson TS et al (2010) Genetic evidence for high-altitude adaptation in Tibet. Science 329:72. https://doi.org/10.1126/science.1189406

Song S et al (2016) Exome sequencing reveals genetic differentiation due to high-altitude adaptation in the Tibetan cashmere goat (Capra hircus). BMC Genomics 17:122. https://doi.org/10.1186/s12864-016-2449-0.

Steinmann K, Richter AM, Dammann RH (2011) Epigenetic silencing of erythropoietin in human cancers. Genes Cancer 2:65–73. https://doi.org/10.1177/1947601911405043

Stobdan T et al (2015) Endothelin receptor B, a candidate gene from human studies at high altitude, improves cardiac tolerance to hypoxia in genetically engineered heterozygote mice. Proc Natl Acad Sci U S A 112:10425–10430. https://doi.org/10.1073/pnas.1507486112

Storz JF, Scott GR, Cheviron ZA (2010) Phenotypic plasticity and genetic adaptation to high-altitude hypoxia in vertebrates. J Exp Biol 213:4125–4136. https://doi.org/10.1242/jeb.048181

Suzuki K et al (2003) Genetic variation in hypoxia-inducible factor 1alpha and its possible association with high altitude adaptation in Sherpas. Med Hypotheses 61:385–389. https://doi.org/10.1016/s0306-9877(03)00178-6

Szpak M et al (2018) FineMAV: prioritizing candidate genetic variants driving local adaptations in human populations. Genome Biol 19:5. https://doi.org/10.1186/s13059-017-1380-2

Tarazona-Santos E et al (2001) Genetic differentiation in South Amerindians is related to environmental and cultural diversity: evidence from the Y chromosome. Am J Hum Genet 68:1485–1496. https://doi.org/10.1086/320601

Tian H, McKnight SL, Russell DW (1997) Endothelial PAS domain protein 1 (EPAS1), a transcription factor selectively expressed in endothelial cells. Genes Dev 11:72–82. https://doi.org/10.1101/gad.11.1.72

Tian H, Hammer RE, Matsumoto AM, Russell DW, McKnight SL (1998) The hypoxia-responsive transcription factor EPAS1 is essential for catecholamine homeostasis and protection against heart failure during embryonic development. Genes Dev 12:3320–3324. https://doi.org/10.1101/gad.12.21.3320

Tucci S et al (2018) Evolutionary history and adaptation of a human pygmy population of Flores Island, Indonesia. Science 361:511–516. https://doi.org/10.1126/science.aar8486

Udpa N et al (2014) Whole genome sequencing of Ethiopian highlanders reveals conserved hypoxia tolerance genes. Genome Biol 15:R36. https://doi.org/10.1186/gb-2014-15-2-r36

Velotta JP, Ivy CM, Wolf CJ, Scott GR, Cheviron ZA (2018) Maladaptive phenotypic plasticity in cardiac muscle growth is suppressed in high-altitude deer mice. Evolution 72:2712–2727. https://doi.org/10.1111/evo.13626

Vernot B, Akey JM (2014) Resurrecting surviving Neandertal lineages from modern human genomes. Science 343:1017–1021. https://doi.org/10.1126/science.1245938

Virues-Ortega J, Garrido E, Javierre C, Kloezeman KC (2006) Human behaviour and development under high-altitude conditions. Dev Sci 9:400–410. https://doi.org/10.1111/j.1467-7687.2006.00505.x

Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS Biol 4:e72. https://doi.org/10.1371/journal.pbio.0040072

Vuckovic D et al (2020) The polygenic and monogenic basis of blood traits and diseases. Cell 182:1214–1231. https://doi.org/10.1016/j.cell.2020.08.008

Wang B et al (2011) On the origin of Tibetans and their genetic basis in adapting high-altitude environments. PLoS One 6:e17002. https://doi.org/10.1371/journal.pone.0017002

Wang G-D et al (2014) Genetic convergence in the adaptation of dogs and humans to the high-altitude environment of the tibetan plateau. Genome Biol Evol 6:2122–2128. https://doi.org/10.1093/gbe/evu162

Weir BS, Hill WG (2002) Estimating F-statistics. Annu Rev Genet 36:721–750. https://doi.org/10.1146/annurev.genet.36.050802.093940

Weitz CA, Garruto RM, Chin CT (2016) Larger FVC and FEV1 among Tibetans compared to Han born and raised at high altitude. Am J Phys Anthropol 159:244–255. https://doi.org/10.1002/ajpa.22873

Wenger RH, Stiehl DP, Camenisch G (2005) Integration of oxygen signaling at the consensus HRE. Sci STKE 2005:re12. https://doi.org/10.1126/stke.3062005re12

West JB (1982) Respiratory and circulatory control at high altitudes. J Exp Biol 100:147

Witt KE, Huerta-Sanchez E (2019) Convergent evolution in human and domesticate adaptation to high-altitude environments. Philos Trans R Soc Lond Ser B Biol Sci 374:20180235. https://doi.org/10.1098/rstb.2018.0235

Wu TY et al (2005) Hemoglobin levels in Tibet: different effect of age and gender for Tibetans vs Han. Comp Clin Pathol 14:25–35. https://doi.org/10.1007/s00580-005-0550-x

Wu DD et al (2018) Pervasive introgression facilitated domestication and adaptation in the Bos species complex. Nat Ecol Evol 2:1139–1145. https://doi.org/10.1038/s41559-018-0562-y

Wuren T et al (2014) Shared and unique signals of high-altitude adaptation in geographically distinct Tibetan populations. PLoS One 9:e88252. https://doi.org/10.1371/journal.pone.0088252

Xiang K 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. https://doi.org/10.1093/molbev/mst090

Xin J et al (2020) Chromatin accessibility landscape and regulatory network of high-altitude hypoxia adaptation. Nat Commun 11:4928. https://doi.org/10.1038/s41467-020-18638-8

Xu S et al (2010) A genome-wide search for signals of high-altitude adaptation in Tibetans. Mol Biol Evol 28:1003–1011

Xu X-H et al (2014) Two functional loci in the promoter of EPAS1 gene involved in high-altitude adaptation of Tibetans. Sci Rep 4:7465–7465. https://doi.org/10.1038/srep07465

Xu J et al (2015) Association between genetic polymorphisms of EDNRA gene and high altitude polycythemia in Tibetans at the Qinghai-Tibetan Plateau. Zhonghua Yi Xue Za Zhi 95:1382–1385

Yang D et al (2016) HMOX2 functions as a modifier gene for high-altitude adaptation in Tibetans. Hum Mutat 37:216–223. https://doi.org/10.1002/humu.22935

Yang J et al (2017) Genetic signatures of high-altitude adaptation in Tibetans. Proc Natl Acad Sci U S A 114:4189–4194. https://doi.org/10.1073/pnas.1617042114

Yasukochi Y, Nishimura T, Motoi M, Watanuki S (2018) Association of EGLN1 genetic polymorphisms with SpO2 responses to acute hypobaric hypoxia in a Japanese cohort. J Physiol Anthropol 37:9. https://doi.org/10.1186/s40101-018-0169-7

Yi X et al (2010) Sequencing of 50 human exomes reveals adaptation to high altitude. Science 329:75–78. https://doi.org/10.1126/science.1190371

Yip G, Heinrich EC, Villafuerte FC, Simonson TS (2018) Heme-oxygenase 2 (HMOX2) variants associated with evolutionary adaptation and hemoglobin concentration in Tibetans are common in Andean Highlanders. FASEB J 32:lb413

Yu G, Wang L-G, Han Y, He Q-Y (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 16:284–287. https://doi.org/10.1089/omi.2011.0118

Yuan Y, Chung CY, Chan TF (2020) Advances in optical mapping for genomic research. Comput Struct Biotechnol J 18:2051–2062. https://doi.org/10.1016/j.csbj.2020.07.018

Yun Sung Cho LH, Hou H, Lee H, Jiaohui X, Kwon S, Sukhun O, Kim H-M, Jho S, Kim S, Shin Y-A, Kim BC, Kim H, Kim C-U, Luo S-J, Johnson WE, Koepfli K-P, Schmidt-Küntzel A, Turner JA, Marker L, Harper C, Miller SM, Jacobs W, Bertola LD, Kim TH, Lee S, Zhou Q, Jung H-J, Xiao X, Gadhvi P, Pengwei X, Xiong Y, Luo Y, Pan S, Gou C, Chu X, Zhang J, Liu S, He J, Chen Y, Yang L, Yang Y, He J, Liu S, Wang J, Kim CH, Kwak H, Kim J-S, Hwang S, Ko J, Kim C-B, Kim S, Bayarlkhagva D, Paek WK, Kim S-J, O'Brien SJ, Wang J, Bhak J (2013) The tiger genome and comparative analysis with lion and snow leopard genomes. Nat Commun 4:2433

Zhang W et al (2014) Hypoxia adaptations in the grey wolf (Canis lupus chanco) from Qinghai-Tibet Plateau. PLoS Genet 10:e1004466. https://doi.org/10.1371/journal.pgen.1004466.

Zhang Q et al (2016) The expression plasticity of hypoxia related genes in high altitude and plains Nanorana parkeri populations. Asian Herpetol Res 7:21–27. https://doi.org/10.16373/j.cnki.ahr.150056.

Zhang C et al (2017a) Differentiated demographic histories and local adaptations between Sherpas and Tibetans. Genome Biol 18:1242. https://doi.org/10.1186/s13059-017-1242-y

Zhang H et al (2017b) Cross-altitude analysis suggests a turning point at the elevation of 4,500 m for polycythemia prevalence in Tibetans. Am J Hematol 92:E552–E554. https://doi.org/10.1002/ajh.24809

Zhang XL et al (2018) The earliest human occupation of the high-altitude Tibetan Plateau 40 thousand to 30 thousand years ago. Science 362:1049–1051

Zhang Q, Yan Q, Yang H, Wei W (2019) Oxygen sensing and adaptability won the 2019 nobel prize in physiology or medicine. Genes Dis 6:328–332. https://doi.org/10.1016/j.gendis.2019.10.006

Zhang D et al (2020) Denisovan DNA in late pleistocene sediments from baishiya karst cave on the Tibetan Plateau. Science 370:584–587. https://doi.org/10.1126/science.abb6320

Zhou D et al (2013) Whole-genome sequencing uncovers the genetic basis of chronic mountain sickness in Andean highlanders. Am J Hum Genet 93:452–462. https://doi.org/10.1016/j.ajhg.2013.07.011

Zhuang J et al (1993) Hypoxic ventilatory responsiveness in Tibetan compared with Han residents of 3,658 m. J Appl Physiol 74:303–311. https://doi.org/10.1152/jappl.1993.74.1.303

Part II
Evolution of Modern Human Populations

Chapter 4
Mitochondrial DNA

Jun Gojobori

Abstract Mitochondria DNA (mtDNA) has distinct features from nuclear DNA. It has a circular DNA and coding its own genome. mtDNA consists of two regions, control region and coding region. Control region has start site for replication and transcription start sites. Coding regions have rRNAs and protein coding genes. Only the mtDNA in an oocyte is transmitted to the offspring, therefore it is transmitted through female lineage only. Thousands of mtDNA copies are in a cell and this makes mtDNA extraction easier. The mutation rate of mtDNA is higher than nuclear DNA. Hyper Variable Regions (HVR) in the control region have even higher mutation rates. Because of these features, mtDNA is frequently used for forensic studies or ancient DNA studies. Closely related mtDNA sequences are grouped into haplogroups. The combination of mutation including recurrent mutations determines the haplogroup. Whole mtDNA sequences are ideal for determining haplogroup. Caution is needed when haplogroups are determined based on control region.

Keywords Mitochondria · Population · Mutation rate · Haplogroups · Bayesian skyline plot · Disease · Missing heritability

4.1 Outline of This Chapter

This chapter is to introduce the mitochondria genome diversity of human. Specific features of mitochondria DNA (mtDNA) are introduced first, and the origin of current human mtDNA diversity is discussed next. The mutation rate of mtDNA is also discussed. The concept of mtDNA haplogroup is very useful and explained here with phylogeny. A method for inferring detail demographic history from mtDNA is shown with an example. The association between mtDNA mutations and disease is discussed in the last section.

J. Gojobori (✉)
The School of Advanced Sciences, SOKENDAI (The Graduate University for Advanced Studies), Hayama, Japan
e-mail: gojobori_jun@soken.ac.jp

N. Saitou (ed.), *Evolution of the Human Genome II*, Evolutionary Studies,
https://doi.org/10.1007/978-4-431-56904-6_4

4.2 Characters of mtDNA and Its Merits for Studying Human Populations

In a cell, there is typically only one nucleus while there is several hundreds or thousands more number of mitochondria. Because each mitochondrion has its own circular genomes (mtDNA), this also means that there is much more number of mtDNA compared with nucleus genomes. Then extracting mtDNA from tissue of human should be much easier than nucleus DNA. Because of this reason, mtDNA was used for earlier days of human genetics study, forensic DNA applications, and ancient DNA study. MtDNA of humans has higher mutation rate than nucleus (Pakendorf and Stoneking 2005). Although mtDNA has only around 16 kbp, it contains more mutations than nucleus DNA sequences of equivalent length because of the higher mutation rate (Table 4.1).

MtDNA is genetically haploid. When all the mtDNA in a cell, in an organ, or in an individual have the same sequence, this state is called as homoplasmy. But existence of more than mtDNA sequence type is possible and it is called as heteroplasmy. The rate of heteroplasmy is known to be higher than previously thought (Li et al. 2010; Ramos et al. 2013). Heteroplasmy can be a result of somatic mutations and does not necessarily inherit to the next generation.

MtDNA is only transmitted through the lineage of female because the only mtDNA in an oocyte is transmitted to the offspring. The mtDNA copies in oocyte are all the same and the mtDNA from sperm will be eliminated during very early development. This will assure homoplasmic status of an embryo. As a result of maternal lineage transmitting and haploid, the effective population size of mtDNA is one-fourth of nucleus. This low effective population size leads to the higher contribution of genetic drift over selection and geographical structure is easier to be formed. Therefore, mtDNA is a useful tool for tracing the lineage and characterizing human populations (Pakendorf and Stoneking 2005).

Table 4.1 The features of human mtDNA compared with nuclear DNA

	mtDNA	Nuclear DNA
Size	1.6×10^4 bp	3×10^9 bp
Form	Circular	Linear
Number of genomes in a cell	Hundreds to thousands	Two
Number of genes	37	More than 20,000
Intron	No	Yes in most gene
Inheritance	Maternal inheritance	Mendelian inheritance for autosomes
Transcription	Polycistron	Each genes
Recombination	No	Yes
Mutation rate	0.168×10^{-6}/year/site (in control region, Santos et al. 2005)	0.5×10^{-9}/year/site (Scally and Durbin 2012)

The evidences of recombination in human mtDNA were once reported (Awadalla et al. 1999). If the recombination between mtDNA in oocyte and mtDNA from sperm occurs, the assumption of single genealogy of mtDNA will be violated. Several lines show no evidence of mtDNA recombination and the apparent signal of recombination can be explained by mutational hotspots (Jorde and Bamshad 2000; Kivisild and Villems 2000; Kumar et al. 2000; Parsons and Irwin 2000; Innan and Nordborg 2002). Later, more evidence of recombination was shown in somatic tissues (Kraytsberg et al. 2004; Zsurka et al. 2005). Because we have much more powerful sequencing methods, we can collect many sample of mtDNA and may find evidence to show recombinant of human mtDNA inherits. However, indirect way to test the existence of recombination using human population data has certain limitation even if it exists (White et al. 2013).

MtDNA consists of two regions, control and coding regions (Fig. 4.1). Control region contains the starting points of replication for one strand and transcription start sites for both strands. These regions interact with nuclear-encoded genes. Displacement loop (D-loop) region is in control region but these two terms are sometimes used synonymously in the literatures. In this control region, two regions are known to have even higher mutation rate. These regions are known to be hypervariable regions (HVR). The region near to the origin is called as HVR I and the other is called as HVR II. These hypervariable regions have ten times higher mutation rate compared with coding region (Pakendorf and Stoneking 2005).

Coding regions contain many genes, 12S and 16S rRNAs, 22 tRNAs, and 13 protein coding genes (Fig. 4.1). Note that the codon table of mtDNA is different from that of nucleus genome which is the consequence of endosymbiotic origin of mitochondria. Mitochondrion produces adenosine-5′-triphosphate (ATP) through a process via the metabolic pathway called as oxidative phosphorylation (OXPHOS). OXPHOS takes place through five enzyme complexes, the mitochondrial respiratory chain. In total, more than 80 subunits comprise mitochondrial respiratory chain and 13 of them are encoded by mtDNA.

First whole mtDNA nucleotide sequences were determined in 1981 (Anderson et al. 1981), and it is called as Cambridge Reference Sequence (CRS). Later, several sequencing errors were detected for CRS and revised version of CRS was published as rCRS (Andrews et al. 1999). The position numbers of mtDNA are often referred from this rCRS and the mutation found in mtDNA often described as the difference from rCRS. Note that CRS is just one mtDNA in European lineage and is not a consensus sequence or the ancestral sequence of *Homo sapiens*. Therefore, the use of reconstructed ancestral mtDNA sequence, Reconstructed Sapiens Reference Sequence (RSRS), is recently proposed (Behar et al. 2012).

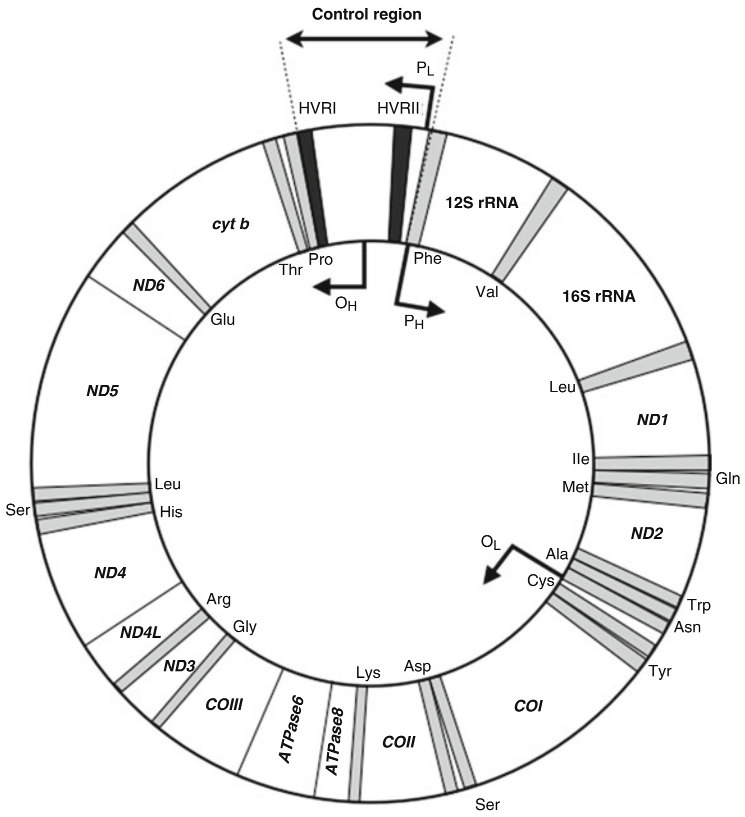


Fig. 4.1 The structure of human mtDNA. It has a length of 16,569 in case of CRS (Anderson et al. 1981). Two dark gray bars shown in control region are hypervariable regions (HVRI and HVRII). Light gray bars show the location of tRNAs together with corresponding amino acids. The regions with gene names between these bars are where the genes locate. One strand of mtDNA is rich in G bases and called as heavy chain while the other is rich in C bases and called as light chain. O_L and O_H denote the replication origin of light chain and heavy chain, respectively. P_L and P_H are the location of promoters of light and heavy chains. The region other than control region is called as coding region

4.3 Emergence of Modern Humans and Establish of Genetic Diversity of Current Human Populations

Fossils of *Australopithecus* and earlier hominids can be found in only Africa. However, the fossils of *Homo* can be found in Europe and Asia in addition to Africa. This leads to the hypothesis of “Out of Africa” of early *Homo*. This hypothesis supposes that the place of origin of hominids is Africa. Furthermore, the evidences of *Homo* in Europe and Asia lead to the hypothesis of “Multiregional origin” of *Homo sapiens*. This explains that the diversity of current humans, such as Africans, Asian, and Europeans was established at the time of early *Homo*’s Out of Africa, around

Table 4.2 The overview of genetic diversity in human

	Data set	Length	*n*	*S*	π
Total	African	16,556	21	358	4.6×10^{-3}
	Non-African	16,555	32	367	2.3×10^{-3}
D-loop	African	1121	21	77	1.8×10^{-2}
	Non-African	1118	32	103	1.1×10^{-2}

Total shows the diversity of whole mtDNA. Length is the number of nucleotides used for obtain the diversity. *n* denotes the number of sample. *S* denotes the number of segregating sites. π shows the nucleotide diversity. Data are referred from Ingman et al. (2000)

2 million years ago. In this hypothesis, it is explained that Asian are the descendants of early *Homo* found in Asia and Europeans are the descendants of early *Homo* in Europe. The alternative hypothesis to explain the current diversity of *Homo sapiens* is "Single origin" hypothesis. This explains that *Homo sapiens* emerged (possibly once) in Africa, after early *Homo* spread out to the Old World. Then *Homo sapiens* experienced secondary Out of Africa and current diversity of humans was established afterward.

If the origin of *Homo sapiens* or hominids is Africa and if there was Out of Africa event, we can predict that the genetic diversity of Africans will be larger than non-Africans. We can also predict that the time to the most common ancestors (tMRCA) of humans will be around one to two million years ago if multiregional origin hypothesis is correct while we can predict that the tMRCA will be around 100–200 thousand years ago if Single origin hypothesis is correct.

MtDNA data suggest the larger genetic diversity within African populations than Asians or Europeans (Table 4.2) in both total mtDNA and D-loop region only (Ingman et al. 2000). This suggests the African origin of *Homo sapiens*. MtDNA also suggest that tMRCA of current human populations is only around 100 thousand years ago. This early date is not consistent with multiregional origin hypothesis and supports the Single origin hypothesis.

The phylogeny of mtDNA (Fig. 4.2) also suggests one important aspect of human genetic diversity. According to this phylogenetic tree, the range of genetic diversity of Asians and Europeans is within the range of diversity of Africans. Lower genetic diversities of Asians and Europeans are concordant with this. Furthermore, these observations of genetic diversity also support the African origin of *Homo sapiens*. In the phylogeny of mtDNA, Asians and Europeans form a clade. This suggests that Asians and Europeans are derived from one of the lineages in Africa. The time of the split of this lineage from the others should correspond to the time for Out of Africa of *Homo sapiens* (Asterisk in Fig. 4.2). According to this tree, this time is estimated to be around 52,000 years ago (Ingman et al. 2000).

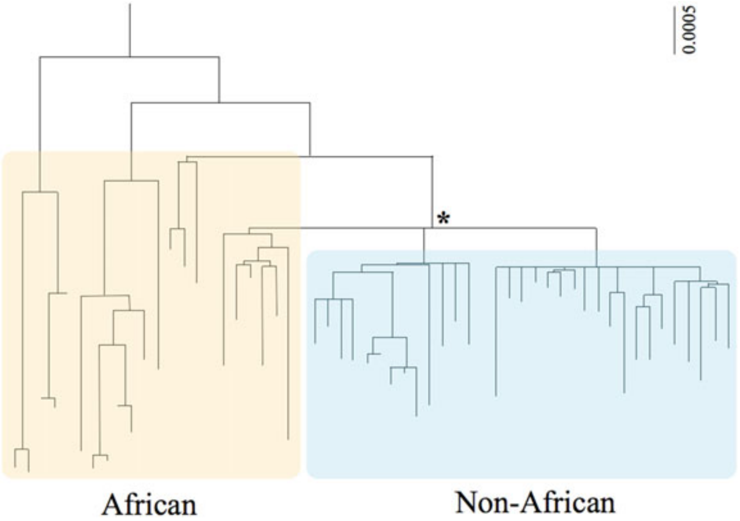


Fig. 4.2 Neighbor-joining phylogeny of 53 complete human mtDNA. MtDNAs are classified into African or non-African. Asterisk shows the branching of non-African lineage from African lineage. Data are referred from Ingman et al. (2000)

4.3.1 Neanderthals and Modern Humans

The fossils of Neanderthals (*Homo neanderthalensis*) suggest that they lived until Late Pleistocene (~4 Kya) and also suggest the co-existence with *Homo sapiens.* Whether there was a gene flow between two groups of hominids can be tested by genetics research. The early studies on this problem could only rely on mtDNA because it has many copies and can be extracted easier than nuclear genome. The first study of Neanderthals of mtDNA was done on HVRI (Krings et al. 1997). As a result, the Neanderthal mtDNA did not fall into the range of *Homo sapiens* diversity and it seemed to be a distinct lineage from modern humans (Fig. 4.3). The divergence time of Neanderthal-modern humans is four times greater than that of MRCA of modern human mtDNAs. Therefore, it was concluded that Neanderthals did not contribute genetically to mtDNA lineages of modern humans. The whole mtDNA genome of Neanderthal was successfully determined later and reached to the same conclusion as the result of HVRI (Green et al. 2008). Note that the evidence of gene flow between Neanderthal and non-Africans was found in nuclear genome (Green et al. 2010).

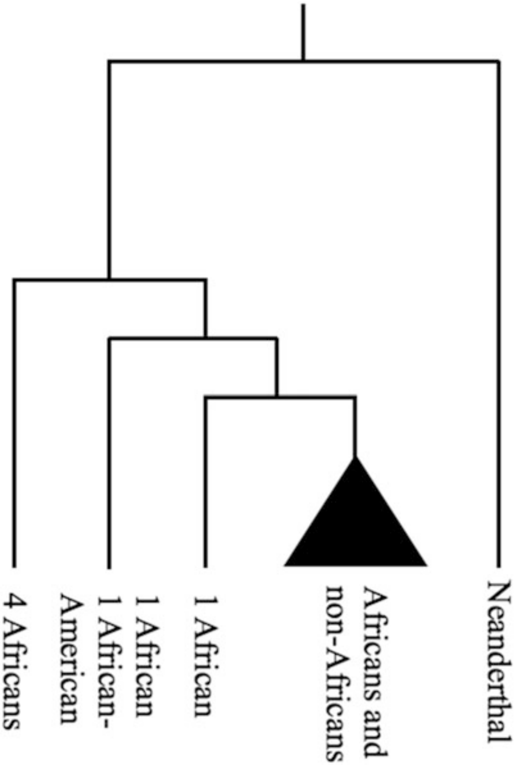


Fig. 4.3 Schematic representation of the mtDNA phylogeny of humans and Neanderthal. Note that mtDNA of Neanderthal is the different lineage from that of humans. Modified from Krings et al. (1997)

4.4 Controversy Between Pedigree Based Estimation and Divergence Based Estimation

Molecular clock is a powerful tool for estimating the divergence time among groups of human or between species (Kumar 2005; Takahata 2007). What we need for estimating divergence time is mutation rate. Typically, reference of independently determined time needed for calibrating the mutation rate. The choice of mutation rate can affect the result of estimated divergence time.

There are mainly two ways to estimate the mutation rate of nucleotide sequence, pedigree method and divergence method. Pedigree method obtains sequences from closely related individuals and counts the number of de novo mutation per generation. On the other hand, divergence method compares the number of nucleotide difference between human and closely related species or among human population and divide by the divergence time to obtain the rate. When the rate of per site per year is needed, the generation time in year is required for pedigree method. Generation time of 25 years is preferable for mtDNA (Fenner 2005). For divergence method, reference divergence time from independent study is needed. There is substantial difference of estimated mtDNA mutation rate in human between these two methods (Howell et al. 2003). Under neutral, the substitution rate should be equal to mutation rate. This is the logical base of estimating mutation rate from substitution data.

Counting a new mutation (de novo mutations) in the context of pedigree method can contain the mutations eventually removed from population. First of all, the somatic mutations are not inherited to the next generation. The mutations found in males have no chance to be transmitted, therefore including these mutations for estimating the mutation rate might lead to the overestimation of the rate. Heteroplasmy is a transient state of mutation fixation, and therefore counting the mutation result from heteroplasmy might also lead to the overestimation of mutation rate (Santos et al. 2005).

The different nature of mtDNA than nuclear can add complications to understand the mutation rate. There are several population levels of mtDNA, in mitochondria, in a cell, in a tissue, and in an individual. A newly emerged mutation in mtDNA might not always transmit to the next generation. During ovulation, the number of mtDNA copies in an oocyte can reach to the order of 10^5. But these copies are identical and result to a homoplasmy state of the fetus. A bottleneck of mtDNA during oocytogenesis likely occurs restricting the transmission of mtDNA to offspring. The number of segregating units in an individual is shown to be on the order of 10 despite high copy number of mtDNA in an oocyte (Brown et al. 2001). Shifting heteroplasmic state to homoplasmic state is needed for a new mutation in mtDNA, which can be considered as a fixation process for mtDNA population in a cell. Santos et al. (2005) examined the discrepancy of mtDNA mutation rate estimated by pedigree method and divergence method. They estimated the mutation rate of mtDNA control region using pedigree and further they exclude the somatic mutations and the mutations exclusively observed in males. They incorporate the bottleneck process during the oocytogenesis, supposing the number of segregating units as 30. They also incorporate the process from heteroplasmy to homoplasmy. As a result of these considerations, the estimated rate from pedigree became similar to the rate from phylogeny.

4.5 Haplogroups

Closely related mtDNAs sequences are expected to form a cluster and can be grouped together. Since mtDNA is a stretch of a non-recombining DNA, this grouped mtDNAs are called as "haplogroups" and a haplogroup is denoted as a single capital of alphabet. The subgroups of haplogroups can be defined and are denoted by number or lower-case letter followed by the first capital. Nesting of haplogroups is allowed. For example, halpogroup M and haplogroup N are two major lineages among non-Africans but these two are branches included in haplogroup L3 which is African lineage. The first haplogroups are described for Native Americans (Torroni et al. 1993). This study revealed four haplogroups, A, B, C, and D in Native Americans. Cladistics notation was once proposed for mtDNA haplogroups (Richards et al. 1998). However, not all the newly defined mtDNA haplogroups follow the suggested rule. In some cases, the same haplogroup name is used for different groups of mtDNA sequences and different haplogroup names are used for the same group of mtDNA sequences (van Oven and Kayser 2009).

When a new mtDNA sequence which has a single nucleotide difference form extant haplogroups is found, a new subgroup of mtDNA haplogroup can be defined. Ultimately, every different mtDNA can have their own subgroup. However, even a pair of mtDNAs derived from two closely related individuals can differ in one nucleotide position. Because haplotype and their haplogroups should be useful for convenient clustering of human mtDNA, researcher should be careful for defining a new subgroup of mtDNA haplogroup.

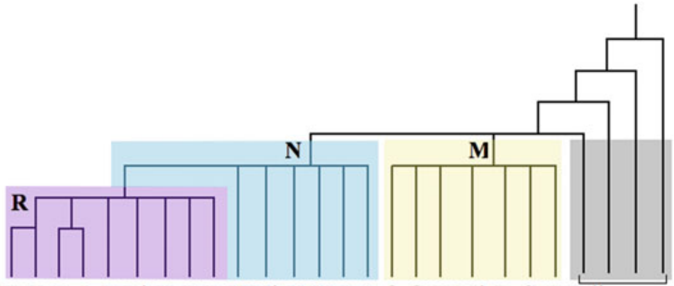
Mutations in an mtDNA determine which haplogroup should that mtDNA is classified, i.e. haplogroup is determined by a certain combination of mutations. Not only nucleotide changes but also indels are used for determination. Note that recurrent mutations are also used for haplotype assigning (Homoplasy). Sequence from whole mtDNA genome is ideal for determining the haplogroups, because not only the mutations in the control regions but also the mutations in the coding regions can be a key for the determination. However, there are studies that rely only on control region for haplogroups determination. Then caution will be needed when these data are taken into account.

4.5.1 Phylogeny of Haplogroups

The evolutionary relationship among haplogroups can be represented by a phylogeny. Because of no recombination, the whole genealogy of human mtDNA can be represented in a single phylogenetic tree (Fig. 4.4). The haplogroup L is considered to be the oldest haplogroup, i.e. haplogroup diverged from the other lineage of haplogroup first in the evolutionary history of human mtDNA, and the haplotype L0 considered to be the oldest. These haplogroups L are seen in only Africans, which is consistent with African origin of *Homo sapiens*. Macrohaplogroups M and N are seen in non-Africans and derived from haplogroups L3. Haplogroup L3 also gives rise to subgroups of L3 which are African lineages. Macrohaplogroup M further gives rise to haplogroup D and lineages including C, E, G, Q, and Z. From macrohaplogroup N, macrohaplogroup of R and lineages leading to A, I, S, W, X, and Y are derived. Macrohaplogroup R includes B, F, H, J, K, P, T, and U. The geographical distributions of these haplogroups are shown in Fig. 4.4.

4.6 Demographic History Estimation Based on mtDNA

Because mtDNA represents single genealogy, simple coalescent model without recombination can be applied to describe its genealogy. Changes of effective population size affect the probability of coalescent event because this probability



	K	U	T	J	P	R*	H	F	B	Y	N*	X	W	I	A	Q	G	E	M*	D	C	Z	L lineages
Europe	K	U		J	P		H					X	W	I									
Middle east	K	U	T			R*	H				N*	X	W	I					M*				L
North Africa		U																	M*				L
Sub-Saharan Africa																							L
South Asia		U	T			R*					N*	X							M*				
East Asia					P			F	B	Y	N*				A		G	E	M*	D	C	Z	
Oceania					P				B		N*					Q			M*				
North America									B			X			A					D	C		
South America									B						A					D	C		

Fig. 4.4 The phylogeny of haplogroups of human mtDNA is shown above. Macrohaplogroups L, M, N, and R are shown in different colors. The simplified geographical distribution of macrohaplogroup is shown below. Data are referred from Jobling et al. (2014)

correlates with a reciprocal of effective population size. If the population experienced population expansion, the coalescent event is more likely to occur in past when the effective population size was smaller. This will result in longer external branches and larger number of singleton sites compared with constant size of effective population. By testing the deviation from expectation under constant effective population size, we can test whether the population experienced change of effective population size (e.g. Tajima's D, Tajima 1989). Note that we can only discuss the effective population size in female lineage and only handle single genealogy when using mtDNA.

Bayesian Skyline Plot (BSP) method (Drummond et al. 2005) is sometime applied to mtDNAs sampled from human population. BSP is implemented in BEAST software (Bouckaert et al. 2014). In BSP, standard Markov chain Monte Carlo (MCMC) sampling procedures are used to infer a posterior distribution of effective population size along the time. By doing this, more detailed information about changes of effective population size can be expected to be referred from mtDNA sequences. An example of BSP from mtDNA is shown below.

4.6.1 mtDNA Research Example Using BSP: Peopling to America

American continent is the last frontier where *Homo sapiens* migrated. It is supposed that the first migrant went through the Bering Strait at the last ice age from Asia (Williams et al. 1985; Greenberg et al. 1986). At that time, the level of sea was lower than current time and the Siberia and Alaska were connected by the land bridge. This land bridge is often referred as "Beringia" and supposed to function as refugia. The first migrants were around Alaska region but further migration into the rest of American continent was blocked by the two large ice sheets. They are called as Laurentide and Cordilleran ice sheets and these collided with each other. After the ice age, these ice sheets started to retreat. The collision of these ice sheets was dissolved and a route connecting Alaska and the rest of American continent emerged. This route was named as "ice-free corridor" and suggested to be the route for peopling to America. However, if coat route was taken, the first migrant did not need to wait until the opening of the ice-free corridor.

Based on the facial morphology of the skull found in archeological sites in America, the discontinuity of current Native Americans and the older skeletal remains (Paleoamericans) are suggested and the ancestry of the first migrant is under debate. Based on the similarity between Clovis culture of America from 13,500 to 12,900 years ago and Solutrean technology in France, Spain, and Portugal, from roughly 23,500 to 20,000 years ago, Solutrean hypothesis proposes that the people with Solutrean technology first colonized the New World and the ancestors of current Native Americans came afterwards (Stanford and Bradley 2002). The study of mtDNA in Native Americans showed the close relationship between current Native Americans and East Asians and proposed the Asian origin of Native Americans (Schurr et al. 1990). But the existence of haplogroup X in Native Americans is not necessarily consistent with Asian origin because X is often observed in Europe (Smith et al. 1999). More than one haplotypes in Native Americans suggest multiple waves of migrations to America and linguistic or dental data suggest three waves of migrations (Greenberg et al. 1986). There are several questions about the first migrant to America. Who were they? Which route did they take? How many waves of migration were there? The detail demographic history inferred from mtDNA genome can answer to these questions.

The coalescent times of Haplogroups A, B, C, and D are almost same and dated to around 20,000 to 15,000 years ago (Fig. 4.5, Fagundes et al. 2008). This shows that the age of mtDNA in Native Americans is old enough and shows that the current Native Americans are the descendants of the first migrant to America. Similar age among four haplogroups suggests that one major wave of migration can explain the current mtDNA diversity of Native American by explaining a population consist of all four haplogroups of mtDNA entered the America continent (Fig. 4.5). Bayesian Skyline Plot shows that the ancestors of current Native Americans experienced population growth around 15,000 years ago (Fig. 4.5). This estimated time overlaps with the end of Last Glacial Maximum (GLM), which is consistent with the scenario

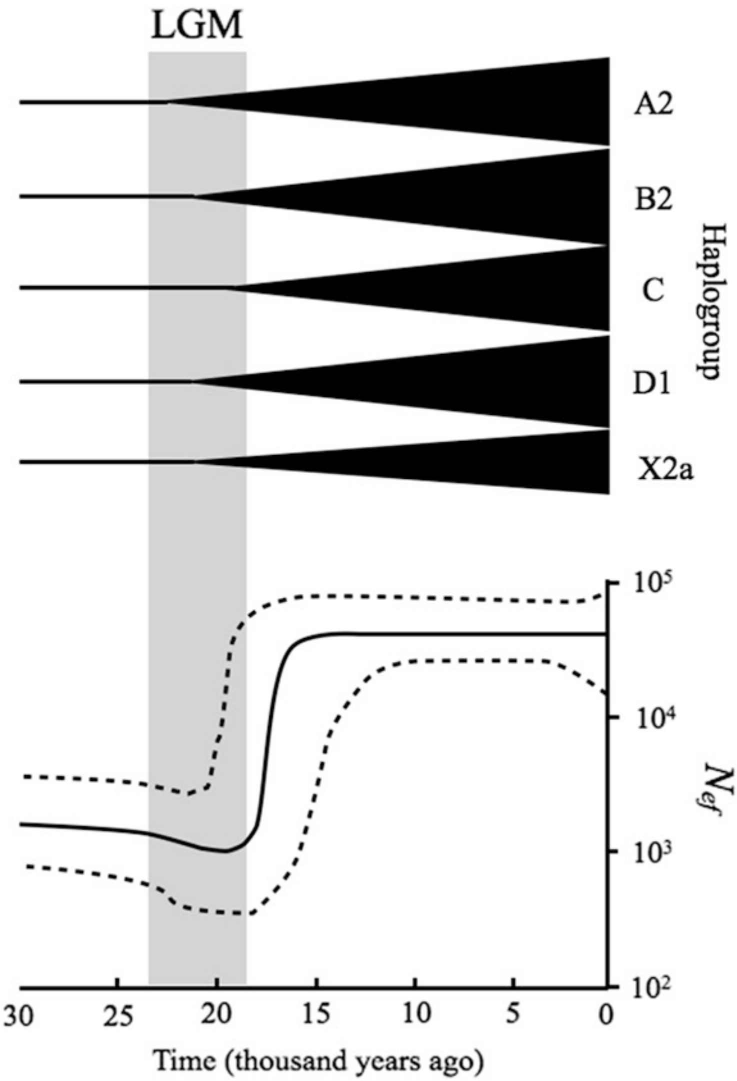


Fig. 4.5 Time to the most recent common ancestor (tMRCAs) of five haplogroups, A, B, C, D, and X2a are well overlapped with the time of last glacial maximum (GLM) and the estimated timing of population expansion. The graph below shows the Bayesian skyline plot (BSP) which represents the estimated changes of female effective population size (N_{ef}) along the time. Modified from Fagundes et al. (2008)

that this timing of population growth can be interpreted as the timing of peopling to America and this age is older than the opening of the ice-free corridor. Therefore the coast route was taken for the first migration to America. In conclusion, mtDNA suggests that the first migrants to America are the ancestors of current Native Americans, there were one major wave of migration and the coast route was taken for the first migration to America. These findings from mtDNA are also confirmed by nucleus genome data (Reich et al. 2012). Ancient DNA of paleoamericans confirms the continuity between them and current Native Americans (Chatters et al. 2014).

4.7 Disease

The 13 protein coding genes in mtDNA encode the subunits of mitochondrial respiratory chain. Therefore, amino acids changes in these genes should affect not only one tissue but also multiple tissues. Disrupting energy conversion process should result severe situation for surviving of an individual. Then we can assume that the mutations in mtDNA genes are deleterious and we might not expect many polymorphisms in these genes. However, we observe a certain frequency of mutations and some of these mutations cause mitochondrial diseases. These diseases include, myopathies, visual and hearing impairment, respiratory disorders, neurological disorders, dementia, and male infertility (Table 4.3). The association of certain haplogroups and diseases is also reported (Table 4.4). Because the subunits of the mitochondrial respiratory chain are not only encoded by mtDNA but also encoded by nuclear genome, the association between mtDNA mutation and disease is not simple (Dowling et al. 2008; Osada and Akashi 2012). Even if an individual inherit the mutation, he or she not necessarily suffers from the disease. While half of the males develop impaired vision, only 10% of females develop the disease in case of Leber hereditary optic neuropathy (LHON, Man et al. 2003). This shows the importance of nuclear genome background on mitochondrial disease. The association between mtDNA mutations and common diseases is reported for diabetes (Patti and Corvera 2010), autism (Haas 2010) and cancer (Wallace 2012). However, these reported associations are not always reproduced by the other study using different

Table 4.3 Examples of mtDNA diseases

Disorder	Phenotype	Type of mutation	Gene	Ref.
LHON (Leber hereditary optic neuropathy)	Optic neuropathy	3460G>A 11778G>A 14484T>C	ND1 ND4 ND6	Howell et al. (1991) Wallace et al. (1988) Johns et al. (1992)
Kearns-Sayre syndrome	Progressive myopathy	4.9 kbp deletion	Several	Moraes et al. (1989)
NIDDM (non-insulin-dependent diabetes mellitus)	Diabetes, deafness	3243A>G	TRNL1	van den Ouweland et al. (1992)
Sensorineural hearing loss	Deafness	1555A>G	RNR1	Prezant et al. (1993)

Table 4.4 Examples of association between mtDNA haplogroup and disease

Disease	Haplogroup	Population	Ref.
Parkinson's disease	J, K	Danish	Gaweda-Walerych et al. (2008)
Parkinson's disease	K	Italian	Ghezzi et al. (2005)
Alzheimer disease	U	Europeans	van der van der Walt et al. (2004)
Asthenozoospermia	H, T	Europeans	Ruiz-Pesini et al. (2000)
Hypertrophic cardiomyopathy	H, J, K	Danish	Hagen et al. (2013)
Sporadic amyotrophic lateral sclerosis	I	Europeans	Mancuso et al. (2004)
Ischemic stroke	K	Europeans	Chinnery et al. (2010)

Some haplogroups were shown to decrease the risk of diseases while the others increase risk

population (Mancuso et al. 2007; Hudson et al. 2012). This also suggests epistasis effect of nuclear genome background.

4.7.1 Missing Heritability and mtDNA Variation

As a result from a lot of the genome wide association study (GWAS), many mutants (SNPs in nucleus genome) are reported to be associated with complex diseases. However these reported SNPs can only explain a small portion of heritability and only have small effect on diseases (odds ratio of 1.1–1.5). This problem is famous as "missing heritability" (Manolio et al. 2009). The attempt to fill this missing heritability by incorporating mtDNA variant information was done by studying over 50,000 European individuals (Hudson et al. 2014). It was expected that the same variant of mtDNA affects different complex diseases because the mutation on mitochondrial respiratory chain should affect multiple tissues. It was shown that the same variant of mtDNA associates with multiple diseases in both ways, and it prevents one disease and increases the risk of the other disease. Overall, the effects of mtDNA variants seem to bias toward increasing disease risk. As mtDNA variation data tend not to be included in GWAS and the less effort is needed to determine whole mtDNA sequence than before, it is suggested to include mitochondrial genome in large-scale genetic association studies for better understanding of complex diseases.

References

Anderson S, Bankier AT, Barrell BG, De Bruijn M (1981) Sequence and organization of the human mitochondrial genome. Nature 290:457–465

Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet 23:147

Awadalla P, Eyre-Walker A, Smith JM (1999) Linkage disequilibrium and recombination in hominid mitochondrial DNA. Science 286:2524–2525

Behar DM, van Oven M, Rosset S, Metspalu M, Loogväli E-L, Silva NM, Kivisild T, Torroni A, Villems R (2012) A "Copernican" reassessment of the human mitochondrial DNA tree from its root. Am J Hum Genet 90:675–684

Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, Suchard MA, Rambaut A, Drummond AJ (2014) BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput Biol 10:e1003537

Brown DT, Samuels DC, Michael EM, Turnbull DM, Chinnery PF (2001) Random genetic drift determines the level of mutant mtDNA in human primary oocytes. Am J Hum Genet 68:533–536

Chatters JC, Kennett DJ, Asmerom Y, Kemp BM, Polyak V, Blank AN, Beddows PA, Reinhardt E, Arroyo-Cabrales J, Bolnick DA, Malhi RS, Culleton BJ, Erreguerena PL, Rissolo D, Morell-Hart S, Stafford TW Jr (2014) Late Pleistocene human skeleton and mtDNA link Paleoamericans and modern native Americans. Science 344:750–754

Chinnery PF, Elliott HR, Syed A, Rothwell PM, Oxford Vascular Study (2010) Mitochondrial DNA haplogroups and risk of transient ischaemic attack and ischaemic stroke: a genetic association study. Lancet Neurol 9:498–503. https://doi.org/10.1016/S1474-4422(10)70083-1

Dowling DK, Friberg U, Lindell J (2008) Evolutionary implications of non-neutral mitochondrial genetic variation. Trends Ecol Evol 23:546–554

Drummond AJ, Rambaut A, Shapiro B, Pybus OG (2005) Bayesian coalescent inference of past population dynamics from molecular sequences. Mol Biol Evol 22:1185–1192

Fagundes NJRN, Kanitz RR, Eckert RR, Valls ACSA, Bogo MRM, Salzano FMF, Smith DGD, Silva WAW, Zago MAM, Ribeiro-dos-Santos AKA, Santos SEBS, Petzl-Erler MLM, Bonatto SLS (2008) Mitochondrial population genomics supports a single pre-Clovis origin with a coastal route for the peopling of the Americas. Am J Hum Genet 82:583–592

Fenner JN (2005) Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. Am J Phys Anthropol 128:415–423

Gaweda-Walerych K, Maruszak A, Safranow K, Bialecka M, Klodowska-Duda G, Czyzewski K, Slawek J, Rudzinska M, Styczynska M, Opala G, Drozdzik M, Canter JA, Barcikowska M, Zekanowski C (2008) Mitochondrial DNA haplogroups and subhaplogroups are associated with Parkinson's disease risk in a Polish PD cohort. J Neural Transm 115:1521–1526

Ghezzi D, Marelli C, Achilli A, Goldwurm S, Pezzoli G, Barone P, Pellecchia MT, Stanzione P, Brusa L, Bentivoglio AR, Bonuccelli U, Petrozzi L, Abbruzzese G, Marchese R, Cortelli P, Grimaldi D, Martinelli P, Ferrarese C, Garavaglia B, Sangiorgi S, Carelli V, Torroni A, Albanese A, Zeviani M (2005) Mitochondrial DNA haplogroup K is associated with a lower risk of Parkinson's disease in Italians. Eur J Hum Genet 13:748–752

Green RE, Malaspinas A-S, Krause J, Briggs AW, Johnson PLF, Uhler C, Meyer M, Good JM, Maricic T, Stenzel U, Prüfer K, Siebauer M, Burbano HA, Ronan M, Rothberg JM, Egholm M, Rudan P, Brajkovic D, Kucan Z, Gusic I, Wikström M, Laakkonen L, Kelso J, Slatkin M, Pääbo S (2008) A complete neandertal mitochondrial genome sequence determined by high-throughput sequencing. Cell 134:416–426

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MHY, Hansen NF, Durand EY, Malaspinas AS, Jensen JD, Marques-Bonet T, Alkan C, Prufer K, Meyer M, Burbano HA, Good JM, Schultz R, Aximu-Petri A, Butthof A, Hober B, Hoffner B, Siegemund M, Weihmann A, Nusbaum C, Lander ES, Russ C, Novod N, Affourtit J, Egholm M, Verna C, Rudan P, Brajkovic D, Kucan Z, Gusic I, Doronichev VB, Golovanova LV, Lalueza-Fox C, la Rasilla de M, Fortea J, Rosas A, Schmitz RW, Johnson PLF, Eichler EE, Falush D, Birney E, Mullikin JC, Slatkin M, Nielsen R, Kelso J, Lachmann M, Reich D, Pääbo S (2010) A draft sequence of the neandertal genome. Science 328:710–722

Greenberg JH, Turner CG, Zegura SL, Campbell L (1986) The settlement of the Americas: a comparison of the linguistic, dental, and genetic evidence. Curr Anthropol 27:477–497

Haas RH (2010) Autism and mitochondrial disease. Dev Disabil Res Rev 16:144–153

Hagen CM, Aidt FH, Hedley PL, Jensen MK, Havndrup O, Kanters JK, Moolman-Smook JC, Larsen SO, Bundgaard H, Christiansen M (2013) Mitochondrial haplogroups modify the risk of developing hypertrophic cardiomyopathy in a Danish population. PLoS ONE 8:e71904

Howell N, Bindoff LA, McCullough DA, Kubacka I, Poulton J, Mackey D, Taylor L, Turnbull DM (1991) Leber hereditary optic neuropathy: identification of the same mitochondrial ND1 mutation in six pedigrees. Am J Hum Genet 49:939–950

Howell N, Smejkal CB, Mackey DA, Chinnery PF, Turnbull DM, Herrnstadt C (2003) The pedigree rate of sequence divergence in the human mitochondrial genome: there is a difference between phylogenetic and pedigree rates. Am J Hum Genet 72:659–670

Hudson G, Sims R, Harold D, Chapman J, Hollingworth P, Gerrish A, Russo G, Hamshere M, Moskvina V, Jones N, Thomas C, Stretton A, Holmans PA, O'Donovan MC, Owen MJ, Williams J, Chinnery PF, GERAD1 Consortium (2012) No consistent evidence for association between mtDNA variants and Alzheimer disease. Neurology 78:1038–1042

Hudson G, Gomez-Duran A, Wilson IJ, Chinnery PF (2014) Recent mitochondrial DNA mutations increase the risk of developing common late-onset human diseases. PLoS Genet 10:e1004369

Ingman M, Kaessmann H, Pääbo S, Gyllensten U (2000) Mitochondrial genome variation and the origin of modern humans. Nature 408:708–713

Innan H, Nordborg M (2002) Recombination or mutational hot spots in human mtDNA? Mol Biol Evol 19:1122–1127

Jobling MA, Hollox E, Hurles M, Kivisild T, Tyler-Smith C (2014) Human evolutionary genetics. Garland Science, New York

Johns DR, Neufeld MJ, Park RD (1992) An ND-6 mitochondrial DNA mutation associated with leber hereditary optic neuropathy. Biochem Biophys Res Commun 187:1551–1557

Jorde LB, Bamshad M (2000) Questioning evidence for recombination in human mitochondrial DNA. Science 288:1931

Kivisild T, Villems R (2000) Questioning evidence for recombination in human mitochondrial DNA. Science 288:1931

Kraytsberg Y, Schwartz M, Brown TA, Ebralidse K, Kunz WS, Clayton DA, Vissing J, Khrapko K (2004) Recombination of human mitochondrial DNA. Science 304:981

Krings M, Stone A, Schmitz RW, Krainitzki H, Stoneking M, Pääbo S (1997) Neandertal DNA sequences and the origin of modern humans. Cell 90:19–30

Kumar S (2005) Molecular clocks: four decades of evolution. Nat Rev Genet 6:654–662

Kumar S, Hedrick P, Dowling T, Stoneking M (2000) Questioning evidence for recombination in human mitochondrial DNA. Science 288:1931

Li M, Schönberg A, Schaefer M, Schroeder R, Nasidze I, Stoneking M (2010) Detecting heteroplasmy from high-throughput sequencing of complete human mitochondrial DNA genomes. Am J Hum Genet 87:237–249

Man PYW, Griffiths PG, Brown DT, Howell N, Turnbull DM, Chinnery PF (2003) The epidemiology of Leber hereditary optic neuropathy in the North East of England. Am J Hum Genet 72:333–339

Mancuso M, Conforti FL, Rocchi A, Tessitore A, Muglia M, Tedeschi G, Panza D, Monsurrò M, Sola P, Mandrioli J, Choub A, DelCorona A, Manca ML, Mazzei R, Sprovieri T, Filosto M, Salviati A, Valentino P, Bono F, Caracciolo M, Simone IL, La Bella V, Majorana G, Siciliano G, Murri L, Quattrone A (2004) Could mitochondrial haplogroups play a role in sporadic amyotrophic lateral sclerosis? Neurosci Lett 371:158–162

Mancuso M, Nardini M, Micheli D, Rocchi A, Nesti C, Giglioli NJ, Petrozzi L, Rossi C, Ceravolo R, Bacci A, Choub A, Ricci G, Tognoni G, Manca ML, Siciliano G, Murri L (2007) Lack of association between mtDNA haplogroups and Alzheimer's disease in Tuscany. Neurol Sci 28:142–147

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TFC, McCarroll SA, Visscher PM (2009) Finding the missing heritability of complex diseases. Nature 461:747–753

Moraes CT, DiMauro S, Zeviani M, Lombes A, Shanske S, Miranda AF, Nakase H, Bonilla E, Werneck LC, Servidei S (1989) Mitochondrial DNA deletions in progressive external ophthalmoplegia and Kearns-Sayre syndrome. N Engl J Med 320:1293–1299

Osada N, Akashi H (2012) Mitochondrial-nuclear interactions and accelerated compensatory evolution: evidence from the primate cytochrome C oxidase complex. Mol Biol Evol 29:337–346

Pakendorf B, Stoneking M (2005) Mitochondrial DNA and human evolution. Annu Rev Genomics Hum Genet 6:165–183

Parsons TJ, Irwin JA (2000) Questioning evidence for recombination in human mitochondrial DNA. Science 288:1931

Patti ME, Corvera S (2010) The role of mitochondria in the pathogenesis of type 2 diabetes. Endocr Rev 31:364–395

Prezant TR, Agapian JV, Bohlman MC, Bu X, Öztas S, Qiu W-Q, Arnos KS, Cortopassi GA, Jaber L, Rotter JI, Shohat M, Fischel-Ghodsian N (1993) Mitochondrial ribosomal RNA mutation associated with both antibiotic-induced and non-syndromic deafness. Nat Genet 4:289–294

Ramos A, Santos C, Mateiu L, Gonzalez MDM, Alvarez L, Azevedo L, Amorim A, Aluja MP (2013) Frequency and pattern of heteroplasmy in the complete human mitochondrial genome. PLoS ONE 8:e74636

Reich D, Patterson N, Campbell D, Tandon A, Mazieres S, Ray N, Parra MV, Rojas W, Duque C, Mesa N, García LF, Triana O, Blair S, Maestre A, Dib JC, Bravi CM, Bailliet G, Corach D, Hünemeier T, Bortolini MC, Salzano FM, Petzl-Erler ML, Acuña-Alonzo V, Aguilar-Salinas C, Canizales-Quinteros S, Tusié-Luna T, Riba L, Rodríguez-Cruz M, Lopez-Alarcón M, Coral-Vazquez R, Canto-Cetina T, Silva-Zolezzi I, Fernandez-Lopez JC, Contreras AV, Jimenez-Sanchez G, Gómez-Vázquez MJ, Molina J, Carracedo A, Salas A, Gallo C, Poletti G, Witonsky DB, Alkorta-Aranburu G, Sukernik RI, Osipova L, Fedorova SA, Vasquez R, Villena M, Moreau C, Barrantes R, Pauls D, Excoffier L, Bedoya G, Rothhammer F, Dugoujon J-M, Larrouy G, Klitz W, Labuda D, Kidd J, Kidd K, Di Rienzo A, Freimer NB, Price AL, Ruiz-Linares A (2012) Reconstructing native American population history. Nature 488:370–374

Richards MB, Macaulay VA, Bandelt H-J, Sykes BC (1998) Phylogeography of mitochondrial DNA in western Europe. Ann Hum Genet 62:241–260

Ruiz-Pesini E, Lapeña AC, Díez-Sánchez C, Pérez-Martos A, Montoya J, Alvarez E, Díaz M, Urriés A, Montoro L, López-Pérez MJ, Enríquez JA (2000) Human mtDNA haplogroups associated with high or reduced spermatozoa motility. Am J Hum Genet 67:682–696

Santos C, Montiel R, Sierra B, Bettencourt C, Fernandez E, Alvarez L, Lima M, Abade A, Aluja MP (2005) Understanding differences between phylogenetic and pedigree-derived mtDNA mutation rate: a model using families from the Azores Islands (Portugal). Mol Biol Evol 22:1490–1505

Scally A, Durbin R (2012) Revising the human mutation rate: implications for understanding human evolution. Nat Rev Genet 13:745–753

Schurr TG, Ballinger SW, Gan YY, Hodge JA, Merriwether DA, Lawrence DN, Knowler WC, Weiss KM, Wallace DC (1990) Amerindian mitochondrial DNAs have rare Asian mutations at high frequencies, suggesting they derived from four primary maternal lineages. Am J Hum Genet 46:613–623

Smith DG, Malhi RS, Eshleman J, Lorenz JG, Kaestle FA (1999) Distribution of mtDNA haplogroup X among Native North Americans. Am J Phys Anthropol 110:271–284

Stanford D, Bradley B (2002) Ocean trails and prairie paths? Thoughts about Clovis origins. In: Jablonski N (ed) The First Americans: the Pleistocene colonization of the new world, vol 27. Memoirs of the California Academy of Sciences, San Francisco, pp 255–271

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595

Takahata N (2007) Molecular clock: an anti-neo-Darwinian legacy. Genetics 176:1–6

Torroni A, Schurr TG, Cabell MF, Brown MD, Neel JV, Larsen M, Smith DG, Vullo CM, Wallace DC (1993) Asian affinities and continental radiation of the four founding Native American mtDNAs. Am J Hum Genet 53:563–590

van den Ouweland JMW, Lemkes HHPJ, Ruitenbeek W, Sandkuijl LA, de Vijlder MF, Struyvenberg PAA, van de Kamp JJP, Maassen JA (1992) Mutation in mitochondrial tRNALeu(UUR) gene in a large pedigree with maternally transmitted type II diabetes mellitus and deafness. Nat Genet 1:368–371

van der Walt JM, Dementieva YA, Martin ER, Scott WK, Nicodemus KK, Kroner CC, Welsh-Bohmer KA, Saunders AM, Roses AD, Small GW, Schmechel DE, Murali Doraiswamy P, Gilbert JR, Haines JL, Vance JM, Pericak-Vance MA (2004) Analysis of European mitochondrial haplogroups with Alzheimer disease risk. Neurosci Lett 365:28–32

van Oven M, Kayser M (2009) Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat 30:E386–E394. https://doi.org/10.1002/humu.20921

Wallace DC (2012) Mitochondria and cancer. Nat Rev Cancer 12:685–698

Wallace DC, Singh G, Lott MT, Hodge JA, Schurr TG, Lezza AM, Nikoskelainen EK (1988) Mitochondrial DNA mutation associated with Leber's hereditary optic neuropathy. Science 242:1427–1430

White DJ, Bryant D, Gemmell NJ (2013) How good are indirect tests at detecting recombination in human mtDNA? G3; genes|genomes|. Genetics 3:1095–1104

Williams RC, Steinberg AG, Gershowitz H, Bennett PH, Knowler WC, Pettitt DJ, Butler W, Baird R, Dowda-Rea L, Burch TA (1985) GM allotypes in native Americans: evidence for three distinct migrations across the Bering land bridge. Am J Phys Anthropol 66:1–19

Zsurka G, Kraytsberg Y, Kudina T, Kornblum C, Elger CE, Khrapko K, Kunz WS (2005) Recombination of mitochondrial DNA in skeletal muscle of individuals with multiple mitochondrial DNA heteroplasmy. Nat Genet 37:873–877

Chapter 5
The Y Chromosome

Francesc Calafell and David Comas

Abstract Most of the length of the Y chromosome escapes recombination with the X chromosome and is strictly paternally inherited. This has profound evolutionary implications and provides (together with the matrilineal mitochondrial DNA) unique tools in human population genetics. Absence of recombination implies that a most parsimonious tree can be easily constructed from SNPs and other slowly mutating polymorphisms. The main branches of this tree, called haplogroups, have distinct geographic distributions and can be used to trace human migrations. Faster evolving polymorphisms, such as microsatellites provide readily time estimates for these migrations and other demographic events. Beyond sets of predefined polymorphism, sequencing of most of the non-recombining portion of the Y chromosome has yielded an accurate picture of the global evolution of this chromosome. Combining the Y chromosome with mitochondrial DNA has revealed sex-specific migrations, particularly in the Colonial period. Since surnames are also patrilineally inherited in many populations, the analysis of Y chromosome variation within surnames has shed light on the dynamics of surnames in populations, but has also contributed to the investigation of notorious lineages, such as the Columbus, Bourbons, and Draculs. However, it also raises the possibility of predicting a surname from an anonymous sample, which may be an important tool in forensic genetics but raises also privacy concerns for participants in genetic studies.

Keywords Y chromosome · Phylogeography · Haplogroup

F. Calafell (✉)
Institut de Biologia Evolutiva, Universitat Pompeu Fabra, Barcelona, Catalonia, Spain
e-mail: francesc.calafell@upf.edu

D. Comas
Institut de Biologia Evolutiva (CSIC-UPF), Universitat Pompeu Fabra, Barcelona, Catalonia, Spain
e-mail: david.comas@upf.edu

N. Saitou (ed.), *Evolution of the Human Genome II*, Evolutionary Studies,
https://doi.org/10.1007/978-4-431-56904-6_5

5.1 The Odd Chromosome and Its History

The Y chromosome is a genomic oddity: it is the squat half of the only mismatched couple in the karyotype, rather smaller than its X partner, and containing fewer genes. As we shall briefly discuss below, these features arose from its function: the presence of the Y chromosome in humans, as in eutherian mammals and other animals, steers development towards the production of a male embryo. The main masculinity switch is a gene called *SRY*. The human Y chromosome only recombines with the X chromosome at two small regions at the ends of the chromosome, called pseudoautosomal regions 1 and 2 (PAR1 and PAR2). The non-recombining region of the Y chromosome (NRY) covers most of its ~23 euchromatic Mb; since it does not recombine with the X chromosome, it is male specific (and thus, it is also denoted MSY) and evolves independently of the X chromosome.

Y chromosomes have arisen multiple times in evolutionary history: once in eutherian mammals, and, independently, multiple times in reptiles, amphibians, and fish (Bachtrog 2013). The sex chromosomes start their evolutionary life as an autosome pair in which a sex-determining gene appears. Any mutation that is beneficial to males but detrimental to females increases its fitness if it is inherited in tight linkage with such sex-determining gene. Thus, lack of recombination in this young Y chromosome is beneficial, and inversions seem to have been selected as a mechanism to inhibit recombination with the X chromosome. This process proceeded in stages until recombination was blocked for most of the length of the Y chromosome (Bachtrog 2013); by comparing the levels of divergence between genes in the Y chromosome and their homologues in the X chromosome, four such stages or strata have been recognized (Lahn and Page 1999). But lack of recombination comes with a price: detrimental mutations are more difficult to eliminate from the Y chromosome, and, on the contrary, beneficial mutations are harder to fixate. Recombination allows decoupling the evolution of a particular mutation from that of its immediate genomic vicinity; recombinant chromosomes lacking a particular deleterious allele will be selected for, increasing the efficiency of natural selection. However, when a deleterious mutation lands on the NRY, negative selection will affect the whole chromosome, including any weakly beneficial mutations. This has led to a drastic purge of genes on the Y chromosome (78 protein-coding genes in the MSY compared with 800 protein-coding genes in the X chromosome); most of the remaining genes have been shown to have male-related functions (Lahn and Page 1997).

The human Y chromosome is made up of five types of sequences: the *pseudoautosomal regions*; a large *heterochromatic block* (~40 Mb); the *X-degenerate region* that derives from the common autosomal ancestor with the X chromosome; the more recently *X-transposed region*, that is unique to humans and was transposed 3–4 million years ago; and the highly repetitive *ampliconic regions*, which make sequencing and assembling the Y chromosome a daunting task. For a detailed discussion on the gene content of the Y chromosome, see Satta and Iwase (2015).

5.2 Not All Y Chromosomes Are Created Equal: Polymorphisms in the Y Chromosome

Genetic polymorphisms in the Y chromosome (as in the rest of the genome) come in different forms, each with their applications in human population genetics, as discussed below. The most commonly used types currently are UEPs (unique event polymorphisms) and STRs (short tandem repeats). Minisatellites (Bouzekri et al. 1998) and probes tested with restriction enzymes (Lucotte and Ngo 1985) are no longer favored given the technical difficulties in genotyping them and the complexity of their analysis.

UEPs are polymorphisms created by mutations at such a slow rate that they can be assumed to have happened just once in the human lineage. They are single nucleotide polymorphisms (SNPs), short indels, and Alu insertions. Given the lack of recombination in the NRY, UEPs can be readily assorted into a maximum parsimony tree; the deepest branches of such a phylogeny (Fig. 5.1) are called haplogroups, and have particular geographical distributions (Fig. 5.2), and, thus, can be (and have been) used in exploring the geographic ancestry of male lineages. The robustness of the phylogeography of NRY UEPs allows to detect easily the few cases in which UEPs are not actually unique but have happened in different branches or have reverted to the ancestral state (see, for example Adams et al. (2006)).

The initial search for diversity in the human Y chromosome was rather disappointing (Jobling and Tyler-Smith 1995). Two large restriction fragment length polymorphism (RFLP) studies screened, on average, 833 bp (Jakubiczka et al. 1989) and 2215 bp (Malaspina et al. 1990) in each of 22 Y chromosomes and found between them only three polymorphisms. Seielstad et al. (1994) sequenced 1.4 kb of DNA in 12–16 chromosomes of diverse geographical origins and a single polymorphic nucleotide substitution was found, and Dorit et al. (1995) did not find any variation in a 729 bp intron in the ZFY gene in 38 males.

The pace of discovery accelerated when denaturing high-performance liquid chromatography (DHPLC) was used to screen for polymorphisms in the Y chromosome. Just 2 years after Jobling and Tyler-Smith (1995) published a summary of the nine UEPs known at the time, Underhill et al. (1997) discovered 19 new polymorphisms. Three years later, the same group had increased the total number of known NRY UEPs to 166, arranged in a phylogeography of 116 haplogroups that is quite similar to the current one (Semino et al. 2000). In parallel, Hammer et al. (2001) used single-stranded conformation polymorphism (SSCP) to detect 43 new variants, which they subsequently genotyped in >2500 men from 50 world populations. At that point, an effort was made to merge the two new large phylogenies and create a new nomenclature for Y chromosome haplogroups (see below)

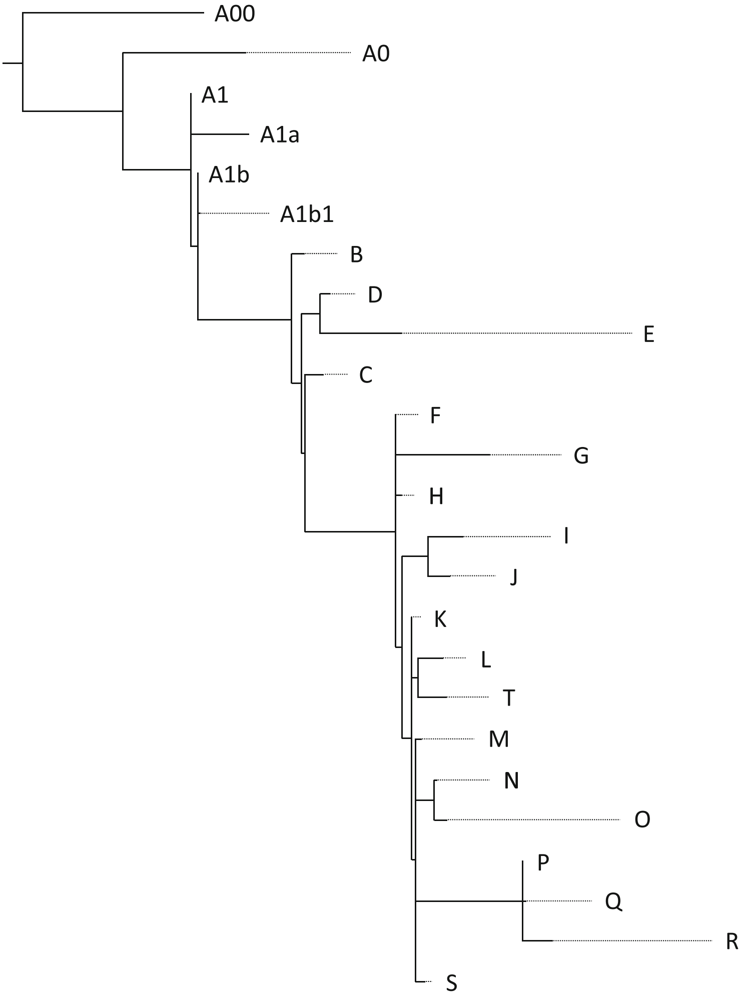


Fig. 5.1 Schematic tree of the Y chromosome haplogroups. Branch lengths are proportional to the number of UEPs in the 2013 ISOGG tree. Dotted lines indicate the length of the deepest branch in each haplogroup (http://www.isogg.org/tree/ISOGG_YDNATreeTrunk.html)

that was modeled on that used for mitochondrial DNA (Calafell et al. 2002; The Y Chromosome Consortium 2002); it was subsequently updated in 2008 (Karafet et al. 2008), and in 2013 (Van Geystelen et al. 2013). In parallel, the International Society for Genetic Genealogy maintains a non-peer-reviewed, continuously updated tree (http://www.isogg.org/tree/ISOGG_YDNATreeTrunk.html).

The diversity, phylogeny, and evolution of many haplogroups were studied in a series of monographic papers that deepened the general knowledge about those haplogroups: A (Cruciani et al. 2011), C (Zhong et al. 2010), E-P2 (Trombetta et al. 2011), E-M215 (Cruciani et al. 2004), E-M78 (Cruciani et al. 2006), G (Rootsi

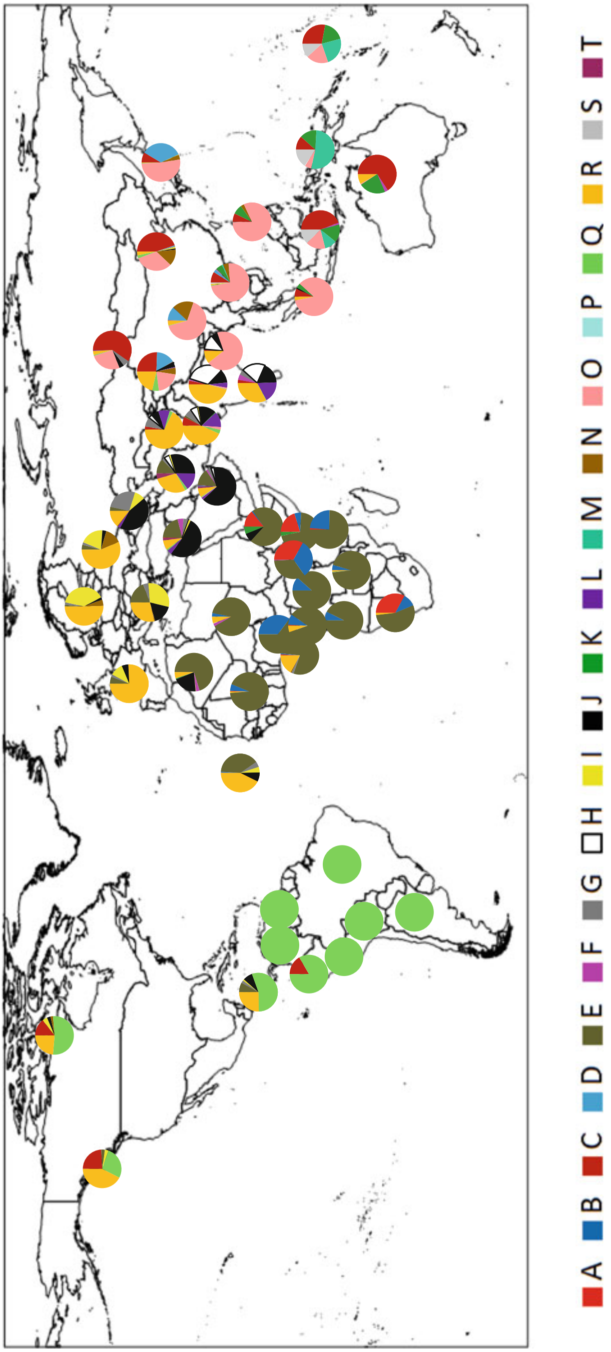


Fig. 5.2 Haplogroup frequencies in selected world populations

et al. 2012), I (Rootsi et al. 2004), J (Semino et al. 2004), N (Rootsi et al. 2007), R1a (Underhill et al. 2010), and R1b-M269 (Myres et al. 2010).

The next leap in the number of known polymorphisms in the Y chromosome happened with the advent of next-generation sequencing technologies, and projects such as the 1000 Genomes Project (http://www.1000genomes.org/).In a pilot project, a 1.8X coverage sequencing of 77 Y chromosomes yielded 2840 high-confidence Y-SNPs, 74% of them new (Durbin et al. 2010). Additionally, Rocca et al. (2012) mined the 1000 Genomes Project data set and established the main branches of the phylogeny of haplogroup R1b-M269 in Western Europe. Subsequently, Y chromosome specific projects have analyzed extensive NRY sequences in population samples. Thus, Wei et al. (2013) focused on ~9 Mb of sequence in 36 worldwide individuals, which allowed them to produce a calibrated phylogeny for the Y chromosome, with dates for its root and some of the main groups. Francalacci et al. (2013) produced low-pass (1.2X) sequences for the same ~9 Mb region in 1204 Sardinian males; they used the Sardinian-specific clades and the known age of the colonization of Sardinia to calibrate the molecular clock and produce a date for the Y chromosome phylogeny (~180,000–200,000 years ago) that is commensurate with the age of the mitochondrial DNA tree. A similar result was reached in a back-to-back paper (Poznik et al. 2013), in which they analyzed a slightly larger sequence (~10 Mb) at a much deeper coverage (~300X) in a smaller sample (69 African, Asian, and Native American males).

Unlike UEPs, STRs mutate at a fast rate (10^{-3}–10^{-4} per meiosis), and, given their stepwise mutations, new mutants are often alleles that already existed in the population; thus, reliable phylogenies cannot be build solely on STRs. However, STR alleles do carry information on the haplogroup background they are found on (and actually, STR variability is partitioned by haplogroups to a greater extent than by populations (Bosch et al. 1999)), to the point that main haplogroups can be reasonable predicted from the STR haplotypes they carry (Athey 2005; Schlecht et al. 2008), but see Larmuseau et al. (2014) for a counterexample. The fast mutation rate in STRs implies that they can be used to time relatively recent events in human evolution, at a prehistoric (Arredi et al. 2004; Alonso et al. 2005; Berniell-Lee et al. 2009; Balaresque et al. 2010; Batini et al. 2011) or historic time scale (Bosch et al. 2003; McEvoy et al. 2006; Adams et al. 2008), or to date specific patrilineal lineages, linked to historic figures or to surnames (Skorecki et al. 1997; Zerjal et al. 2003; McEvoy and Bradley 2006; Hammer et al. 2009; King and Jobling 2009a; Martinez-Cruz et al. 2012; Martínez-González et al. 2012).

5.3 Climbing Down the Y Tree

As mentioned above, the UEPs provide a sound scaffold to produce a phylogenetic tree for the Y chromosome. The main branches are called haplogroups, but, in turn, given the tree structure of the phylogeny, they are subdivided in subhaplogroups, which in turn are further subdivided. The nomenclature of haplogroups (The Y

Chromosome Consortium 2002) starts with a capital Roman letter (A, B,..., T); subhaplogroups are indicated with numbers (A1, A2,...), the next division uses a lowercase letter (A1a, A1b), and, subsequently, numbers and lowercase letters alternate (see, for instance, E1b1b1b2a1 or R1b1a2a1a2a1a1a1). Given how unwieldy this nomenclature can become, haplogroups are often designated with a deep main branch (e.g., E or R1b), plus the most derived UEP that subhaplogroup carries; then, the latter examples become E-M34 and R1b-M153, respectively. In this nomenclature, a final star indicates a paragroup. A Y chromosome falls in a paragroup if it belongs in a haplogroup for which further UEPs have been described but that it is ancestral for those (for instance, E-M78 contains E-V12, E-V13, E-V22, and E-V65; if a Y chromosome is derived for M78 but ancestral for V12, V13, V22, and V65, then it belongs to paragroup E-M78*).

In this section, we will describe the structure, geographical distribution, and history of the main branches of the Y chromosome phylogeny. Further details may be found in the chapters in this book devoted to specific geographic regions. When first described, the deepest branch in the Y tree was labeled A. It is found at moderate frequencies in populations of Eastern and Southern Africa, and at lower frequencies in Central and Northern Africa. Cruciani et al. (2011) discovered new polymorphisms and found that clade A1b split first from the root; it has been subsequently renamed A0. But its position close to the last common ancestor did not last for long: Mendez et al. (2013) discovered that an African-American man carried the ancestral states at all the UEPs known that far; after sequencing a 300-Kb fragment of his Y chromosome, the authors found a new branch of the tree, the closest to the root, which they called A00. It is also found at a very low frequency in Cameroon.

Haplogroup B is also found only in African populations or in populations of African descent. It reaches moderate frequencies in West and Central Africa, and some of its branches are particularly prevalent in Pygmies (Batini et al. 2011).

The next major division of the Y chromosome tree brings together haplogroups D and E, which share one of the first UEPs to be discovered, namely the eponymous YAP (Y Alu Polymorphism). However, the geographical distribution of both haplogroups is quite different. Haplogroup D is found mainly in Japan and Tibet, while different branches of haplogroup E are the most frequent haplogroups in almost all African populations, and reach the Middle East and Southern Europe. E-M81 is prevalent in North Africa and has been used to trace the medieval North African migrations into Iberia and Sicily (Bosch et al. 2001; Adams et al. 2008); similarly, E-V38 may track Bantu migration from central to southern Africa.

Haplogroup C has a vast geographical distribution, from Southern Europe, across Eurasia, and into native North Americans and Polynesians. Several branches of this haplogroup reach their highest frequencies in Central Asia, Japan, New Guinea, and Australia. Within haplogroup C, Zerjal et al. (2003) identified a lineage carried by 8% of men in a large region of Asia, and that, given its age and origin, could represent the descendants of Genghis Khan.

Next to branch is haplogroup F; it contains a few minor branches dispersed over Eurasia, but it also harbors all other haplogroups from G to T. In fact, most men outside of Africa carry some branch of haplogroup F. It has been recently discovered that the first such haplogroup to branch off F is haplogroup G (Poznik et al. 2013). It reaches its peak frequency in the Caucasus, but it is found, probably because of the Neolithic expansion, from Western Europe to South Asia. Haplogroup H, until recently thought to be a sister clade to haplogroup G, is prevalent in South Asia, and its presence in European Roma (Gypsy) men bears testimony to their Indian ancestry.

Two Y chromosome haplogroups that are indeed sister clades are I and J. Haplogroup I is found mostly in Europe, with branches that are among the most frequent among men in Scandinavia, the Balkans, and Sardinia. Haplogroup J peaks in the Middle East, and is found throughout southern Europe, North Africa, and South Asia. The famous "Cohen haplotype" belongs to haplogroup J (Skorecki et al. 1997; Hammer et al. 2009).

The rest of Y chromosome haplogroups is grouped under superhaplogroup K, which, like haplogroup F, has a few minor branches with a sparse distribution found in Oceania and Australia, and only at low frequency in South Asia and the Malay Archipelago. The first clade to branch off K consists of the sister haplogroups L and T: the former is prevalent in South Asia, the Middle East, and Central Italy, while the latter has a sparse distribution from the Middle East to SW Europe.

Another linked pair of haplogroups is constituted by N and O, which represent much of the male lineages in East Asia; N has a more northerly distribution, reaching Northeast Europe to the west, while O is more prevalent to the south, reaching Polynesia. Haplogroup M is also found in New Guinea, extending to Melanesia and Polynesia. A similar, but more restricted distribution, is that of haplogroup S, which is most prevalent in the highlands of New Guinea.

We are reaching the final superhaplogroup: P, which, unlike F and K, has no minor branches. The paragroup P* is present at low to moderate frequencies among South Asian populations. P has two main branches, namely Q and R. Haplogroup Q is found sparsely in Europe and the Middle East, at moderate frequencies in northern Asia, but, with the occasional exception of haplogroup C, is the sole lineage found in native Americans. Finally, haplogroup R has three main branches: R1a, R1b, and R2. R1a is found at high frequencies from Central, Eastern, and Northern Europe, to Central and South Asia. R1b is most prevalent in Western Europe, with peaks >75% in Iberia and Ireland; finally, R2 is much less frequent and mostly restricted to the Middle East.

5.4 La Donna è Mobile … Ma Non Troppo

The Y chromosome diversity tracks the demographic history of men, and, in particular, their migration patterns and rates, and their effective population size. The matrilineal counterpart of the Y chromosome is mitochondrial DNA (mtDNA; see Chap. 6 in this book): although present both in males and females, it is inherited from the mother (with some extremely rare exceptions). The comparison of the diversity and phylogeography of NRY and the mtDNA has been used to learn about the specificities of the demographic history of each sex.

Back in 1987, one of the first genetic pieces of evidence to be found for the Out of Africa model of human evolution came from the mtDNA phylogeography. The mtDNA sequences (analyzed by restriction mapping) were found to coalesce to a common ancestor that may have lived in Africa 200,000 years ago (Cann et al. 1987). This ancestor was promptly dubbed "mitochondrial Eve," and it should be clear that she was not the only woman living at that time, but rather just the one to whom all current humans can trace their maternal ancestry. "Y chromosome Adam" took a little longer to find, but it came with a surprise: he lived only 59,000 years ago (Thomson et al. 2000), a much more recent timeframe than that of his female counterpart. A difference in effective population sizes between the sexes can help resolve this discrepancy. Effective population size, the theoretical figure that governs the intensity of genetic drift, may be lower in men than in women, because the variance in reproductive success may be higher in men. That would be particularly the case in polygynous societies, where a few men can afford many wives (and children), and others may be left childless. However, this may be a recent phenomenon in human history (certainly unheard of in hunter-gatherer societies), and it is unclear for how long male effective population size needs to be reduced to produce such a large discrepancy. The time to the most recent common ancestor (TMRCA) for mtDNA and the NRY seem to have converged recently to a figure of 120,000–200,000 years ago (Soares et al. 2009; Poznik et al. 2013), so differences in effective population sizes between men and women, albeit important in other aspects (see below), need no longer be invoked to resolve any gap between the TMRCAs.

Y chromosome variants tend to be more localized geographically than those of mtDNA and the autosomes. The fraction of variation within human populations for Y chromosome SNPs (as opposed to variation between populations) was estimated globally at 35.5% (Seielstad et al. 1998), versus 80–85% for the autosomes and mtDNA (Barbujani et al. 1997; Romualdi et al. 2002). Note that Jorde et al. (2000) did not find this difference, since they were comparing NRY STRs to mitochondrial and autosomal SNPs; the mutation pattern in STRs, in which the same alleles are produced numerous times, runs counter to the demographic differentiation processes. A higher female than male migration rate (via patrilocality, the tendency for a wife to move into her husband's natal household) explains most of this discrepancy; polygyny may also contribute (Dupanloup et al. 2003). Luca Cavalli-Sforza, the distinguished population geneticist who was the senior author of Seielstad et al. (1998) summarized this pattern in his native Italian: *la donna è mobile.* At a local scale, Pérez-Lezaun et al. (1999) found that the patterns of mtDNA and NRY differentiation between low- and high-altitude communities in central Asia suggested a strong bottleneck in the initial colonization of the high-altitude habitat, but that subsequent female migration from the lowlands had replenished mtDNA, but not NRY, diversity. Similarly, comparison of the mtDNA and NRY allowed to conclude that social mobility among castes in India is easier for women than for men

(Wooding et al. 2004). And African Pygmy women are more likely than men to marry Bantu farmers (Verdu et al. 2009).

However, at a continental scale, the comparison of NRY and the mtDNA tips the scales in favor of male mobility. For instance, many urban mestizo populations throughout Latin America are the result of triple admixture from European, African, and Native American ancestries, although the proportions differ by sex. The contributions of Native American women and European men are larger than those of Native American men and European women (Torroni et al. 1994; Bosch et al. 2003; Mendizabal et al. 2008; Corach et al. 2010; Núñez et al. 2010; Guerra et al. 2011); among the first European colonizers of the Americas, men were much more frequent than women. Obviously, that sexual asymmetry stemmed also from the power inequality, manifest also in the fact that European admixture into African Americans is much larger for the NRY than for mtDNA (Stefflova et al. 2009; Lao et al. 2010; Battaggia et al. 2012; Torres et al. 2012). Similar situations have happened in other continents and historic times, such as in the fifteenth century rediscovery of the Canary Islands (Flores et al. 2003; Gonzalez et al. 2003), the French colonization of the Réunion Island (Berniell-Lee et al. 2008), the incredible journey of the Southeast Asian seafarers into Madagascar (Hurles et al. 2005; Tofanelli et al. 2009), or the dual colonization of Iceland with Scandinavian men and Irish women (Helgason et al. 2000a; Helgason et al. 2000b).

5.5 Male Lines

The Y chromosome is carried by men, who inherit it from their fathers. Thus, the NRY behaves as a marker for paternal lines of descent, as surnames do in many cultures. It is expected, then, that all the descendants of a founder of a surname carry copies of the same Y chromosome, modified only by mutation. In the usual time depth of surnames (about 700 years in most western European countries, although they are more recent in Japan and more ancient in China), most mutations that can be detected within the Y chromosomes of carriers of the same surname will be in STRs; given the ascertainment scheme used to find the SNPs usually typed in the NRY, the descendants of the founder of a surname are very unlikely to carry detectable SNP variability in their Y chromosomes (obviously, this is not the case if the whole NRY is sequenced; we are referring specifically to the panels of SNPs that are normally used to classify Y chromosomes into haplogroups). Conversely, STR and SNP variation can be used to sort the Y chromosomes into groups of descendants from different surname founders. After the medieval founding of surnames, Y chromosomes may have introgressed into a surname by at least three different mechanisms: false paternity, adoption, and the anomalous inheritance of the maternal surname. It may be particularly difficult to distinguish a founder's Y chromosome from a Y chromosome that introgressed into a surname early in its history (King and Jobling 2009b).

Sykes and Irven (2000) pioneered the study of Y chromosome diversity in a surname, namely Sykes, which opened the floodgates for genetic genealogy; thousands of men, particularly in the USA, have found their kin in some men carrying their same surname, with the help of companies such as Family Tree DNA (King and Jobling 2009b). Of a broader scientific interest is the analysis of an entire surname system, such as that of England (King and Jobling 2009a) or Ireland (McEvoy and Bradley 2006). In the English case, King and Jobling (2009a) found that surname frequency is driven by polyphyletism; that is, the more frequent surnames are common because they were founded numerous times: Smith is an obvious example, but the patronym surnames (those that derive from first names, such as Jones, Davies or Williams) are both likely to have been founded multiple times and to be quite frequent in the UK. King and Jobling also estimated the rate of Y chromosome introgression into a surname at about 2% per generation. The Irish case (McEvoy and Bradley 2006) was strikingly different: some frequent surnames, such as O'Sullivan and Ryan, were founded by a single ancestor, and came to be frequent in Ireland by differences in fecundity engendered by the social power and prestige associated with the name and its bearers in the past.

5.6 A Royal Mystery in Three Acts

Some particular male lineages have been investigated by their historical connections. Thus, it has been established that Thomas Jefferson fathered one of his slaves' sons (King et al. 2007). The Italian Colombo and Catalan Colom surnames have been investigated to decide whether it would be genetically possible to validate an alternate, Catalan hypothesis for Columbus' origin (Martínez-González et al. 2012). It has also been established that the current bearers of the Basarab surname are not the descendants of the most famous bearer of that surname, count Vlad Dracul (Martinez-Cruz et al. 2012). And then, a bizarre story unfolded in three acts. Enter first an antique gourd that emerged and was brought to the attention of a group of geneticists. It was pyrographed with portraits of French revolutionaries and with an inscription that stated that the gourd contained a handkerchief that had been dipped in the blood of Louis XVI of France immediately after he was guillotined. Indeed, the gourd contained a black residue with human DNA consistent with a blue-eyed male. mtDNA and NRY profiles were also obtained (Lalueza-Fox et al. 2011). In the second act, a severed, mummified head that appeared in the Parisian attic of a tax collector was anatomically shown to be compatible with Henri IV's missing head. A partial NRY STR profile could be retrieved (Charlier et al. 2013), and showed one mismatch with his seven-generation descendant, Louis XVI. The mismatch could have been caused by a mutation in these seven generations. But, in the (final?) third act, three living Bourbons enter the stage. They share a common ancestor that is intermediate between Henri IV and Louis XVI; they are also shown (Larmuseau et al. 2013) to share the same Y chromosome haplotype, which is different either from that of the severed head and that in the handkerchief. These results can be explained by forgery, contamination, or false paternity in various possible combinations.

5.7 Hacking the Y Chromosome

The fact that the NRY and the surname are inherited together opens the door to intriguing possibilities, such as the prediction of the surname from the observation of a particular NRY haplotype, which would be a welcome tool in crime investigation. Given the high rate of Y chromosome introgression into a surname, many false positives can be expected. However, Gitschier (2009) succeeded in inferring the surnames of 20 of the 30 Utah males in the CEU sample, by accessing their genotypes from the Hapmap project and their surnames from the http://www.ysearch.org/, where both STR haplotypes and surnames are stored. It can be argued that this was possible due to both the relative closed nature of the Latter-Day Saint community in Utah and to their devotion to genealogy. But Gymrek et al. (2013) systematized this search and estimated that the probability of assigning a correct surname to the STR haplotype of a European-American male is ~12%, and, that, if information on year of birth and US state is known, then the list of potential matches averages only 12 men. They demonstrated this approach by recovering Craig Venter's name from the NRY STRs in his publicly available, identified genome (they failed in two other public genomes). It should be noted that this exercise was conducted by using publicly available databases and with minimal informatic sophistication.

References

Adams SM, King TE, Bosch E, Jobling MA (2006) The case of the unreliable SNP: recurrent back-mutation of Y-chromosomal marker P25 through gene conversion. Forensic Sci Int 159:14–20

Adams SM, Bosch E, Balaresque PL et al (2008) The genetic legacy of religious diversity and intolerance: paternal lineages of Christians, Jews, and Muslims in the Iberian Peninsula. Am J Hum Genet 83:725–736

Alonso S, Flores C, Cabrera V et al (2005) The place of the Basques in the European Y-chromosome diversity landscape. Eur J Hum Genet 13:1293–1302

Arredi B, Poloni ES, Paracchini S et al (2004) A predominantly neolithic origin for Y-chromosomal DNA variation in North Africa. Am J Hum Genet 75:338–345

Athey TW (2005) Haplogroup prediction from Y-STR values using an allele- frequency approach. J Genet Geneal 1:1–7

Bachtrog D (2013) Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. Nat Rev Genet 14:113–124

Balaresque P, Bowden GR, Adams SM et al (2010) A predominantly neolithic origin for European paternal lineages. PLoS Biol 8:e1000285

Barbujani G, Magagni A, Minch E, Cavalli-Sforza LL (1997) An apportionment of human DNA diversity. Proc Natl Acad Sci USA 94:4516–4519

Batini C, Ferri G, Destro-Bisol G et al (2011) Signatures of the pre-agricultural peopling processes in sub-Saharan Africa as revealed by the phylogeography of early Y chromosome lineages. Mol Biol Evol 28:2603–2613

Battaggia C, Anagnostou P, Boschi I et al (2012) Detecting sex-biased gene flow in African-Americans through the analysis of intra- and inter-population variation at mitochondrial DNA and Y- chromosome microsatellites. Balk J Med Genet 15:7–14

Berniell-Lee G, Plaza S, Bosch E et al (2008) Admixture and sexual bias in the population settlement of La Reunion Island (Indian Ocean). Am J Phys Anthropol 136:100–107

Berniell-Lee G, Calafell F, Bosch E et al (2009) Genetic and demographic implications of the bantu expansion: insights from human paternal lineages. Mol Biol Evol 26:1581–1589

Bosch E, Calafell F, Santos FR et al (1999) Variation in short tandem repeats is deeply structured by genetic background on the human Y chromosome. Am J Hum Genet 65:1623–1638

Bosch E, Calafell F, Comas D et al (2001) High-resolution analysis of human Y-chromosome variation shows a sharp discontinuity and limited gene flow between northwestern Africa and the Iberian Peninsula. Am J Hum Genet 68:1019–1029

Bosch E, Calafell F, Rosser ZH et al (2003) High level of male-biased Scandinavian admixture in Greenlandic Inuit shown by Y-chromosomal analysis. Hum Genet 112:353–363

Bouzekri N, Taylor PG, Hammer MF, Jobling MA (1998) Hypervariable digital DNA codes for human paternal lineages: MVR-PCR at the Y-specific minisatellite, MSY1 (DYF155S1). Hum Mol Genet 7:643–653

Calafell F, Comas D, Bertranpetit J (2002) Why names. Genome Res 12:219–221

Cann R, Stoneking M, Wilson AC (1987) Mitochondrial DNA and human evolution. Nature 325:31–36

Charlier P, Olalde I, Sole N et al (2013) Genetic comparison of the head of Henri IV and the presumptive blood from Louis XVI (both kings of France). Forensic Sci Int 226:38–40. https://doi.org/10.1016/j.forsciint.2012.11.018

Corach D, Lao O, Bobillo C et al (2010) Inferring continental ancestry of argentineans from autosomal, Y-chromosomal and mitochondrial DNA. Ann Hum Genet 74:65–76

Cruciani F, La Fratta R, Santolamazza P et al (2004) Phylogeographic analysis of haplogroup E3b (E-M215) y chromosomes reveals multiple migratory events within and out of Africa. Am J Hum Genet 74:1014–1022

Cruciani F, La Fratta R, Torroni A et al (2006) Molecular dissection of the Y chromosome haplogroup E-M78 (E3b1a): a posteriori evaluation of a microsatellite-network-based approach through six new biallelic markers. Hum Mutat 27:831–832

Cruciani F, Trombetta B, Massaia A et al (2011) A revised root for the human y chromosomal phylogenetic tree: the origin of patrilineal diversity in Africa. Am J Hum Genet 88:814–818

Dorit R, Akashi H, Gilbert W (1995) Absence of polymorphism at the ZFY locus on the human Y chromosome. Science 268:1183–1185

Dupanloup I, Pereira L, Bertorelle G et al (2003) A recent shift from polygyny to monogamy in humans is suggested by the analysis of worldwide Y-chromosome diversity. J Mol Evol 57:85–97

Durbin RM, Abecasis GR, Altshuler DL et al (2010) A map of human genome variation from population-scale sequencing. Nature 467:1061–1073

Flores C, Maca-Meyer N, Perez JA et al (2003) A predominant European ancestry of paternal lineages from canary islanders. Ann Hum Genet 67:138–152

Francalacci P, Morelli L, Angius A et al (2013) Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny. Science 341:565–569

Gitschier J (2009) Inferential genotyping of Y chromosomes in latter-day saints founders and comparison to Utah samples in the HapMap project. Am J Hum Genet 84:251–258

Gonzalez AM, Brehm A, Perez JA et al (2003) Mitochondrial DNA affinities at the Atlantic fringe of Europe. Am J Phys Anthropol 120:391–404

Guerra DC, Perez CF, Izaguirre MH et al (2011) Gender differences in ancestral contribution and admixture in Venezuelan populations. Hum Biol 83:345–361

Gymrek M, McGuire AL, Golan D et al (2013) Identifying personal genomes by surname inference. Science 339:321–324

Hammer MF, Karafet TM, Redd AJ et al (2001) Hierarchical patterns of global human Y-chromosome. Diversity 18:1189–1203

Hammer MF, Behar DM, Karafet TM et al (2009) Extended Y chromosome haplotypes resolve multiple and unique lineages of the Jewish priesthood. Hum Genet 126:707–717

Helgason A, Sigureth ardottir S, Gulcher JR et al (2000a) mtDNA and the origin of the Icelanders: deciphering signals of recent population history. Am J Hum Genet 66:999–1016

Helgason A, Sigureth ardottir S, Nicholson J et al (2000b) Estimating Scandinavian and Gaelic ancestry in the male settlers of Iceland. Am J Hum Genet 67:697–717

Hurles ME, Sykes BC, Jobling MA, Forster P (2005) The dual origin of the Malagasy in island Southeast Asia and East Africa: evidence from maternal and paternal lineages. Am J Hum Genet 76:894–901

Jakubiczka S, Arnemann J, Cooke HJ et al (1989) A search for restriction fragment length polymorphism on the human Y chromosome. Hum Genet 84:86–88

Jobling MA, Tyler-Smith C (1995) Fathers and sons: the Y chromosome and. Hum Evol 11:449–456

Jorde LB, Watkins WS, Bamshad MJ et al (2000) The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. Am J Hum Genet 66:979–988

Karafet TM, Mendez FL, Meilerman MB et al (2008) New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. Genome Res 18:830–838

King TE, Jobling MA (2009a) Founders, drift, and infidelity: the relationship between Y chromosome diversity and patrilineal surnames. Mol Biol Evol 26:1093–1102

King TE, Jobling MA (2009b) What's in a name? Y chromosomes, surnames and the genetic genealogy revolution. Trends Genet 25:351–360

King TE, Bowden GR, Balaresque PL et al (2007) Thomas Jefferson's Y chromosome belongs to a rare European lineage. Am J Phys Anthropol 132:584–589

Lahn BT, Page DC (1997) Functional coherence of the human Y chromosome. Science 278:675–680

Lahn BT, Page DC (1999) Four evolutionary strata on the human X chromosome. Science 286:964–967

Lalueza-Fox C, Gigli E, Bini C et al (2011) Genetic analysis of the presumptive blood from Louis XVI, King of France. Forensic Sci Int Genet 5:459–463. https://doi.org/10.1016/j.fsigen.2010.09.007

Lao O, Vallone PM, Coble MD et al (2010) Evaluating self-declared ancestry of U.S. Americans with autosomal, Y-chromosomal and mitochondrial DNA. Hum Mutat 31:1875–1893

Larmuseau MH, Delorme P, Germain P et al (2013) Genetic genealogy reveals true Y haplogroup of House of Bourbon contradicting recent identification of the presumed remains of two French Kings. Eur J Hum Genet 22(5):681–687

Larmuseau MHD, Vanderheyden N, Van Geystelen A et al (2014) Recent radiation within Y-chromosomal Haplogroup R-M269 resulted in high Y-STR haplotype resemblance. Ann Hum Genet 78:92–103. https://doi.org/10.1111/ahg.12050

Lucotte G, Ngo NY (1985) p49F, a highly polymorphic probe that detects TaqI RFLPs on the human Y chromosome. Nucleic Acids Res 13:8285

Malaspina P, Persichetti F, Novelletto A et al (1990) The human Y chromosome shows a low level of DNA polymorphism. Ann Hum Genet 54:297–305

Martinez-Cruz B, Ioana M, Calafell F et al (2012) Y-chromosome analysis in individuals bearing the Basarab name of the first dynasty of Wallachian kings. PLoS One 7:e41803. https://doi.org/10.1371/journal.pone.0041803

Martínez-González LJ, Martínez-Espín E, Álvarez JC et al (2012) Surname and Y chromosome in southern Europe: a case study with Colom/Colombo. Eur J Hum Genet 20:211–216. https://doi.org/10.1038/ejhg.2011.162

McEvoy B, Bradley DG (2006) Y-chromosomes and the extent of patrilineal ancestry in Irish surnames. Hum Genet 119:212–219

McEvoy B, Brady C, Moore LT, Bradley DG (2006) The scale and nature of Viking settlement in Ireland from Y-chromosome admixture analysis. Eur J Hum Genet 14:1288–1294

Mendez FL, Krahn T, Schrack B et al (2013) An African American paternal lineage adds an extremely ancient root to the human Y chromosome phylogenetic tree. Am J Hum Genet 92:454–459

Mendizabal I, Sandoval K, Berniell-Lee G et al (2008) Genetic origin, admixture, and asymmetry in maternal and paternal human lineages in Cuba. BMC Evol Biol 8:213

Myres NM, Rootsi S, Lin AA et al (2010) A major Y-chromosome haplogroup R1b Holocene era founder effect in Central and Western Europe. Eur J Hum Genet 19:95–101

Núñez C, Baeta M, Sosa C et al (2010) Reconstructing the population history of Nicaragua by means of mtDNA, Y-chromosome STRs, and autosomal STR markers. Am J Phys Anthropol 143:591–600

Pérez-Lezaun A, Calafell F, Comas D et al (1999) Sex-specific migration patterns in central Asian populations revealed by the analysis of Y-chromosome STRs and mtDNA. Am J Hum Genet 65:208–219

Poznik GD, Henn BM, Yee MC et al (2013) Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. Science 341:562–565

Rocca RA, Magoon G, Reynolds DF et al (2012) Discovery of Western European R1b1a2 Y chromosome variants in 1000 genomes project data: an online community approach. PLoS One 7:e41634

Romualdi C, Balding D, Nasidze IS et al (2002) Patterns of human diversity, within and among continents, inferred from biallelic DNA polymorphisms. Genome Res 12:602–612

Rootsi S, Magri C, Kivisild T et al (2004) Phylogeography of Y-chromosome haplogroup I reveals distinct domains of prehistoric gene flow in Europe. Am J Hum Genet 75:128–137

Rootsi S, Zhivotovsky LA, Baldovic M et al (2007) A counter-clockwise northern route of the Y-chromosome haplogroup N from Southeast Asia towards Europe. Eur J Hum Genet 15:204–211

Rootsi S, Myres NM, Lin AA et al (2012) Distinguishing the co-ancestries of haplogroup G Y-chromosomes in the populations of Europe and the Caucasus. Eur J Hum Genet 20:1275–1282

Satta Y, Iwase M (2015) Genes on X and Y chromosomes. In: Evolution of the human genome I. Evolutionary studies. Springer, Tokyo

Schlecht J, Kaplan ME, Barnard K et al (2008) Machine-learning approaches for classifying haplogroup from Y chromosome STR data. PLoS Comput Biol 4:e1000093

Seielstad MT, Hebert JM, Lin AA et al (1994) Construction of human Y-chromosomal haplotypes using a new polymorphic A to G transition. Hum Mol Genet 3:2159–2161

Seielstad MT, Minch E, Cavalli-Sforza LL (1998) Genetic evidence for a higher female migration rate in humans. Nat Genet 20:278–280

Semino O, Passarino G, Oefner PJ et al (2000) The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y chromosome perpective. Science 290:1155–1159

Semino O, Magri C, Benuzzi G et al (2004) Origin, diffusion, and differentiation of Y-chromosome haplogroups E and J: inferences on the neolithization of Europe and later migratory events in the Mediterranean area. Am J Hum Genet 74:1023–1034

Skorecki K, Selig S, Blazer S et al (1997) Y chromosomes of Jewish priests. Nature 385:32

Soares P, Ermini L, Thomson N et al (2009) Correcting for purifying selection: an improved human mitochondrial molecular clock. Am J Hum Genet 84:740–759

Stefflova K, Dulik MC, Pai AA et al (2009) Evaluation of group genetic ancestry of populations from Philadelphia and Dakar in the context of sex-biased admixture in the Americas. PLoS One 4:e7842

Sykes B, Irven C (2000) Surnames and the Y chromosome. Am J Hum Genet 66:1417–1419

The Y Chromosome Consortium (2002) A nomenclature system for the tree of human Y-chromosomal binary haplogroups. Genome Res 12:339–348

Thomson R, Pritchard JK, Shen P et al (2000) Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. Proc Natl Acad Sci U S A 97:7360–7365

Tofanelli S, Bertoncini S, Castri L et al (2009) On the origins and admixture of Malagasy: new evidence from high-resolution analyses of paternal and maternal lineages. Mol Biol Evol 26:2109–2124

Torres JB, Doura MB, Keita SO, Kittles RA (2012) Y chromosome lineages in men of west African descent. PLoS One 7:e29687

Torroni A, Chen YS, Semino O et al (1994) mtDNA and Y-chromosome polymorphisms in four native American populations from southern Mexico. Am J Hum Genet 54:303–318

Trombetta B, Cruciani F, Sellitto D, Scozzari R (2011) A new topology of the human Y chromosome haplogroup E1b1 (E-P2) revealed through the use of newly characterized binary polymorphisms. PLoS One 6:e16073

Underhill PA, Jin L, Lin AA et al (1997) Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. Genome Res 7:996–1005

Underhill PA, Myres NM, Rootsi S et al (2010) Separating the post-glacial coancestry of European and Asian Y chromosomes within haplogroup R1a. Eur J Hum Genet 18:479–484

Van Geystelen A, Decorte R, Larmuseau MH (2013) Updating the Y-chromosomal phylogenetic tree for forensic applications based on whole genome SNPs. Forensic Sci Int Genet 7:573–580

Verdu P, Austerlitz F, Estoup A et al (2009) Origins and genetic diversity of pygmy hunter-gatherers from Western Central Africa. Curr Biol 19:312–318

Wei W, Ayub Q, Chen Y et al (2013) A calibrated human Y-chromosomal phylogeny based on resequencing. Genome Res 23:388–395

Wooding S, Ostler C, Prasad BV et al (2004) Directional migration in the Hindu castes: inferences from mitochondrial, autosomal and Y-chromosomal data. Hum Genet 115:221–229

Zerjal T, Xue Y, Bertorelle G et al (2003) The genetic legacy of the mongols. Am J Hum Genet 72:717–721

Zhong H, Shi H, Qi XB et al (2010) Global distribution of Y-chromosome haplogroup C reveals the prehistoric migration routes of African exodus and early settlement in East Asia. J Hum Genet 55:428–435

Chapter 6
Africa

David Comas and Francesc Calafell

Abstract The higher genetic diversity present in African human groups is the result of our African origin and the subsequent bottleneck that took place during the colonization of the rest of the continents. However, the genetic diversity within Africa is structured by geography, lifestyle, and culture. The analysis of the genetic structure within Africa might reveal the population history of human groups, including some controversial issues such as the geographic origin of our species, some population splits and admixtures among groups, and the dating of several demographic events. In addition, adaptation to broad environmental landscapes within Africa has left a genetic signal in some human groups. In the present chapter, we summarize part of the vast genetic complexity described in African populations, focusing in the genetic structure in extant groups and emphasizing the urgent need to enlarge our knowledge at continental level.

Keywords Africa · Human populations · Hunter-gatherers · Khoisan-speakers · Pygmies · Bantu languages · Genetic diversity

6.1 The African Landscape

Africa is one of the largest continents, with ~20% of the emerged land, and harbors a great environmental and cultural diversity.

The geographical landscape has modeled the movements and distribution of human groups and it is nowadays characterized by a tectonic fault that crosses the continent from north to south (the Rift Valley, from Egypt to the Mozambique coast)

D. Comas (✉)
Institut de Biologia Evolutiva (CSIC-Universitat Pompeu Fabra), Barcelona, Catalonia, Spain
e-mail: david.comas@upf.edu

F. Calafell
Institut de Biologia Evolutiva, Universitat Pompeu Fabra, Barcelona, Catalonia, Spain
e-mail: francesc.calafell@upf.edu

N. Saitou (ed.), *Evolution of the Human Genome II*, Evolutionary Studies,
https://doi.org/10.1007/978-4-431-56904-6_6

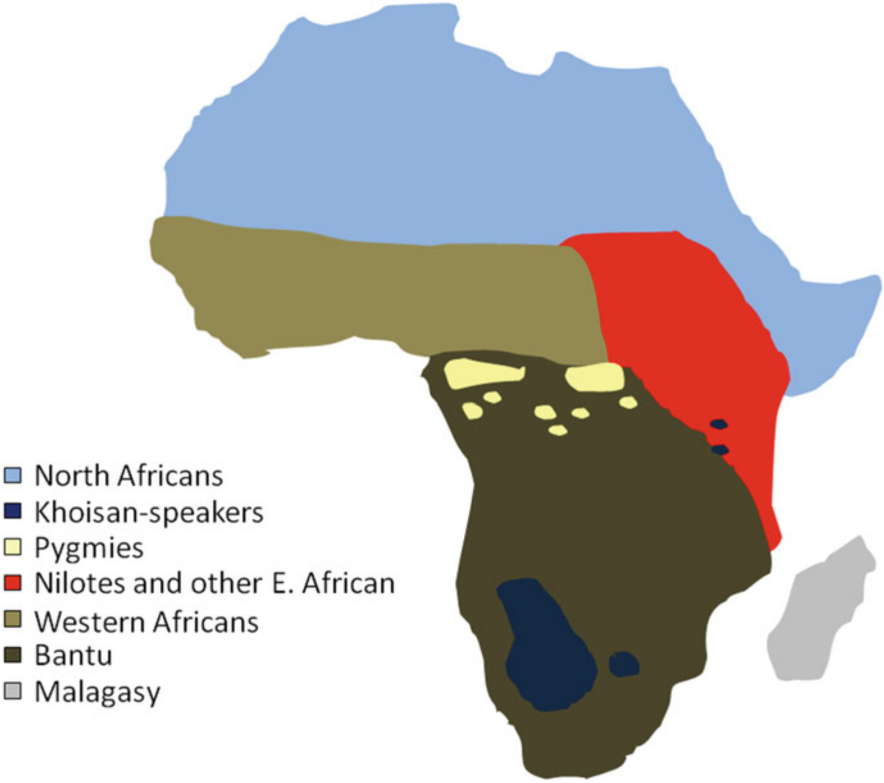


Fig. 6.1 Geographic distribution of human African groups following the classification proposed by Hiernaux (1974)

that separates the easternmost part of the continent; the deserts of Sahara (an extensive region dividing the continent in a north-south pattern), Namib and Kalahari (both in southern Africa); and the intertropical area occupied by the dense rainforest.

The cultural landscape includes around one-third of the extant world languages spoken in continental Africa; they are grouped, by the most inclusive classification, in four linguistic families: Afro-Asiatic, Nilo-Saharan, Niger-Kordofanian, and Khoisan (Greenberg 1963), although many linguists dispute the integrity of the Nilo-Saharan, Niger-Kordofanian, and Khoisan families (Sands 2009). Diverse traditional subsistence models (hunter-gatherers, farmers and pastoralists) are found throughout the continent.

These geographical and cultural landscapes, together with anthropological traits, were the bases that Hiernaux (1974) used to provide a rough description of human African groups despite the limitations of this classification, which Hiernaux himself acknowledged (Fig. 6.1). As a first distinction, North Africans, separated from the rest of the African groups by the Sahara desert, were found to be akin to Middle Eastern and European populations. As for sub-Saharan groups, Khoisan-speakers (including southern African groups as well as Hadza and Sandawe from Tanzania) and Pygmies (in the equatorial belt) were characterized by their distinctive physical traits as well as by their hunter-gatherer subsistence lifestyle. Eastern African groups, with diverse languages, and most with a pastoral lifestyle, were defined according to their elongated morphological traits explained as a result of adaptation to dry heat. The Western and Bantu-speaking groups formed a diverse group of peoples speaking Niger-Kordofanian languages spread all over Africa, being the spread of Bantu one of the largest past human cultural expansions. Finally, Hiernaux (1974) described the Malagasy groups as a blend of Indonesian and African traits as a result of the admixture of South-East Asians and continental Africans and influences from the Eastern coastal Arabic trade.

6.2 African Genetic Diversity in a Global Scale

Despite the extensive genetic diversity found in Africa, until recently most genetic studies were limited and mainly focused on the comparison of the diversity of the continent with non-African peoples to prove the African origin of humankind, ignoring most of the diversity within the continent. Nonetheless, the correlation between genetic diversity within Africa and the classification of human groups according to their lifestyle or language is remarkable and it helps us to understand the population history of African groups and to refine demographic and adaptation aspects provided by other disciplines.

All genetic studies, from classical markers to complete genomes, show greater diversity in African populations (Cavalli-Sforza et al. 1994; Tishkoff et al. 2009; Henn et al. 2011), which has been explained by our African origin and the larger population sizes of African groups. Genetic evidence places the cradle of humankind in Africa around 200,000 years ago (Ingman et al. 2000; Francalacci et al. 2013), which has allowed African populations to accumulate changes in their genomes for a larger period of time. In addition, the larger effective population size of these groups has allowed maintaining more variants in African populations. The demographic bottlenecks resulting of the out-of-Africa expansion of humans (Henn et al. 2012b) might have limited the genetic diversity elsewhere. As a consequence of these facts, genetic variants and haplotypes found in other continents are a subsample of those found in Africa (Tishkoff et al. 1996, 1998; Melé et al. 2012), and the deepest branches in almost all human gene genealogies are occupied by African populations (Cann et al. 1987; Vigilant et al. 1991; Cavalli-Sforza et al. 1994; Ingman et al. 2000; Li et al. 2008; Tishkoff et al. 2009; Keinan et al. 2009). It is worth noting that in these studies, African samples refer to the sub-Saharan part of the continent, being North African groups considered as part of the out-of-Africa group.

One remarkable contribution from genome studies to the field of human evolution in last years has been the detection of admixture of ancient archaic hominins with human groups. The detected admixture of Neanderthals with out-of-Africa populations, including North Africa, reveals that around 2–3% of the genomes of extant non-sub-Saharan groups might come from archaic hominins (Green et al. 2010; Sánchez-Quinto et al. 2012). In addition, some southeast Asians also show admixture with another archaic hominin, the Denisovans (Reich et al. 2010). These studies did not detect admixture in extant Sub-Saharan populations neither with Neanderthals nor with Denisovans, but unusual patterns of linkage disequilibrium and deep divergence have suggested introgression in some sub-Saharan groups from unknown ancient hominins (Hammer et al. 2011), suggesting a high promiscuity of our ancestors with other hominins at a global scale.

6.3 Genetic Diversity Within Africa

As mentioned above, few genetic diversity analyses up to now have been performed at intracontinental level. The monumental compilation of classical genetic markers by Cavalli and coauthors (1994) showed genetic structure at continental level, with sub-Saharan populations differentiated from those of the North, and genetic differentiation between Eastern and Western population groups.

The genetic structure within Africa has been also assessed through the analysis of uniparental markers (i.e., mitochondrial DNA and Y chromosome). As a continent-wide synthesis, Salas et al. (2002) analyzed the mitochondrial DNA (mtDNA) control region in several African groups, which showed a clear geographic structure of their genetic diversity. For instance, some hunter-gatherer groups (such as Pygmies or Southern Khoisan-speakers) carried specific haplogroups, and Eastern and Western African populations exhibited different lineages, pointing to an intracontinental genetic structure modeled by diverse demographic processes. More recently, Behar et al. (2008), using the whole mtDNA sequence, reached similar conclusions. Unfortunately, no comparable Y-chromosome analysis exists at a continental level. However, general and local surveys performed with African Y-chromosome diversity (Underhill et al. 2000; Macaulay et al. 2005; Berniell-Lee et al. 2009; Cruciani et al. 2011; Scozzari et al. 2012) point to specific ancient lineages in hunter-gatherers as well as to some geographic structure of lineages, although the genetic structure of the Y chromosome is not so pronounced as what has been described for the mtDNA. This difference has been attributed to the homogenizing effect of the male-driven Bantu expansion in paternal lineages (Berniell-Lee et al. 2009).

The landmark compilation of hundreds of autosomal markers in more than 2500 individuals from 120 populations by Tishkoff et al. (2009) has provided the largest geographic coverage of autosomal genetic diversity in Africa. Their conclusions confirm the genetic structure in Africa, suggesting a probable ancestral connection between hunter-gatherers groups, and a complex diverse genetic pattern in Eastern Africa, which contrasts with the homogeneous genetic landscape shown by Western and Bantu-speaking groups. However, the genetic link found between Pygmies and Khoisan-speakers has been challenged: an analysis of more than half a million autosomal markers performed in a more limited sample set (Jakobsson et al. 2008) pointed to this hunter-gatherer link, whereas Henn et al. (2011) did not find evidence of such genetic connection. These discrepancies point to the lack of adequate studies on the internal diversity of the African continent, both by the extent of the genome analyzed and by the populations covered, to resolve the demographic history of the extant populations. The advent of new generation sequencing techniques will provide a refined view of the population history of the whole continent, as shown already by Lachance et al. (2012), who sequenced the genomes of just a few hunter-gatherer individuals.

6.4 Local Genetic Diversity Within Africa

In the following section, we will summarize the genetic diversity of African population groups as defined by Hiernaux (1974), who used the geographical, cultural, and anthropological traits described above. Although this classification might have its caveats and should be regarded as fluid, it is a simple way to assess the population questions and descriptions that have been posed by genetic studies.

6.4.1 North Africans, the Misfit Within the Continent

Physical anthropology, as exemplified by the work of Hiernaux (1974), distinguishes North Africans from the rest of the continental groups, and links them to Western Eurasians rather to sub-Saharan Africans, although the population history of North Africans has not been totally independent of the rest of the continent. The peopling of North Africa has been limited by the geographic barriers of the Mediterranean Sea to the north and the Sahara Desert to the south, which have channeled human movements in an East-West direction. Recent archaeological records point to an ancient occupation of North Africa earlier than 120,000 years ago (Barton et al. 2009; Garcea 2010) with several cultural strata until the arrival of the Neolithic and historical migrations. Besides the complex cultural replacements in the area, the Berber-speaking groups have been identified with an ample consensus as the autochthonous people in the region, in contrast with subsequent migrations from neighboring areas (Bosch et al. 2001; Henn et al. 2012a).

Classical genetic markers and uniparental genomes have shown a closer relationship of North Africans to Eurasians than to sub-Saharans (Cavalli-Sforza et al. 1994; Bosch et al. 1997, 2001; Plaza et al. 2003; Fadhlaoui-Zid et al. 2011, 2013). In addition, these studies have not found genetic differences between Berber- and Arab-speaking groups in NW Africa, suggesting that the Arabization of North Africa that took place between the seventh and eleventh centuries was mainly a cultural rather than a demographic process (although, as discussed below, some Middle Eastern contribution is indeed found in N. Africa). Uniparental haplogroups found in North Africa are those found in Eurasia with sub-Saharan contributions (Plaza et al. 2003; Fadhlaoui-Zid et al. 2011, 2013). However, some specific haplogroups have been described in North Africa such as mtDNA haplogroup U6 or Y-chromosome E1b1b1b-M81 (Rando et al. 1998; Bosch et al. 2001), that might have been remnants of the first North African settlers or specific developments in the region.

Genome-wide analyses in North African populations have shown a complex demographic pattern and a clear difference with sub-Saharan groups (Li et al. 2008; Tishkoff et al. 2009; Henn et al. 2012a). This complex scenario is characterized by an autochthonous component in extant North African populations that might have been introduced in Africa from the Middle East in pre-Holocene times (Henn et al. 2012a). This recent autochthonous component is further evidence for a discontinuity between the first peoples in North Africa more than 120,000 years ago and the extant populations. In addition, this component seems to be more frequent in isolated Berber groups, in agreement with the idea of Berber-speakers being the autochthonous Holocene peoples of the region. Other genetic components are seen in North Africa, such as a sub-Saharan component that might have been introduced in North Africa in recent times agreeing with the sub-Saharan slave trade; a Middle Eastern component that might be linked to the Neolithization and/or Arabization processes; and a European component as a result of gene flow across the Mediterranean (Henn et al. 2012a). In addition, a North African contribution to southern Europe, specifically in the Iberian Peninsula, has been described as one of the putative processes that might have yield higher genetic diversity in southern Europe (Botigué et al. 2013).

It is noteworthy that North African genetic diversity has been usually ignored in global studies, where usually the Human Genome Diversity Panel (HGDP) (Li et al. 2008) Mozabite (a Berber population from Algeria) sample is included. It is even worrying that in most of the analyses performed at a global scale with the HGDP samples, Mozabites are often pooled with Near Eastern samples since the populations of both regions speak primarily Afro-Asiatic languages, regardless of their population structure.

6.4.2 Khoisan-Speakers, the Deepest Branch in the Humankind Tree

The Khoisan language family comprises click languages spoken in Southwestern Africa and by two small groups in Tanzania, the Hadza and the Sandawe. A recent survey of the phonemic diversity in human populations (Atkinson 2011) has proposed the Khoisan family as the first branch of the human linguistic tree, suggesting a South African origin for our species (but see the reply by linguists Berdicevskis and Piperski in an e-letter to *Science*). The analysis of classical genetic markers (Cavalli-Sforza et al. 1994) showed a differentiation of Southern African Khoisan-speakers in the third principal component. More than two thirds of the mtDNA lineages found in southern Khoisan-speakers belong to haplogroups L0d and L0k, which have been described as one of the deepest roots of the human mtDNA tree, although their presence in southern populations has been explained as a result of an ancient migration from Eastern Africa (Behar et al. 2008). Sublineages of L0d have also been found at low frequencies in Eastern Africa (Gonder et al. 2007), suggesting an ancient link between Eastern and Southern Africa. Regarding Y-chromosome haplogroups, lineages within the deepest branches of the tree (A and B haplogroups) are found in Southern Khoisan-speakers (A2, A3b1, and B2b). The presence of a phylogenetically close lineage (A3b2) in Eastern Africa contributes evidence for a South-East African connection. In fact, the presence of B2b lineages in Southern and Eastern Khoisan-speakers as well as in Pygmies (Batini et al. 2011a) suggests an ancient link between all hunter-gatherer groups. Nonetheless, the deepest branches of the Y-chromosome tree (A1) are not found in extant Southern Khoisan-speaking groups (Cruciani et al. 2011), which questions the origins of human paternal lineages. In addition, recent data point to an extreme ancient root of the Y chromosome (named A00) that is not found in southern hunter-gatherers (Mendez et al. 2013), although the methods used to estimate the time of divergence have been criticized (Elhaik et al. 2014).

The higher genetic diversity in Eastern African groups and the genetic link between hunter-gatherer groups suggested by the autosomal analysis performed by Tishkoff et al. (2009) has been challenged by Henn et al. (2011) and Schlebusch et al. (2012) by the inclusion of more Khoisan-speaking groups in the recent analyses. Henn et al. (2011) point to an origin of modern humans in South Africa rather than in East Africa, although the deep analysis of several Khoisan groups shows a more complex pattern of admixture and genetic structure that challenges the exact location of the origin of our species (Pickrell et al. 2012; Schlebusch et al. 2012).

Despite not being included in the 1000 genomes project, some studies have produced the whole genome sequencing of Khoisan-speaking individuals (Schuster et al. 2010; Lachance et al. 2012). These studies have confirmed the extreme diversity found in Khoisan-speaking individuals, providing a high percentage of variation not found in any other population analyzed up to date.

6.4.3 The Pygmies, the Hunter-Gatherers of the African Equatorial Forest

Pygmies are hunter-gatherer groups scattered in the dense forest of Central Africa. It has been suggested that extant African hunter-gatherers (Pygmies and Khoisan-speakers) might have had a larger distribution range before other population groups, such as Bantu-speakers, might have had displaced them to their extant locations. Two main groups of Pygmies exist: Eastern (located in the Democratic Republic of Congo and Rwanda) and Western (Cameroon, Republic of Congo, Gabon, and the Central African Republic). In addition to these, other groups, often called "pygmoids," might be the result of admixture between Pygmy and neighboring populations (Hiernaux 1974; Cavalli-Sforza 1986). Despite this classification, the definition of the groups is not simple since some have had differential contacts with neighboring populations, and in contrast with Khoisan-speaking groups, Pygmies do not have a language in common, since they speak languages borrowed from their neighbors (Niger-Kordofanian, Nilo-Saharan) that might have replaced their ancestral languages (Ruhlen 1994). One characteristic common to all African Pygmies is their short height; they fall in the extreme of the stature distribution in humans (the genes associated to this characteristic trait are discussed in the following section).

As discussed above, genetic studies have suggested Khoisan-speakers and Pygmies as the first split of the human tree with an ancient connection between African hunter-gatherers groups (Jakobsson et al. 2008; Li et al. 2008; Tishkoff et al. 2009; Batini et al. 2011a), although this idea has been challenged in favor of a first split of Khoisan-speakers rather than Pygmy groups (Henn et al. 2011; Schlebusch et al. 2012).

It has been extensively debated whether Eastern and Western Pygmy groups have common or independent origins. The studies performed with uniparental lineages have shown different distribution of haplogroups in Eastern and Western Pygmies. As mentioned above, Pygmy groups share with Khoisan-speaking groups the Y-chromosome B2b lineage (Wood et al. 2005; Berniell-Lee et al. 2009; Batini et al. 2011a), which is very frequent in Pygmies, suggesting a possible common root among African hunter-gatherers, although this can also be interpreted as a recent contact between both groups. In contrast, substantial differences are found in mtDNA haplogroups suggesting an ancient maternal separation between Western (characterized by L1c lineages) and Eastern (characterized by L0a, L2a, and L5 lineages) Pygmies (Destro-Bisol et al. 2004; Batini et al. 2007; Quintana-Murci et al. 2008; Batini et al. 2011b). These discrepancies shown by the analyses of uniparental markers seem to be resolved by the studies of autosomal markers that have pointed a common origin and a split of both Pygmy groups around 25,000 years ago (Patin et al. 2009; Batini et al. 2011b), after the divergence of Pygmies and non-Pygmies around 60,000 years ago (Patin et al. 2009; Verdu et al. 2009; Batini et al. 2011b). These studies have found little gene flow between both groups of Pygmies after the split, in contrast with extensive gene flow from/to neighboring non-Pygmy populations.

6.4.4 East African Groups, Source and Sink of Genetic Diversity

East African groups have been characterized by their high genetic diversity within and among populations, which has been interpreted either as the origin of our species or as a melting pot of migrations from different sources. The presence of the oldest fossil representatives of our species in East Africa has been interpreted as evidence for the origin of humans in this African area (Stringer and Andrews 1988), although the difficulty of finding fossil remains in other geographical areas due to the composition of the soil might have biased this conclusion. Nonetheless, most genetic studies performed with worldwide samples have shown a pattern of decreasing genetic diversity from East Africa, pointing to a clear evidence of an early expansion out of East Africa that might have taken place through the southern end of the Red Sea (Jakobsson et al. 2008; Melé et al. 2012). The origin of humans in East Africa and the subsequent expansion out of Africa fit with the data provided by the analysis of the mtDNA that shows a high diversity of lineages, including some of the oldest branches of the mtDNA tree (L0d, L0f, L5) (Gonder et al. 2007; Behar et al. 2008) as well as deep branches of the diversity found out of Africa (M1, N1) (Quintana-Murci et al. 1999).

However, all the genetic studies that focus on East Africa have identified external gene flow into this region, mostly from the West and the North. The detailed compilation of autosomal markers by Tishkoff et al. (2009) revealed a complex scenario in East Africa with a large number of genetic components. Different linguistic groups in East Africa, including Afro-Asiatic, Nilo-Saharan, and Niger-Kordofanian, share a common ancestral component at high frequencies regardless their linguistic affiliation that is also found at lower frequencies in northern populations in Sudan or even in North Africans. Besides this component, a common component found all over Africa associated to Niger-Kordofanian speakers is also present in East Africa, although, as expected, is more frequent in the local Niger-Kordofanian speakers.

Due to the high diversity found in this region, international genome projects such as HapMap or the 1000 genomes have included samples from East Africa in order to provide a good representation of the human diversity. HapMap III has included Nilo-Saharan Maasai and Bantu Luhya in their analyses, whereas the 1000 genomes have also included a Luhya sample.

6.4.5 Western and Bantu-Speakers, the Major Component in Africa

The major linguistic group in Africa in the classical Greenberg (1963) classification, the Niger-Kordofanian family, which includes the Bantu branch, is spread all over sub-Saharan Africa and includes populations from Western, Central, Eastern, and Southern Africa. The distribution of these languages has been explained by the expansion of the Bantu branch that took place around 5000 years ago from the area of the extant border of Nigeria and Cameroon to the south of the continent in a fast manner, divided in two main streams (East and West) and reaching the southern coast in less than 3000 years (Phillipson 1993; Newman 1995; Vansina 1995). This expansion might have been related to the expansion of agriculture and iron technology, and might have been one of the largest cultural and demographic human movements (Diamond 1997). However, some scholars have noted the simplified models and conclusions of this expansion, and have highlighted the relevance of local migration processes in the extant diversity found in sub-Saharan Africa (Lwanga-Lunyiigo 1976; Ehret 2001; Schoenbrun 2001).

Most analyses have shown that Western and Bantu-speaking groups are genetically homogeneous and this has been explained by their common and recent origin. Uniparental markers have revealed a similar frequency composition of major haplogroups (E1b1a and B2a for the Y chromosome; and L0a, L2a, L2b, and L3e for the mtDNA) as a result of the Bantu expansion (Pereira et al. 2001; Cruciani et al. 2002; Salas et al. 2002; Plaza et al. 2004; Wood et al. 2005; Berniell-Lee et al. 2009; Montano et al. 2011). However, these studies have found some local differences, not only in their specific haplogroup composition (Salas et al. 2002) but also in the amount of assimilation and admixture with hunter-gatherer groups, suggesting a sex-biased gene flow due to social rules (Destro-Bisol et al. 2004). Autosomal polymorphisms (Tishkoff et al. 2009) have shown a striking homogeneity of Niger-Kordofanian groups with small influences from other African groups, being this major component highly represented in African Americans as a result of the Atlantic slave trade from the sixteenth to the nineteenth centuries, although other components have also been described, such as a Southeastern component in Mozambique (Sikora et al. 2011). The presence of this major African component has fostered the inclusion of some Western African populations in international projects: since the beginning, the Yoruba from Nigeria were included in the HapMap project; and several Western groups (from Nigeria, Sierra Leone and Gambia) besides Yoruba have also been included in the 1000 genomes. However, it is noteworthy that despite the spread of Bantu languages in Africa, only the Eastern Bantu Luhya have been included in these international genome efforts.

6.4.6 Madagascar, an Island Between Two Distant Continents

The peopling of Madagascar, the largest African island in the southeastern coast of the continent, is an outlier within the African landscape. The Malagasy language belongs to the Austronesian family spoken in southeast Asia. Linguistic data point to an initial peopling of the island from southeast Asia through a long range maritime expansion in historical times. Posterior African gene flow from the continent was added to the initial colonizers and subsequent migrations from the African continent and the Arab slave trade across the Indic Ocean modeled the extant Malagasy population. Uniparental markers agree with the admixture in the island of Southeast Asian and African haplogroups, and point to a population sexual bias, being Southeast Asian maternal lineages and African paternal lineages more prevalent in the Malagasy population (Tofanelli et al. 2009). Recent autosomal data on some

Malagasy groups have provided fresh evidence for the admixture of these two distant (Austronesian and Bantu) components (Pierron et al. 2014).

6.5 Some Adaptive Variants in African Groups

In the previous sections we have assessed the genetic distribution and diversity of African populations at a genome level, which provides good evidence about the demographic history of African groups. However, the African environmental diversity has forced humans to adapt and several genetic variants have been shown to have been selected for in some African populations. These variants are difficult to pinpoint for several reasons: they are related to specific genes or regulatory regions rather than to the whole genome, the effect of some of them is minor, the adaptive effect is unknown, and it can be masked by demographic effects.

One of the paradigms of adaptation in African groups is the resistance to malaria, caused by *Plasmodium* and having the *Anopheles* mosquito as a vector of transmission. The cohabitation of some human groups with the parasite has driven the selection of some genetic variants in order to combat the disease. The most well-known variant is hemoglobin S, an amino acid change (valine by glutamic acid at position 6 of the hemoglobin beta-chain) that causes sickle-cell disease in homozygotes but provides resistance to the disease in heterozygous individuals (Luzzatto 2012). Other variants that have been shown to provide to resistance to malaria are those in the *G6PDH* gene, and also in the *DARC* gene, which codes for the Duffy blood group (Hamblin and Di Rienzo 2000; Sabeti et al. 2002). Other variants related to protection against infectious diseases in African groups have been detected such as those related to VIH (Heeney et al. 2006).

Another interesting phenotype that has been associated to adaptive variants in African groups is the persistence of lactase in adulthood, which allows to properly digest lactose. In human groups, some regulatory variants of the lactase gene (*LCT*) have been selected since they allowed adults to digest lactose from milk (Bersaglieri et al. 2004). Some pastoralist populations in Eastern Africa have shown the presence of some such variants in the *LCT* gene (Tishkoff et al. 2007; Ranciaro et al. 2014), although these variants are different to those found in Eurasia, suggesting convergent evolution (Enattah et al. 2007, 2008).

Other genetic adaptations are more difficult to recognize since the phenotype under selection is much more complex and determined by a large number of genetic variants. This might be the case of the African Pygmy phenotype characterized by short stature plus other physical features such as peppercorn hair and lighter skin color compared to other African groups. The putative adaptation of the Pygmy phenotype is controversial (Perry and Dominy 2009) and several hypothesis have been postulated to explain its advantages, such as better thermoregulatory adaptation to the rainforest (Cavalli-Sforza 1986), adaptation to food limitations (Hart and Hart 1986), mobility in the dense equatorial forest (Diamond 1991) or adaptation for an earlier reproductive age as a result of shortened lifespan (Migliano et al. 2007).

Several studies have found genetic variants that might be related to the Pygmy phenotype (Lachance et al. 2012; Mendizabal et al. 2012; Jarvis et al. 2012) although there are discrepancies on the genetic pathways that might have been pivotal for the adaptation of the phenotype.

Other hunter-gatherer groups besides Pygmies have been analyzed in order to detect genetic signals of adaptations. Schuster et al. (2010) described some genetic variants related to bone mineral density and toxic metabolizers that might be related to longevity in Southern Khoisan-speakers; and Lachance et al. (2012) pointed to genes involved in immunity, metabolisms, olfactory and taste perception, reproduction, and wound healing that might be related to adaptation in Pygmies and Eastern Khoisan-speakers.

Another complex phenotype that might have been subject of adaptation is the genetic adaptation to high altitudes, which has been investigated in the Ethiopian highlands (Scheinfeldt and Tishkoff 2012) and where several genes that might be involved in the adaptation to altitude were found. However, these genes are different from the genetic variants described in non-African highland populations such as Tibet or Andes (Bigham et al. 2010), suggesting a process of convergent adaptation to high altitudes, although some of the genes found in different populations belong to the same genetic pathway (HIF-1).

In this brief review, we have seen how the vastness of Africa reflects both on the complexity of the demographic history of the human populations inhabiting the continent, as well as in the particular adaptations that many of those populations have needed to survive in their environments. Still, we have glimpsed a tiny fraction of the diversity in both accounts, and it is expected that new research will unveil additional layers of complexity.

References

Atkinson QD (2011) Phonemic diversity supports a serial founder effect model of language expansion from Africa. Science 332:346–349

Barton RNE, Bouzouggar A, Collcutt SN et al (2009) OSL dating of the Aterian levels at Dar es-Soltan I (Rabat, Morocco) and implications for the dispersal of modern Homo sapiens. Quat Sci Rev 28:1914–19431

Batini C, Coia V, Battaggia C et al (2007) Phylogeography of the human mitochondrial L1c haplogroup: genetic signatures of the prehistory of Central Africa. Mol Phylogenet Evol 43:635–644

Batini C, Ferri G, Destro-Bisol G et al (2011a) Signatures of the pre-agricultural peopling processes in sub-Saharan Africa as revealed by the phylogeography of early Y chromosome lineages. Mol Biol Evol 28:2603–2613

Batini C, Lopes J, Behar DM et al (2011b) Insights into the demographic history of African pygmies from complete mitochondrial genomes. Mol Biol Evol 28:1099–1110

Behar DM, Villems R, Soodyall H et al (2008) The dawn of human matrilineal diversity. Am J Hum Genet 82:1130–1140

Berniell-Lee G, Calafell F, Bosch E et al (2009) Genetic and demographic implications of the bantu expansion: insights from human paternal lineages. Mol Biol Evol 26:1581–1589

Bersaglieri T, Sabeti PC, Patterson N et al (2004) Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet 74:1111–1120

Bigham A, Bauchet M, Pinto D et al (2010) Identifying signatures of natural selection in Tibetan and Andean populations using dense genome scan data. PLoS Genet 6:e1001116. https://doi.org/10.1371/journal.pgen.1001116

Bosch E, Calafell F, Pérez-Lezaun A et al (1997) Population history of North Africa: evidence from classical genetic markers. Hum Biol 69:295–311

Bosch E, Calafell F, Comas D et al (2001) High-resolution analysis of human Y-chromosome variation shows a sharp discontinuity and limited gene flow between northwestern Africa and the Iberian Peninsula. Am J Hum Genet 68:1019–1029

Botigué LR, Henn BM, Gravel S et al (2013) Gene flow from North Africa contributes to differential human genetic diversity in southern Europe. Proc Natl Acad Sci USA 110:11791–11796

Cann R, Stoneking M, Wilson AC (1987) Mitochondrial DNA and human evolution. Nature 325:31–36

Cavalli-Sforza LL (1986) African pygmies. Academic Press, Orlando

Cavalli-Sforza LL, Menozzi P, Piazza A (1994) History and geography of human genes. Princeton University Press, Princeton

Cruciani F, Santolamazza P, Shen P et al (2002) A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. Am J Hum Genet 70:1197–1214

Cruciani F, Trombetta B, Massaia A et al (2011) A revised root for the human y chromosomal phylogenetic tree: the origin of patrilineal diversity in Africa. Am J Hum Genet 88:814–818

Destro-Bisol G, Donati F, Coia V et al (2004) Variation of female and male lineages in sub-Saharan populations: the importance of sociocultural factors. Mol Biol Evol 21:1673–1682

Diamond JD (1991) Anthropology. Why are pygmies small? Nature 354:111–112

Diamond JD (1997) Guns, germs, and steel: the fates of human societies. W.W.Norton & Company, New York

Ehret C (2001) Bantu expansions: re-envisioning a central problem of early African history. Int J Afr Hist Stud 34:5–41

Elhaik E, Tatarinova TV, Klyosov AA, Graur D (2014) The "extremely ancient" chromosome that isn't: a forensic bioinformatic investigation of Albert Perry's X-degenerate portion of the Y chromosome. Eur J Hum Genet 22:1111. https://doi.org/10.1038/ejhg.2013.303

Enattah NS, Trudeau A, Pimenoff V et al (2007) Evidence of still-ongoing convergence evolution of the lactase persistence T-13910 alleles in humans. Am J Hum Genet 81:615–625. https://doi.org/10.1086/520705

Enattah NS, Jensen TGK, Nielsen M 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. https://doi.org/10.1016/j.ajhg.2007.09.012

Fadhlaoui-Zid K, Rodríguez-Botigué L, Naoui N et al (2011) Mitochondrial DNA structure in North Africa reveals a genetic discontinuity in the Nile Valley. Am J Phys Anthropol 145:107–117. https://doi.org/10.1002/ajpa.21472

Fadhlaoui-Zid K, Haber M, Martínez-Cruz B et al (2013) Genome-wide and paternal diversity reveal a recent origin of human populations in North Africa. PLoS One 8:e80293. https://doi.org/10.1371/journal.pone.0080293

Francalacci P, Morelli L, Angius A et al (2013) Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny. Science 341:565–569

Garcea EAA (2010) The spread of Aterian peoples in North Africa. Oxbow Books, Oxford

Gonder MK, Mortensen HM, Reed FA et al (2007) Whole-mtDNA genome sequence analysis of ancient African lineages. Mol Biol Evol 24:757–768

Green RE, Krause J, Briggs AW et al (2010) A draft sequence of the Neandertal genome. Science 328:710–722

Greenberg J (1963) The languages of Africa. Indiana University Press, Bloomington

Hamblin MT, Di Rienzo A (2000) Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. Am J Hum Genet 66:1669–1679. https://doi.org/10.1086/302879

Hammer MF, Woerner AE, Mendez FL et al (2011) Genetic evidence for archaic admixture in Africa. Proc Natl Acad Sci USA 108:15123–15128

Hart TB, Hart JA (1986) The ecological basis of hunter-gatherer subsistence in African rain forest. The Mbuti of eastern Zaire. Hum Ecol 14:29–55

Heeney JL, Dalgleish AG, Weiss RA (2006) Origins of HIV and the evolution of resistance to AIDS. Science 313:462–466. https://doi.org/10.1126/science.1123016

Henn BM, Gignoux CR, Jobin M et al (2011) Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. Proc Natl Acad Sci USA 108:5154–5162

Henn BM, Botigue LR, Gravel S et al (2012a) Genomic ancestry of north Africans supports back-to-Africa migrations. PLoS Genet 8:e1002397

Henn BM, Cavalli-Sforza LL, Feldman MW (2012b) The great human expansion. Proc Natl Acad Sci USA 109:17758–17764

Hiernaux J (1974) The peoples of Africa. Weidenfeld & Nicholson, London

Ingman M, Kaessmann H, Gyllensten U, Pääbo S (2000) Mitochondrial genome variation and the origin of modern humans. Nature 408:708–713

Jakobsson M, Scholz SW, Scheet P et al (2008) Genotype, haplotype and copy-number variation in worldwide human populations. Nature 451:998–1003

Jarvis JP, Scheinfeldt LB, Soi S et al (2012) Patterns of ancestry, signatures of natural selection, and genetic association with stature in Western African pygmies. PLoS Genet 8:e1002641. https://doi.org/10.1371/journal.pgen.1002641

Keinan A, Mullikin JC, Patterson N, Reich D (2009) Accelerated genetic drift on chromosome X during the human dispersal out of Africa. Nat Genet 41:66–70. https://doi.org/10.1038/ng.303

Lachance J, Vernot B, Elbers CC et al (2012) Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. Cell 150:457–469

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

Luzzatto L (2012) Sickle cell anaemia and malaria. Mediterr J Hematol Infect Dis 4:e2012065. https://doi.org/10.4084/MJHID.2012.065

Lwanga-Lunyiigo S (1976) The Bantu problem reconsidered. Curr Anthropol 17:282–286

Macaulay V, Hill C, Achilli A et al (2005) Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. Science 308:1034–1036

Melé M, Javed A, Pybus M et al (2012) Recombination gives a new insight in the effective population size and the history of the old world human populations. Mol Biol Evol 29:25–30. https://doi.org/10.1093/molbev/msr213

Mendez FL, Krahn T, Schrack B et al (2013) An African American paternal lineage adds an extremely ancient root to the human Y chromosome phylogenetic tree. Am J Hum Genet 92:454–459

Mendizabal I, Marigorta UM, Lao O, Comas D (2012) Adaptive evolution of loci covarying with the human African Pygmy phenotype. Hum Genet 131:1305–1317

Migliano AB, Vinicius L, Lahr MM (2007) Life history trade-offs explain the evolution of human pygmies. Proc Natl Acad Sci USA 104:20216–20219

Montano V, Ferri G, Marcari V et al (2011) The Bantu expansion revisited: a new analysis of Y chromosome variation in Central Western Africa. Mol Ecol 20:2693–2708

Newman J (1995) The peopling of Africa: a geographic interpretation. Yale University Press, New Haven

Patin E, Laval G, Barreiro LB et al (2009) Inferring the demographic history of African farmers and pygmy hunter-gatherers using a multilocus resequencing data set. PLoS Genet 5:e1000448

Pereira L, Macaulay V, Torroni A et al (2001) Prehistoric and historic traces in the mtDNA of Mozambique: insights into the Bantu expansions and the slave trade. Ann Hum Genet 65:439–458

Perry GH, Dominy NJ (2009) Evolution of the human pygmy phenotype. Trends Ecol Evol 24:218–225

Phillipson DW (1993) African archaeology. Cambridge University Press, Cambridge

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

Pierron D, Razafindrazaka H, Pagani L et al (2014) Genome-wide evidence of Austronesian-Bantu admixture and cultural reversion in a hunter-gatherer group of Madagascar. Proc Natl Acad Sci USA 111:936–941. https://doi.org/10.1073/pnas.1321860111

Plaza S, Calafell F, Helal A et al (2003) Joining the pillars of Hercules: mtDNA sequences show multidirectional gene flow in the western Mediterranean. Ann Hum Genet 67:312–328

Plaza S, Salas A, Calafell F et al (2004) Insights into the western Bantu dispersal: mtDNA lineage analysis in Angola. Hum Genet 115:439–447

Quintana-Murci L, Semino O, Bandelt HJ et al (1999) Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. Nat Genet 23:437–441. https://doi.org/10.1038/70550

Quintana-Murci L, Quach H, Harmant C et al (2008) Maternal traces of deep common ancestry and asymmetric gene flow between Pygmy hunter-gatherers and Bantu-speaking farmers. Proc Natl Acad Sci USA 105:1596–1601

Ranciaro A, Campbell MC, Hirbo JB et al (2014) Genetic origins of lactase persistence and the spread of pastoralism in Africa. Am J Hum Genet 94:496–510. https://doi.org/10.1016/j.ajhg.2014.02.009

Rando JC, Pinto F, Gonzÿlez AM et al (1998) Mitochondrial DNA analysis in Northwestern African populations reveals genetic exchanges with European, near-Eastern, and sub-Saharan populations. Ann Hum Genet 62:531–550

Reich D, Green RE, Kircher M et al (2010) Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468:1053–1060

Ruhlen M (1994) The origin of language: tracing the evolution of the mother tongue. John Wiley and Sons, New York

Sabeti PC, Reich DE, Higgins JM et al (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832–837. https://doi.org/10.1038/nature01140

Salas A, Richards M, De la Fe T et al (2002) The making of the African mtDNA landscape. Am J Hum Genet 71:1082–1111

Sánchez-Quinto F, Botigue LR, Civit S et al (2012) North African populations carry the signature of admixture with Neandertals. PLoS One 7:e47765

Sands B (2009) Africa's linguistic diversity. Lang Linguist Compass 3(2):559–580

Scheinfeldt LB, Tishkoff SA (2012) Living the high life: high-altitude adaptation. Genome Biol 11:133

Schlebusch CM, Skoglund P, Sjodin P et al (2012) Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. Science 338:374–379

Schoenbrun D (2001) Representing the Bantu expansions: What's at stake? Int J Afr Hist Stud 34:1–4

Schuster SC, Miller W, Ratan A et al (2010) Complete Khoisan and Bantu genomes from southern Africa. Nature 463:943–947

Scozzari R, Massaia A, D'Atanasio E et al (2012) Molecular dissection of the basal clades in the human Y chromosome phylogenetic tree. PLoS One 7:e49170. https://doi.org/10.1371/journal.pone.0049170

Sikora M, Laayouni H, Calafell F et al (2011) A genomic analysis identifies a novel component in the genetic structure of sub-Saharan African populations. Eur J Hum Genet 19:84–88. https://doi.org/10.1038/ejhg.2010.141

Stringer CB, Andrews P (1988) Genetic and fossil evidence for the origin of modern humans. Science 239:1263–1268

Tishkoff SA, Dietzch E, Speed W et al (1996) Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science 271:1380–1387

Tishkoff SA, Goldman A, Calafell F et al (1998) A global haplotype analysis of the myotonic dystrophy locus: implications for the evolution of modern humans and for the origin of myotonic dystrophy mutations. Am J Hum Genet 62:1389–1402

Tishkoff SA, Reed FA, Ranciaro A et al (2007) Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet 39:31–40

Tishkoff SA, Reed FA, Friedlaender FR et al (2009) The genetic structure and history of Africans and African Americans. Science 324:1035–1044

Tofanelli S, Bertoncini S, Castri L et al (2009) On the origins and admixture of Malagasy: new evidence from high-resolution analyses of paternal and maternal lineages. Mol Biol Evol 26:2109–2124

Underhill PA, Shen P, Lin AA et al (2000) Y chromosome sequence variation and the history of human populations. Nat Genet 26:358–361

Vansina J (1995) New linguistic evidence and the “Bantu expansion”. J Afr Hist 36:173–195

Verdu P, Austerlitz F, Estoup A et al (2009) Origins and genetic diversity of pygmy hunter-gatherers from Western Central Africa. Curr Biol 19:312–318

Vigilant L, Stoneking M, Harpending H et al (1991) African populations and the evolution of mitochondrial DNA. Science 253:1503–1507

Wood ET, Stover DA, Ehret C et al (2005) Contrasting patterns of Y chromosome and mtDNA variation in Africa: evidence for sex-biased demographic processes. Eur J Hum Genet 13:867–876

Chapter 7
Peopling and Population Structure of West and South Asia

Analabha Basu and Partha P. Majumder

Abstract South Asia was peopled on an early wave of human migration from out-of-Africa. This wave passed through the coastline of Saudi Arabia and India to Andaman and Nicobar archipelago and possibly to Australia as well. West Asia, on the other hand, served as a corridor for early and historical human migrations with various ancestries resulting in diverse settlements. At least four ancestral components are discernible in the extant populations of mainland Indian subcontinent. Some ancient tribal groups of Andaman and Nicobar archipelago carry signatures of an ancestry that is distinct from those found in India.

Keywords Southern exit route · Ancestral expansion · Mitochondrial DNA · Y-chromosomal DNA · Nuclear DNA · Genetic diversity · Population affinity

7.1 Introduction

Peopling of Asia is debated. Surprisingly, at the center of the debate lies the supervolcanic eruption that occurred sometime between 69,000 and 77,000 years ago at the site of present-day Lake Toba in Sumatra, Indonesia. This catastrophic super-eruption was at least 100 times more powerful than any volcanic eruption in recent history and had an enormous effect on the life on earth, particularly in Asia. One debate on the earliest peopling of Asia is centered on whether anatomically modern humans (AMHs) migrating out-of-Africa (OoA) arrived and settled in Asia prior to the Toba eruption and were they able to survive the ravages of the eruption even if they encountered a severe population bottleneck.

What sparked the debate of pre-Toba OoA migration are archeological finds from the Jurreru River valley of Kurnool district in Andhra Pradesh of southern India (Mellars 2006; Petraglia et al. 2007). Archeologists have discovered lithic

A. Basu · P. P. Majumder (✉)
National Institute of Biomedical Genomics, Kalyani, India
e-mail: ppm1@nibmg.ac.in

N. Saitou (ed.), *Evolution of the Human Genome II*, Evolutionary Studies,
https://doi.org/10.1007/978-4-431-56904-6_7

assemblage, resembling Middle Paleolithic technology, buried in layers pre-dating the Toba eruption (Petraglia et al. 2007). Furthermore, stone tools found in the deep layers of Toba ashes were also very similar to those pre-dating the eruption. This prompted the inference that the catastrophe did not wipe out the entire population living in the Jurreru valley, even though it may have created a severe bottleneck. This inference is far from being widely accepted because there is no evidence yet that the hominid population that used the pre-Toba tools were AMHs and not any of our archaic cousins. There is also no conclusive evidence that the people living in the Jurreru valley after the Toba eruption are the descendants of the population who lived there prior to the eruption.

Archeological evidence from the Jurreru river valley is not unique in terms of antiquity and evidence in favor of existence of Paleolithic technology throughout Asia. To make the academic debate more complex, archeological remains have been found in the Arabian peninsula and the Levant dating about 100,000 YBP indicating pre-Toba human civilization in the region. However, none of these archeological findings can conclusively distinguish between AMH and their hominid cousins.

The timing of the presence of AMHs in Asia is also central to understanding the early peopling of Australia and the Sahul. Molecular genetic data provides evidence that the AMHs were living in Australia after the OoA migration, before they populated modern day Europe. It remains unclear whether Asia was peopled on a single wave of OoA migration. There may have been multiple waves.

Jared Diamond has argued that because of its geography and an east to west spread, populations and technologies in Eurasia have spread like wildfire, as it calls for limited adaptation to new altitudes and environments (Diamond 1999). This echoes with a school of population genetic inference of a rapid colonization of Asia and Australia after the OoA migration (Macaulay et al. 2005; Quintana-Murci et al. 2004; Thangaraj et al. 2005; Forster 2004; Forster and Matsumura 2005). The most plausible route of rapid migration, named as the “costal-express,” postulates beachcombing from the horn of African through Southern Arabian peninsula along the coastal rim of Indian Ocean. Though there is a general consensus developing about this “Southern-Exit” route, whether this has been the only exit OoA to Asia is debated (Lahr and Foley 1994, 1998; Oppenheimer 2012; Macaulay et al. 2005).

Archeological remains of Paleolithic age are sparse throughout Asia. Posterior to these early exits OoA, both West and South Asia have experienced population bursts and multiple complex admixture and migration in the Neolithic age. These post-agriculture population demographic changes, largely contributed to the molecular variation that we observe in contemporary populations of West and South Asia.

7.2 Inferences from Uniparental Markers

Archeological artifacts offer little power to differentiate between early settlements of archaic humans compared to AMHs. Furthermore, accurate carbon dating cannot go beyond 40 K years (Oppenheimer 2012). Hence the debate encompassing early Paleolithic remains, speculated to be ~100–130 kya, are debatable to be attributed to early OoA migrations of AMH. The "African Eve" paper in 1987 using high-resolution restriction mapping compared the uniparental mtDNA variation of Africans compared with non-Africans and revealed a clear clustering of the African mtDNA in a tree cluster with deep branching, and all non-Africans in a separate cluster (with a few Africans "misclassified" as well), which is rooted to a branch coming out of the African cluster (Cann et al. 1987). This substantiated the hypothesis that all AMHs outside of Africa have originated from a successful migration of AMHs out-of-Africa and all AMHs colonizing the world have an African ancestor. This conclusion was soon substantiated and made more robust using data on the hyper-variable segment1 of mtDNA (Vigilant et al. 1991) and on complete mtDNA (Ingman et al. 2000). A further analysis of the mtDNA also revealed that all non-Africans coalesce to an ancestor who has an mtDNA L3 macro-haplogroup (groups of chromosomes bearing similar sets of mutations).

Which route did the first humans take when they moved out-of-Africa into Asia? Intuitively, one would argue that AMHs would have been walking along the river Nile, across the Sinai Peninsula ("northern exit route"). This seems logical because this feat could be achieved with very little adaptation to different climatic conditions, most rudimentary advancement of technology, as well as, it was walking over flat river valley. However, the modern theory of human dispersal OoA postulates that there could have been two waves instead of one (Lahr and Foley 1994; Underhill et al. 2001): one via a northern exit route through the Nile valley and another, via a southern exit route from the Horn of Africa across the mouth of the Red Sea along the coastline of India to southeastern Asia and Australia (Lahr and Foley 1994; Oppenheimer 2004; Quintana-Murci et al. 1999; Macaulay et al. 2005). Fossil records and evidence from genetic data show that islands in South-East Asia and Southern Pacific including Australia (Oppenheimer 2004; Macaulay et al. 2005) were populated at least as early as the oldest record suggestive of the presence of AMHs outside Africa via the northern exit route. It is also noted that, the archeological evidence of the northern exit route is associated with the Upper Paleolithic blade-dominated technology, whereas the remnants of the Southern-exit route are associated with much simpler Middle Paleolithic technology.

mtDNA data is consistent with a single successful migration resulting in a rapid spread of AMHs throughout a large part of Southern Levant, rims of the Indian Ocean to Bali and then eventually to Australia/New Guinea and the Bismarck Archipelago. The mtDNA haplogroup diversity of Africa is unique and the L1, L2, and L3 haplogroups are widespread throughout the African continent but is almost non-existent outside of Africa. When calibrated against the chimp mtDNA as an outgroup, the L1 happens to be the most ancient of all human haplotypes estimated to be arising ~190 ka and is ancestral to both L2 and L3 (Oppenheimer 2004; Soares et al. 2012). A phylogeny based on the complete mtDNA sequence data of 52 individuals randomly selected around the world revealed that the L3 is ancestral to the M and N haplogroups found widespread around the world (Oppenheimer 2004). The M and N haplogroups outside of Africa both have similar ages, which is slightly older than the within Africa derivatives of their immediate founding ancestor L3–195 (Oppenheimer 2004; Metspalu et al. 2004). The argument that AMH exit OoA proceeded initially along coastlines has depended largely on the rapidity of this movement, as inferred from genetic phylogeography, because only three key founder mtDNA haplotypes (M, N, and R) give rise to multiple region-specific branches en route. The molecular clock dating of the L3 haplogroup at 83 ka (61–86 ka) (Oppenheimer 2004) serves as an upper limit of the first OoA migration of AMHs. A much more recent reanalysis with 369 complete sequence of mtDNA L3 haplogroup within Africa estimates the origin of the L3 haplogroup as younger than 70 ka, thus ruling out any successful OoA migration of AMHs before the Toba eruption (Soares et al. 2012). The inferred origin of the N haplogroup is in the Levant and that of the M is in India and all the three mtDNA haplogroups (M, N, and R) have star-like structures, indicating population expansion, along the coastline of Indian Ocean. Their widespread spatial co-distribution along with expansion times close to the L3 haplogroup substantiates the southern exit route and rapid colonization along the coastline and inland, in face of sparse archeological data. Archeological evidence to resolve this controversy has been scanty, primarily because the coastlines of that period have become deeply submerged because of the rapid rise of sea levels after the last glacial period (Lahr and Foley 1998; Oppenheimer 2012).

Both the southern and the northern OoA exit routes have passed through Saudi Arabia. There is evidence of Neolithic expansions in the Arabian peninsula. There is also archeological evidence of Lower Paleolithic Oldowan industries and Middle Paleolithic technologies (Petraglia 2003; Petraglia and Alsharekh 2003). These indicate the presence of Homo erectus, ergaster, and possibly even neanderthalensis in this region. Analysis of mtDNA collected from this region revealed near-exclusive contributions of L, M, and N lineages from clades that have roots in Africa and south Asia. Yemen has the largest contribution of L lineages to this region, and may have served as the gateway of entry of people from Africa to this region (Kivisild et al. 2004). No primitive M or N lineages have been found in this region; therefore, it has been suggested that this region has been "more a receptor human migrations" (Abu-Amero et al. 2008).

Some recent archeological finds from India have also indicated that the major route of dispersal to India from out-of-Africa was through the southern route (Mellars 2006). The strongest genetic evidence in favor of an early southern exit into Indian subcontinent comes from the mtDNA signatures of Indian and other Asian populations. A vast majority of the populations of this region harbor derivatives of the mitochondrial DNA lineages (M and N), which are closely related to the Africa specific L3 lineage (Mountain et al. 1995; Basu et al. 2003; Bamshad et al. 1998, 2001; Kong et al. 2003; Sun et al. 2006). The Austro-Asiatic speaking tribal populations of India whom the anthropologists consider as autochthones of India show extensive diversity in their mtDNA, harboring many M-subhaplogroups and an expansion time of ~60 kya (Basu et al. 2003). The southern exit hypothesis is also supported by analyses of mtDNA data from Andaman Islands (Endicott et al. 2003; Kivisild et al. 2003; Thangaraj et al. 2005) and New Guinea (Tommaseo-Ponzetta et al. 2002; Forster et al. 2001).

Excavation sites in Israel and Lebanon, dated between 45 and 50 kya (the probable date of dispersal through the southern exit route was substantially earlier), which till date are the earliest evidence of migration of AMHs out of the African continents via the northern exit route, show the proximity and the territorial overlap of these modern humans with archaic humans, primarily the Neanderthals. However, all extant AMH populations outside of Africa show evidence that 1-4% of their genomes to be inherited from ancient admixture with Neanderthals and other archaic human cousins (Green et al. 2010). Even more surprising is the fact that the archaic admixture proportions are slightly more in East Asians as compared to present-day Europeans (Green et al. 2010; Reich et al. 2010).

Further, the Y-chromosomal lineages, C and D both defined by the same single nucleotide polymorphism M168 (Underhill and Kivisild 2007), originating outside of Africa. are found only in Asian continent and Oceania (Underhill and Kivisild 2007; Kivisild et al. 2003; Endicott et al. 2003), but not in Europe or North Africa. Whereas the most common ancient Y-lineage among Eurasian populations is derived from the East African F lineage, with an expansion time in Levant ~45 kya (Underhill and Kivisild 2007; Oppenheimer 2004, 2012). The migration of this lineage from Northeast Africa to the Levant and subsequently to the west populating Europe and to the east for rest of Eurasia between 40 and 20 ka serves as the dominant model of AMH migrations through the "northern exit route." The mtDNA haplogroup R, which is a descendent of the haplogroup N, is the ancestor to the widespread U-haplogroup. The mtDNA lineage U, which is likely to have arisen in Central Asia, has a high frequency in Eastern Europe, West, Central, and South Asia. The origin of the U-haplogroup ~50 kya indicates that the rise of the haplogroup has been rapid. The distribution of the sub-haplogroups of U in India poses difficult unanswered questions about the global phylogeny of the mtDNA. In India we find a clinal variation of the U-haplogroup consistent with large-scale migration from West and Central Asia. However, this lineage is composed of two deep sublineages, U2i and U2e, with an estimated split 50,000 years ago. The sublineage U2i is found in high frequency in India (particularly among tribes; 77%), but not in Europe (0%), whereas U2e is found in high frequency in Europe (10%), but not in India (it has very low frequencies among castes, but not among tribals). Thus, a substantial fraction of the U lineage, specifically, the U2i sublineage, may be indigenous to India (Kivisild et al. 1999; Basu et al. 2003). Analysis of the complete mtDNA genome sequence has revealed a large number of sequence variants within major haplogroups within Indian populations, many of which, however, are infrequent (Palanichamy et al. 2004).

The Levant and the West Asia have been the cradle of civilization. Although at least 11 regions of the Old and New World were involved as independent centers of origin of agriculture, encompassing geographically isolated regions on most continents the earliest development was around 11,500 years ago separately in both the Fertile Crescent and at Chogha Golan in modern day Iran about 9800 years ago (Balter 2013; Larson et al. 2014). This discovery of agriculture has hugely increased the carrying capacity of human populations. This led to exponential population growth in epicenters of the discovery and also led to rapid and extensive population migrations. As an example, the subgroups descending from the Y-chromosome haplogroup F (G, H, I, J and K) are found in >80% of the world's population, but almost exclusively outside of sub-Saharan Africa. All these haplogroups arose in different parts of West, Central, and South Asia but have expanded and are widespread in large part of Eurasia, South Asia, East Asia and have also spread out to Oceania and Australia. These migrations from West, Central, and South Asia were all possibly associated with demic diffusion of agriculture, meaning the actual movement of people in the carriage of an idea or technology. This also probably brought a qualitative change in human migration behavior where Indian subcontinent is a model example. The extent of genetic variation of female lineages (mtDNA) in India is rather restricted (Roychoudhury et al. 2001; Basu et al. 2003), indicating a small founding group of females. In contrast, the variation of male lineages (Y-chromosomal) is very high (Basu et al. 2003; Sengupta et al. 2006). This pattern may be indicative of sex-biased gene flow into India with more male immigrants than female (Bamshad et al. 1998), possibly starting with the Neolithic age gaining significance by large-scale male migration through invasions and wars. This indicates a common spread of the root haplotypes of haplogroups M, N, and R between 70 kya and 60 kya along the southern exit route. West and Central Asian populations are supposed to have been the major contributors to the Indian gene pool, particularly to the North Indian gene pool, and the migrants had supposedly moved into India through what is now Afghanistan and Pakistan. Using mtDNA variation data collated from various studies, we have previously shown (Basu et al. 2003) that populations of Central Asia and Pakistan show the lowest genetic distance with the North Indian populations (F_{ST} 0.017), higher distances (F_{ST} 0.042) with the South Indian populations, and the highest values (F_{ST} 0.047) with the northeast Indian populations; F_{ST} is a standard measure of genetic difference among populations derived from a common source population. Thus, northern Indian populations are genetically closer to Central Asians than populations of other geographical regions of India.

This phenomenon of large-scale sex-biased post-agriculture migration and displacement, coupled with rapid population growth, obscures ancient genetic signatures and results in the quick introduction of high genetic variability, often mimicking extreme natural selection (Zerjal et al. 2003). The success of some of the Y-chromosomal haplotypes that arose in Central Asia to spread across vast regions of Eurasia (Zerjal et al. 2003), as well as South and Southeast Asia, is indicative of the "success" of the cultural and technological dominance of west Eurasia and Central Asia (Zerjal et al. 2003; Underhill, Myres et al. 2010).

Similar to mtDNA, the Y-chromosomal gene pool in the Arabian peninsula is mostly of African ancestry and the diversity can be ascribed to Levantine expansion.

Coalescence time for the most prominent J1-M267 haplogroup in Saudi Arabia (11.6 ± 1.9 kya) is similar to that obtained previously for Yemen (11.3 ± 2 kya) but significantly older that those estimated for Qatar (7.3 ± 1.8 kya) and UAE (6.8 ± 1.5 kya). These facts and time estimates are consistent with the inference made from mtDNA analysis that this area has mainly been a recipient of gene flow from its African and Asian surrounding areas (Abu-Amero et al. 2009).

7.3 Inferences from the Nuclear Genome

Uniparental markers have provided a broad overview of major migrations of AMHs and their 150–200 k years of history and evolution. It also has distinct advantages, the mtDNA being inherited from a mother to the offspring delineates female specific migration while the paternally inherited Y-chromosome is testimonial to the male specific migrations. For the same reason the uniparental markers also have a smaller effective population size; mtDNA has approximately half of the effective population size compared to autosomes, whereas the Y-chromosome has approximately one-fourth. A smaller effective population size would also mean that it is much more susceptible to random genetic drift. This in effect would mean that a lot of variation of the uniparental markers would be lost just by random fluctuation in allele frequencies. Surprisingly, this has also been used as an advantage in our inferences of population genetics, as Stephen Oppenheimer testifies (Oppenheimer 2012): "This difference in drift between the two kinds of loci can be directly demonstrated. A good example is the observation that although autosomal loci do show a regular, progressive loss of diversity (e.g., in measures of heterozygosity) with increasing distance east from Ethiopia to the Americas, consistent with a serial founder effect (Ramachandran et al. 2005), this decay effect is linear and, unlike uniparental loci, shows no sudden step fall on leaving Africa. By inference, this lack of an initial step-drop in diversity means that many autosomal loci probably preserved (i.e., carried-over) multiple alleles on exit from Africa. In contrast, the finding of only single recent African lineages represented on mtDNA and NRY outside Africa (L3 and M168, respectively) itself suggests a very severe initial founding drift effect in those loci immediately after exit."

However, besides the advantages the incompleteness of using the uniparental markers for inferences of individuals' genetic ancestry and history is also profound. The uniparental markers, each represents a single locus in the entire genome and is representative of a very small portion of individual's ancestry. To tease out intricate details of migration, particularly to infer about historic and pre-historic admixture events, the autosomal markers are hugely more informative.

One of earliest efforts to catalog and interpret worldwide human variation was by Rosenberg et al. (2002) using the Human Genome Diversity Project (HGDP) panel, where they analyzed 377 autosomal microsatellite loci in 1056 individuals from 52 diverse population groups of the world. This dataset included populations like Mozabite, Bedouin, Druze, Palestinian from West Asia (Middle East in the paper) and populations like Balochi, Brahui, Makrani, Sindhi, Pathan, Burusho, Hazara, Uygur, and Kalash from Central-South Asia (Indian Subcontinent). Using the autosomal microsatellite markers, the within-population differences among individuals account for 93–95% of the genetic variation. An unsupervised clustering algorithm, used to identify population structure in the data, could identify five major genetic clusters in the worldwide data and that correlated strongly with the geographical locations of the extant populations. The major clusters which were identified were: Africa, Eurasia, East-Asia, Oceania, and the Americas. Interestingly it could not separate Europe, Middle East, and Central South Asia from each other (Fig. 1 of Rosenberg et al. 2002). The clustering algorithm would only identify Europe, Middle East, and Central South Asia when they were separated out from the rest of the world (Fig. 2 of Rosenberg et al. 2002). Subsequent to that, 6 years later, a study using the (almost) same set of individuals, but 650,000 common single nucleotide polymorphisms (SNPs), revealed seven clusters instead of five (Li et al. 2008). The larger informativeness in the SNP data could clearly identify ancestry specific to Middle East and Central-South Asia, thus teasing them apart as separate clusters from the other major populations of the world (Fig. 1 of Li et al. 2008). What this analysis also revealed is that all the West Asian (Middle-Eastern) populations are significantly admixed, primarily with large component genetic inputs from the European and Central South Asian populations.

A total of 2150 individuals drawn from different regions of Saudi Arabia were genotyped for about 650,000 single nucleotide and insertion/deletion polymorphisms (Al-Saud et al. 2015). Analyses of these data joint with comparable data from other populations of the Middle East and elsewhere revealed that the Saudi Arabian population is stratified into subgroups, with different subgroups showing genetic affinity with populations of disparate regions. This again strengthens the inference drawn from uniparental markers that the Arabian peninsula has been a walk-through region of diverse peoples, with a shallow antiquity. Subsequent isolation possibly resulted in the formation of subgroups with distinctive genetic profiles within the region.

The Indian subcontinent, occupying the center-stage of Paleolithic and Neolithic migrations, has been under-represented in genomewide studies of variation, but again holds a key in our understanding of how various migrations, different in origin, direction, magnitude, and time have shaped the history of contemporary populations in West and Southeast Asia. Analyses of data on 405 SNPs from a 5.2 Mb region on chromosome 22 in 1871 individuals from diverse 55 Indian populations have revealed that Indian populations form a genetic bridging West Asian and East Asian populations (Consortium 2008). The HUGO Pan Asian SNP Consortium's study (Abdulla, et al. 2009) also showed that most of the Indian populations shared ancestry with West Asian populations, which is consistent with the recent observations and our understanding of the expansion of Indo-European speaking populations. The study also provided evidence that the peopling of India (and also of southeast Asia) was via a single primary wave of migration out-of-Africa (Abdulla et al. 2009).

Using over 500,000 biallelic autosomal SNPs, Reich et al. (2009) have also found a north to south gradient of genetic proximity of Indian populations to western Eurasians/central Asians. In general, the central Asian populations were found to be genetically closer to the higher-ranking caste populations than to the middle- or lower-ranking caste populations. Among the higher-ranking caste populations, those of northern India are, however, genetically much closer ($F_{ST} = 0.016$) than those of southern India ($\mathrm{F}_{\mathrm{ST}} = 0.031$). Phylogenetic analysis of Y-chromosomal data collated from various sources yielded a similar picture.

Reich et al. (2009) have also proposed an elegant model where that extant populations of India were "founded" by two hypothetical ancestral populations, one ancestral north Indian (ANI) and another ancestral south Indian (ASI). Presumably, these ancestral populations were derived from ancient humans who entered India via the southern and the northern exit routes from out-of-Africa. All extant Indian populations are derived from admixture between the two ancestral populations, with the ANI contribution being higher among extant north Indian populations and that of ASI being higher among extant south Indian populations. In a more recent study (Moorjani et al. 2013), these investigators have shown that between 1900–4200 ybp, there was extensive admixture among Indian population groups, followed by a shift to endogamy. This model is simplistic, but intuitive and consistent with findings of earlier studies. It is simplistic because the origins of populations in northeastern region of India cannot be explained by this model since many past studies (cited earlier in this essay) have indicated genetic inputs into these populations from populations of southeast Asia. A more recent study (Basu et al. 2016) analyzing data on more than 80 k SNPs on 367 individuals from 18 mainland and 2 island populations has identified 2 more components in the mainland of the Indian subcontinent besides the ANI and ASI. These 2 newly identified ancestral components are primarily seen among Tibeto-Burman (TB) speaking tribal populations of northeast India and Austro-Asiatic (AA) speaking tribal populations of central and east India, and the authors have identified them as (AAA for Ancestral Austro-Asiatic and ATB as Ancestral Tibeto-Burman). Of these, the authors have shown that the ANI is co-ancestral to Central South Asians and the ATB to be co-ancestral to the East Asians, particularly the Southeast Asians of the HGDP [Fig. 3 and Supplementary Fig. 7 of Basu et al. (2016)]. The absence of significant resemblance with any of the neighboring populations is indicative of the ASI and the AAA being early settlers in India, possibly arriving on the "southern exit" wave out-of-Africa. Differentiation between the ASI and the AAA possibly took place after their arrival in India [Supplementary Fig. 2 of Basu et al. (2016)]. The ANI and the ATB can clearly be rooted to the CS-Asians and E-Asians [Fig. 3 of Basu et al. (2016)], respectively; they likely entered India through the NW and NE corridors, respectively. Ancestral populations seem to have occupied geographically separated habitats.

An even more contemporary approach, aided by the recent development of the technology of massively parallel genome sequencing has enabled researchers to look into genomes of 69 Europeans who lived between 8000–3000 years ago by enriching ancient DNA libraries for a target set of almost 400,000 polymorphisms

(Haak et al. 2015). This study has enabled our understanding of how migrations from West Asia and Near East, aided possibly by a powerful language, has influenced the genetic architecture of contemporary Europe. Using this DNA derived from ancient sources and comparing that with data from contemporary populations, the researchers show that the populations of Western and Far Eastern Europe followed opposite trajectories between 8000 and 5000 years ago. They also show that at the beginning of the Neolithic period, most of Europe was occupied by a relatively homogeneous group of farmers, whereas Russia was inhabited by a distinctive population of hunter-gatherers. By 6000–5000 years ago, the scenario changed and farmers throughout much of Europe had more hunter-gatherer ancestry than their predecessors, but in Russia, the Yamnaya steppe herders of this time were descended not only from the preceding eastern European hunter-gatherers, but also from a population of Near Eastern/West Asian ancestry. This steppe ancestry persisted in all sampled central Europeans until at least 3000 years ago, and is ubiquitous in present-day Europeans. These results provide support for a steppe origin of at least some of the Indo-European languages of Europe.

References

Abdulla MA, Ahmed I et al (2009) Mapping human genetic diversity in Asia. Science 326 (5959):1541–1545

Abu-Amero KK, Larruga JM, Cabrera VM, Gonzalez AM (2008) Mitochondrial DNA structure in the Arabian peninsula. BMC Evol Biol 8:45. https://doi.org/10.1186/1471-2148-8-45.

Abu-Amero KK, Hellani A, Gonzalez AM, Larruga JM, Cabrera VM, Underhill PA (2009) Saudi Arabian Y-chromosome diversity and its relationship with nearby regions. BMC Genet 10:59. https://doi.org/10.1186/1471-2156-10-59

Al-Saud H, Wakil SM, Meyer BF, Falchi M, Dzimiri N (2015) The genetic structure of the Saudi Arabian population. In: Abstract of presentation made in the annual meeting of the American Society of Human Genetics, ASHG 2015

Balter M (2013) Ancient DNA. Farming's tangled European roots. Science 342(6155):181–182

Bamshad MJ, Watkins WS et al (1998) Female gene flow stratifies Hindu castes. Nature 395 (6703):651–652

Bamshad M, Kivisild T et al (2001) Genetic evidence on the origins of Indian caste populations. Genome Res 11(6):994–1004

Basu A, Mukherjee N et al (2003) Ethnic India: a genomic view, with special reference to peopling and structure. Genome Res 13(10):2277–2290

Basu A, Sarkar-Roy N, Majumder PP (2016) Genomic reconstruction of the history of extant populations of India reveals five distinct ancestral components and a complex structure. Proc Nat Acad Sci USA 113:1594–1599. https://doi.org/10.1073/pnas.1513197113

Cann RL, Stoneking M et al (1987) Mitochondrial DNA and human evolution. Nature 325 (6099):31–36

Consortium IGV (2008) Genetic landscape of the people of India: a canvas for disease gene exploration. J Genet 87(1):3–20

Diamond JM (1999) Guns, germs, and steel: the fates of human societies. Norton, New York

Endicott P, Gilbert MT et al (2003) The genetic origins of the Andaman islanders. Am J Hum Genet 72(1):178–184

Forster P (2004) Ice ages and the mitochondrial DNA chronology of human dispersals: a review. Philos Trans R Soc Lond B Biol Sci 359(1442):255–264

Forster P, Matsumura S (2005) Evolution. Did early humans go north or south? Science 308 (5724):965–966

Forster P, Torroni A et al (2001) Phylogenetic star contraction applied to Asian and Papuan mtDNA evolution. Mol Biol Evol 18(10):1864–1881

Green RE, Krause J et al (2010) A draft sequence of the Neandertal genome. Science 328 (5979):710–722

Haak W, Lazaridis I et al (2015) Massive migration from the steppe was a source for indo-European languages in Europe. Nature 522(7555):207–211

Ingman M, Kaessmann H et al (2000) Mitochondrial genome variation and the origin of modern humans. Nature 408(6813):708–713

Kivisild T, Bamshad MJ et al (1999) Deep common ancestry of indian and western-Eurasian mitochondrial DNA lineages. Curr Biol 9(22):1331–1334

Kivisild T, Rootsi S et al (2003) The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations. Am J Hum Genet 72(2):313–332

Kivisild T, Reidla M, Metspalu E, Rosa A, Brehm A, Pennarun E, Parik J, Geberhiwot T, Usanga E, Villems R (2004) Ethiopian mitochondrial DNA heritage: tracking gene flow across and around the gate of tears. Am J Hum Genet 75:752–770

Kong QP, Yao YG et al (2003) Phylogeny of east Asian mitochondrial DNA lineages inferred from complete sequences. Am J Hum Genet 73(3):671–676

Lahr MM, Foley R (1994) Multiple dispersals and modern human origins. Evolution Anthropol: Issues News Rev 3(2):48–60

Lahr MM, Foley RA (1998) Towards a theory of modern human origins: geography, demography, and diversity in recent human evolution. Am J Phys Anthropol 27:137–176

Larson G, Piperno DR et al (2014) Current perspectives and the future of domestication studies. Proc Natl Acad Sci U S A 111(17):6139–6146

Li JZ, Absher DM et al (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319(5866):1100–1104

Macaulay V, Hill C et al (2005) Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. Science 308(5724):1034–1036

Mellars P (2006) Going east: new genetic and archaeological perspectives on the modern human colonization of Eurasia. Science 313(5788):796–800

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

Moorjani P, Thangaraj K et al (2013) Genetic evidence for recent population mixture in India. Am J Hum Genet 93(3):422–438

Mountain JL, Hebert JM et al (1995) Demographic history of India and mtDNA-sequence diversity. Am J Hum Genet 56(4):979–992

Oppenheimer S (2004) Out of Eden: the peopling of the world. Constable, London

Oppenheimer S (2012) Out-of-Africa, the peopling of continents and islands: tracing uniparental gene trees across the map. Philos Trans R Soc Lond B Biol Sci 367(1590):770–784

Palanichamy MG, Sun C et al (2004) Phylogeny of mitochondrial DNA macrohaplogroup N in India, based on complete sequencing: implications for the peopling of South Asia. Am J Hum Genet 75(6):966–978

Petraglia M (2003) The lower Paleolithic of the Arabian peninsula: occupations, adaptations, and dispersals. J World Prehist 17:141–179

Petraglia M, Alsharekh A (2003) The middle Palaeolithic of Arabia: implications for modern human origins, behaviour and dispersals. Antiquity 77:671–684

Petraglia M, Korisettar R et al (2007) Middle Paleolithic assemblages from the Indian subcontinent before and after the Toba super-eruption. Science 317(5834):114–116

Quintana-Murci L, Semino O et al (1999) Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. Nat Genet 23(4):437–441

Quintana-Murci L, Chaix R et al (2004) Where west meets east: the complex mtDNA landscape of the southwest and central Asian corridor. Am J Hum Genet 74(5):827–845

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

Reich D, Thangaraj K et al (2009) Reconstructing Indian population history. Nature 461 (7263):489–494

Reich D, Green RE et al (2010) Genetic history of an archaic hominin group from Denisova cave in Siberia. Nature 468(7327):1053–1060

Rosenberg NA, Pritchard JK et al (2002) Genetic structure of human populations. Science 298 (5602):2381–2385

Roychoudhury S, Roy S et al (2001) Genomic structures and population histories of linguistically distinct tribal groups of India. Hum Genet 109(3):339–350

Sengupta S, Zhivotovsky LA et al (2006) Polarity and temporality of high-resolution y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of central Asian pastoralists. Am J Hum Genet 78(2):202–221

Soares P, Alshamali F et al (2012) The expansion of mtDNA Haplogroup L3 within and out of Africa. Mol Biol Evol 29(3):915–927

Sun C, Kong QP et al (2006) The dazzling array of basal branches in the mtDNA macrohaplogroup M from India as inferred from complete genomes. Mol Biol Evol 23(3):683–690

Thangaraj K, Chaubey G et al (2005) Reconstructing the origin of Andaman islanders. Science 308 (5724):996

Tommaseo-Ponzetta M, Attimonelli M et al (2002) Mitochondrial DNA variability of West New Guinea populations. Am J Phys Anthropol 117(1):49–67

Underhill PA, Kivisild T (2007) Use of y chromosome and mitochondrial DNA population structure in tracing human migrations. Annu Rev Genet 41:539–564

Underhill PA, Passarino G et al (2001) The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. Ann Hum Genet 65(Pt 1):43–62

Underhill PA, Myres NM et al (2010) Separating the post-glacial coancestry of European and Asian Y chromosomes within haplogroup R1a. Eur J Hum Genet 18(4):479–484

Vigilant L, Stoneking M et al (1991) African populations and the evolution of human mitochondrial DNA. Science 253(5027):1503–1507

Zerjal T, Xue Y et al (2003) The genetic legacy of the Mongols. Am J Hum Genet 72(3):717–721

Chapter 8
Europe

Jaume Bertranpetit and Guido Barbujani

Abstract Many aspects of the population history of Europe have been reconstructed from genetic and genomic evidence, drawing inference both from levels and patterns of present genetic diversity and from ancient DNA data. The main events that have been proposed to have shaped the European gene pool are probably as few as four, including: (1) the replacement of Neanderthals by anatomically modern humans around 40,000 years BP, with very limited, but nonzero levels of introgression, (2) the Paleolithic peopling of the continent, followed by a demographic contraction at the last glacial maximum; (3) a large demographic replacement associated with the Neolithic expansion of early farmers, that would have affected the genetic makeup of Europeans more strongly in the Southeast than in the Northwest; and (4) successive migration phenomena, mostly from the East, with a peak in the Bronze age. The relative importance in the present gene pool of the Paleolithic and Neolithic substrates, both having entered Europe from the Southeast, has been a matter of debate, with an important increase of evidence favoring a strong impact of the Neolithic. Several open questions remain to be addressed, including the relative role of cultural and migrational phenomena in causing the Indo-European linguistic expansion. Recent evidence of ancient DNA, with its diversity in time and space, makes the picture more complex and opens the possibility of other, more recent events, having left their signature in the European gene pool.

Keywords Population history · Europe · Neanderthal replacement · Paleolithic substratum · Neolithic · Wave of advance · Human migrations

J. Bertranpetit (✉)
Institut de Biologia Evolutiva, Universitat Pompeu Fabra, Barcelona, Catalonia, Spain
e-mail: jaume.bertranpetit@upf.edu

G. Barbujani
Department of Life Sciences and Biotechnologies, University of Ferrara, Ferrara, Italy

N. Saitou (ed.), *Evolution of the Human Genome II*, Evolutionary Studies,
https://doi.org/10.1007/978-4-431-56904-6_8

8.1 Introduction: Genes and the Neolithic

The reconstruction of population history may be seen as nothing but a small scale reconstruction of a biological evolutionary process. In fact the success of genetics in reconstructing evolutionary processes is largely due to the combination of a powerful basis of theoretical population genetics with a fast-growing and by now very extensive amount of genetic and genomic data. These data must be properly analyzed, since not all regions evolve at the same pace. There are genome regions with higher mutation rates than others (mitochondrial DNA mutates faster than nuclear DNA), there are specific DNA loci with different mutation rates (repeated DNA, like microsatellites or minisatellites are more prone to mutating than other DNA regions), and there are different factors shaping the genome (like recombination or gene conversion as opposed to mutation), all resulting in different speeds in shaping genome differentiation. Thus we have a wide spectrum of possibilities when raising evolutionary questions by looking at the genome to find places and patterns that indicate a deep evolutionary time and others a recent historical time.

The study of the origins of the first European settlers has been of paramount interest, partly due to the relative abundance of genetic data, partly to the placements and roots of most western science and scientist. Moreover Europe is the continent with the largest amount of studies on the human past, be them studies on archaeology, physical anthropology or historical linguistics (see Pinhasi et al. 2012; Barbujani 2013 and references therein). After long-standing academic discussions, archaeologists have been able to approximately define the main transitions, showing a strong communality all over Europe for the main cultural changes; linguists, on the other hand, have achieved a good classification of extant languages and a fair knowledge of many of the extinct ones. Population genetics, trying to be much powerful than the anthropometry had been in previous times, began to show differences in allele frequencies for many of the protein polymorphisms that were described as gene products and that are now called “classical genetic markers,” including blood groups. For many years, the accumulation of information on the genes of many human populations, most of them in Europe, did not have a clear goal in its endeavor: in most cases it was just a matter of description and classification, with a boost of calculating genetic distances. In no case was it clear whether the goal was to understand the specific genetic system or population.

In what might be regarded as the very first step towards genomic analyses, L.L. Cavalli-Sforza revolutionized the field introducing a very simple idea: instead of considering the geographic variation of each of the genetic variant (or marker) studied, he proposed to jointly analyze genetic variation by estimating synthetic functions by principal component analysis (PCA). In a seminal paper (Menozzi et al. 1978) followed by a large and well elaborated book (Cavalli-Sforza et al. 1994) the main ideas about how to use genetics for reconstructing human population history were put forward. In fact, Cavalli-Sforza’s work set the foundations for the modern notion that the present genetic structure of the human population must be interpreted as the footprint left by our distant past: the main changes were produced when the populations were small (mainly surviving, at low densities, as hunter-gatherers) and the factors allowing strong demographic changes had to be considered in relative terms, that is, as proportions of the population being subjected to that evolutionary or demographic phenomenon. The genetic impact of a population increase or of an admixture process will depend on the proportion of people involved; these changes will tend to be more profound in small populations than in large ones. That it is why in some cases the origin of current patterns of genome diversity can be identified in our distant past, and in other cases in relatively recent events.

As we just said, in 1978, Menozzi et al. (1978) first summarized the variation of many allele frequencies in Europe by PCA. Using data on 38 independent alleles at nine protein loci (ABO, Rh, MNS, Le, Fy, Hp, PGMi, HLA-A, and HLA-B), they identified a broad Southeast to Northwest cline spanning from the Near east to the Atlantic coasts of Europe. This first principal component had a strong weight as it accounted for 27% of the total genetic variance in the whole region for many independent loci. This way, for the first time, we had available an interpretation of genetic diversity in terms of population history, independent of the specific genetic markers being used. The most important innovation was that Menozzi et al. (1978) noticed a general correspondence between their PCA map and archaeological maps showing the diffusion of farming activities from the Near East into Europe, and proposed that the latter explain the former. Thus they advanced an interpretation of the genetic data in terms of the demographic impact that would have had a well documented consequence in the archaeological record: the Neolithic transition. The main basis for the interpretation was that the cultural change in the Neolithic would have had a very strong demographic impact, allowing in food-producing populations (through cultivation of plants or breeding of animals) a demographic growth that would have a major genetic impact, lasting until present times.

The impact of the Neolithic transition from food collection to food production in shaping the genetic constitution of the European populations has been, ever since then, a central point of research. Some studies confirmed its crucial importance; others proposed that other historical events had a greater effect in shaping genetic diversity in modern of Europeans. We will come back to this question with the data given by later genetic technologies.

8.2 The Making of the European Paleolithic Populations

To better understand the demographic transitions occurring in anatomically modern Europeans, it is convenient to briefly mention the likely previous stages of human occupation of Europe.

Let us first discuss the initial peopling of Europe by hominins, a neologism referring to the group consisting of modern humans, extinct human species, and all our immediate ancestors (including members of the genera *Homo*, *Australopithecus*, *Paranthropus,* and *Ardipithecus*). In recent years, paleontologists have been discovering older and older remains: while some years ago human remains dated around one million years ago were considered very old, the discovery in Dmanisi, Georgia of 1.8 million years human fossils (Lordkipanidze et al. 2013) shows that the debate about the first stages of early hominin spread in Europe is still open. But this debate is far away from our main purpose of a genetic history of Europe: the most ancient samples for which it has been possible to retrieve DNA date back to around 40,000 years ago and encompass both Neanderthals and modern humans. These two groups have been considered by paleontology two very distinct clades and for many years anthropologists disagreed on whether there was a clear discontinuity between the two groups, or whether Neanderthals could be regarded as ancestral, at least in part, to the later populations.

Thanks to the genetic analysis of ancient samples, there is now compelling evidence that around 40,000 years ago (between 45,000 and 35,000, depending on the places), when anatomically modern humans arrived in Europe, usually associated with the Upper Paleolithic cultures, they had some interbreeding with Neanderthals. This interbreeding would have contributed some 1–3% of the present genetic pool of Europeans (Pääbo 2015 and references therein). Thus, the pre-Neolithic populations in Europe would consist of the descendants of modern humans that, having originated in Africa, spread all over the word and in their expansion partially interbred with ancient hominin populations; for Europe, the Neanderthals. The interaction between the two different groups has been and will be a matter of debate, but what is clear in the more sophisticated analysis is that the genome of modern humans who lived around 40,000 years ago in Eurasia (but not in Africa) shows the signature of Neanderthal interbreeding. Because the fragments of likely Neanderthal origin in the genome are quite long, and have not been broken yet by genetic recombination, a reasonable conclusion is that interbreeding between the two human forms occurred just a few thousand years before (Fu et al. 2014). More recently (Fu et al. 2015) this admixture has been highlighted at much higher levels (between 6 and 9% of the genome with a likely Neanderthal origin) in an individual from present Romania, dated between 37,000 and 42,000 years ago, with long DNA fragments, again indicative of a recent Neanderthal admixture. The most recent findings suggest a complex process of extinction of Neanderthals, with levels of admixture with modern humans varying in space and time.

Even if we do not have an exact date, it is likely that by 35,000 years the modern humans would have expanded all over Europe, Neanderthals as specific populations would have vanished (the very late specimens from South Iberia are dated at 29,000 years BP) and the cultural landscape would be mostly occupied by Upper Paleolithic cultures (mainly Aurignacian). Controversy still exists on the interpretation of the previous cultures, Châtelperronian (Central-South France) and neighboring Uluzzian (Central-Southern Italy), as they have both Upper Paleolithic characteristics and may be associated with Neanderthals according to morphological evidence; future studies will have to clarify the transition (see e.g. the mass spectrometry data by Higham et al. 2014, bringing to 40,000 years ago the likely date of Neanderthal extinction).

Be that as it may, (1) the modern European genomes seem to contain components that may be traced back to Neanderthal ancestors (see Villanea and Schreiber 2019), but (2) the differences among European populations have hardly anything to do with a geographically variable Neanderthal contribution. Rather, they reflect the numerous demographic changes occurring during and after the first, Paleolithic colonization of the continent.

8.3 The Genetic Legacy of Paleolithic Europe

As we are seeing, our understanding of the genetic history of Europeans is based on studies of DNA of both present individuals and ancient archeological human remains. No doubt, the analysis of the former is much easier and it has been possible to accumulate a lot of data for many different populations. In turn, these data have been interpreted in the light of sophisticated evolutionary algorithms connecting the patterns observed in the genome to their likely generating processes.

Mitochondrial DNA (mtDNA) was for many years the primary source of information on human molecular diversity (that is, describing the DNA and not its products, as it was done with the classical genetic markers), and has long been the only one available for comparing populations at the large geographical scale. mtDNA is basically unaffected by recombination, a process that shuffles the genetic information in the autosomes, which makes it simple to summarize genetic diversity by evolutionary trees or networks. This way, a well known and fully accepted unique tree of all the mtDNA diversity in the world became available. With mtDNA (as well as with data on Y-chromosome diversity) it is also easy to estimate the age of mutations shared by the groups of variants, that is, of the age of the main branches of the evolutionary tree, or haplogroups. In the initial analyses of mtDNA, very old ages were estimated for the main European haplogroups; later they were corrected depending on mutation rate and the definition of the haplogroups, but very few of the main haplogroups had ages younger than 10,000 years (Richards et al. 2000 and references therein). The presence in Europe of haplogroups having a very old date led to the proposal of a fundamental Paleolithic background for the present European populations.

Richards et al. (2000) made two strong assumptions: (1) each mitochondrial cluster can be assigned, in its entirety, to one of the main migration phases (i.e. Neolithic and Paleolithic), and (2) the age of each cluster approximates very closely the timing of the migration event. Under those assumptions these studies concluded that the main mitochondrial variants in Europe predate the Neolithic expansion, and hence that the genetic structure of Europe has been determined in Paleolithic times (Barbujani 2013). It was just a small fraction of other haplogroups that had a younger age; when this was in the order of 10,000 years, it was allocated to a Neolithic origin and as the frequency of these variants over Europe was 20%, that was also considered to be a measure of the fraction of the Europeans' ancestors who entered the continent in Neolithic times. These initial mtDNA contributions did not consider any spatial variation within Europe and, what is worse, did not make an appropriate differentiation between gene trees and population trees.

This approach (often referred as a phylogeographic analysis) was strongly influential, mainly because on the one hand, it was easy to obtain mtDNA sequences and attribute them to the proper haplogroup, making it a popular method among evolutionary laboratories; on the other hand, there was not much theoretical development in the understanding of the mtDNA variation. Now, the main conclusions drawn from mtDNA analysis have been dismissed both with criticisms of the methods and with new and more detailed data available beyond the frequencies of haplogroups defined a priori by the researchers. Barbujani (2013) summarized the criticisms: the estimated age of some European mitochondrial haplogroups, and hence the entire mitochondrial genealogy, predate the arrival in Europe of the first anatomically modern humans (some 40,000 years ago), which can only mean that the relevant mutation events occurred out of Europe. Furthermore, the definition of haplogroups is arbitrary, depending on what one considers their founding mutation; by splitting a haplogroup in subhaplogroups of inferior order, clusters can be constructed with any lower time-depth; and finally, the average coalescence time of two sequences sampled from two diverging populations is older, or much older, than the split of the groups. Unless a group colonizing a new territory passes through a strong and long-lasting bottleneck, part of its initial diversity will be maintained. Therefore, the coalescence times inferred from samples of its descendants will be close to the coalescence times of the population of origin, consistently overestimating the age of the derived populations. In short: people, not haplogroups, migrate, and hence inferences about population history must be based on measures of genetic diversity between populations, not between molecules (Barbujani 2013).

It is interesting to note the attention that another historical event has had in the interpretation of population history: as it may be easily concluded, during the last glacial maximum (some 18,000 years ago, much before the Neolithic), an ice sheet occupied an important part of Northern Europe and pushed the human populations to the South, mainly into the three main peninsulas: Iberian, Italian, and the Balkans. When the ice retracted, the populations were able to expand northwards and re-occupy most of the Europe. The biological consequences of this scenario have been well documented in the distribution of many species along the continent, as indicated by the influential work of the late Godfrey M Hewitt (2011 and references therein). Several studies tried to identify markers in the mtDNA gene pool that could be interpreted this way, among which one that proposed a specific genetic relationship between Iberia and Scandinavia, affecting even the Saami, based on the presence and age of a given haplogroup (Torroni et al. 2001). This interpretation has been repeated in many papers (see for example, Soares et al. (2010)) without a reassessment that, fortunately, is now possible with data beyond the partial mtDNA in present populations and interpretation that go beyond the ages of particular haplogroups. Both the possibility of having information of other parts of the genome (in fact mtDNA behaves as just a single locus in our genome) and the retrieval of DNA from ancient samples (Fu et al. 2012) have revolutionized the field; among other things, it is now evident that the basic assumptions of the Paleolithic model were wrong.

8.4 The Genetic Legacy of Neolithic Europe

As we shall see, recent analyses of ancient DNA have proved that the European gene pool, and specifically the mitochondrial gene pool, profoundly changed since Paleolithic times (Brandt et al. 2013). However, long before these ancient DNA data became available, an articulated model of the impact of Neolithic gene flow in Europe had been put forward (Menozzi et al. 1978; Cavalli-Sforza et al. 1994). What is referred to as the Neolithic model is based on a complex framework with five main factors: (1) an origin of agriculture in the fertile crescent, some 10,000 years ago; (2) the existence of allele-frequency differences at some loci between Near Eastern farmers and European hunter-gatherers; (3) a demographic growth in farming communities, prompted by the greater availability of resources; (4) a gradual dispersal of farmers towards North and West, looking for new arable land; and (5) a lower rate of population increase for hunting-gathering than for farming people, even when, after the farmers' dispersal, they came to occupy the same territories (Barbujani 2013). In time, this process is going to generate continental clines and this is the reason of the initial interpretation of Cavalli-Sforza et al. (1994) in which the spread of farming technologies in Europe was regarded mainly as a process of movement of the people carrying those technologies, and not as just the spread of ideas through cultural contact. The importance of this view has now been supported by several independent lines of evidence, only some of which we mention here.

Fu et al. (2012) used contemporary mtDNA datasets with limited ancient genetic data. Short stretches of ancient mitochondrial DNA (mtDNA) from skeletons of pre-Neolithic hunter-gatherers as well as early Neolithic farmers showed a marked genetic discontinuity between Paleolithic and Neolithic Europe. Pre-Neolithic Europe appeared very homogeneous, with most individual sequences falling within haplogroup U. Even for haplogroup H, which was sometimes interpreted as the signature of Paleolithic expansions in Europe (Achilli et al. 2004), there is no evidence of its pre-Neolithic occurrence. By contrast, the number and frequency of haplogroups observed in the Neolithic samples matched closely the observations on modern Europeans. Also, Fu et al. (2012) discovered a signal of a past population expansion approximately 9000 years ago for mtDNA typical of early farmers. The observed changes over time suggest that the spread of agriculture in Europe involved the expansion of farming populations into Europe followed by the eventual assimilation of resident hunter-gatherers.

Sánchez-Quinto et al. (2012) typed more than 20,000 genomic SNPs in two 7000 year old Mesolithic Iberian hunter-gatherers. Accurate statistical analyses, based on Approximate Bayesian Computation, showed that a model in which Neolithic farmers have a greater demographic growth rate and largely replace preexisting Paleolithic people is more than 3000 times as likely as a model of genetic continuity in Europe since Paleolithic times. Their results indicate that the European population underwent substantial changes with the arrival of farming technologies.

Fernández et al. (2014) presented mitochondrial DNA data of the original Near Eastern Neolithic communities with the aim of providing an adequate background for the interpretation of Neolithic genetic data from European samples. Sixty-three skeletons from Pre Pottery Neolithic B sites dating between 8700 and 6600 cal. B.C. were analyzed and 15 validated mitochondrial DNA profiles were recovered. Comparisons performed among the ancient datasets allowed the authors to identify derived mitochondrial DNA haplogroups as potential markers of the Neolithic expansion, whose genetic signature would have reached both the Iberian coasts and the Central European plain. A significant conclusion was that the Neolithic expansion would have had an important pioneer seafaring colonization component, a notion also proposed to explain the fast expansion of the cardial Neolithic in the Western Mediterranean (Zilhâo 2001).

Globally speaking, the recent ancient DNA data confirm a genetic discontinuity between Mesolithic hunter-gatherers and early farmers (Keller et al. 2012; Sánchez-Quinto et al. 2012; Skoglund et al. 2012; Mathieson et al. 2018; Lazaridis et al. 2014), but add complexity to the picture, identifying at least two main, and distinct genetic components in the Europeans' genomes. One has been tentatively attributed to farmers originating in Southeastern Europe or in the Levant, the other to Northwest European hunter-gatherers (Skoglund et al. 2012). As we shall see, a third component has emerged in successive analyses.

8.5 More Complex Models, More Complex Facts

New genomic data keep accumulating, adding nuances to the picture. With the present evidence, no doubt that a single process accounting for the European genetic structure is too much of an oversimplification. The sequence analysis of ancient samples is giving a more articulate picture than could be imagined during the long-lasting Paleolithic versus Neolithic controversy. Lazaridis et al. (2014) proposed that three ancestral populations significantly contributed to the genomic variation of present-day Europeans. They sequenced the genomes of a 7000 year old farmer from Germany and eight 8000 year old hunter-gatherers from Luxembourg and Sweden and analyzed them along with other ancient genomes and 2345 contemporary humans. The results show that most present-day Europeans are likely to derive from at least three recognizable source populations: west European hunter-gatherers, who contributed ancestry to all Europeans but not to Near Easterners; ancient north Eurasians related to Upper Paleolithic Siberians, who contributed to both Europeans and Near Easterners; and early European farmers, who were mainly of Near Eastern origin but also harbored west European hunter-gatherer related ancestry, probably resulting from admixture between the two groups in the early stages of the Neolithic demic diffusion. Lazaridis et al. (2014) modelled these populations' deep relationships and showed that early European farmers had some 44% ancestry from a "basal Eurasian" population that split before the diversification of other non-African lineages. This result (and others that will come in the future) will modify the general models with the complexity that we know had to exist in the European Paleolithic populations.

Much earlier, Dupanloup et al. (2004) estimated admixture coefficients from autosomal, mitochondrial and Y-linked polymorphisms, using a model regarding the European populations as derived from hybridization among up to four potential parental populations. Two main genome components became apparent, presumably corresponding to the contributions of the first, Paleolithic Europeans, and the early Neolithic farmers, the second component decreasing from east to west, as predicted by a model in which the alleles of Neolithic immigrants from the Near East got diluted during an expansion towards the Northwest.

In the study of ancient mtDNA several results are also of special interest. Haak et al. (2015) successfully extracted and sequenced the mtDNA from 24 out of 57 Neolithic skeletons from various locations in Germany, Austria, and Hungary, remains dated to the LBK (also known as linear pottery culture or *Linearbandkeramik*) period (7000–7500 years ago). They found that 25% (6 out of 24) of the samples are of a distinctive and rare N1a lineage of the mtDNA well-known phylogeny. Furthermore, five of these six individuals display different N1a haplotypes and they were widespread in the LBK area. Europeans today have a 150 times lower frequency (0.2%) of this mtDNA type, suggesting that these first Neolithic people did not have a strong impact on the genetic background of the modern European female lineages. They proposed that small pioneer farming groups carried farming into new areas of Europe and that, once the technique had taken root, the surrounding hunter-gatherers adopted the new culture and then outnumbered the original farmers, diluting their N1a frequency to the low modern value. Thus, this result and its interpretation would support the cultural diffusion model of the Neolithic technologies, where the farming culture itself spread without the people originally carrying these ideas. They proposed that within the current debate on whether Europeans are genetically of Paleolithic or Neolithic origin, their data lends weight to the arguments for a Paleolithic origin of Europeans. But, using a similar approach in South Europe, Sampietro et al. (2007) produced mitochondrial DNA sequences from 11 Neolithic remains from Granollers (Catalonia, northeast Spain) dated to 5500 years BP. Phylogeographic analysis showed that the haplogroup composition of the Neolithic population was very similar to that found in modern populations from the Iberian Peninsula, suggesting a long-time genetic continuity, at least since Neolithic times (as later confirmed by Fu et al. (2012)); this result contrasts with that of Haak et al. (2015) for Central Europe and suggests that the Neolithic dispersals had different intensities and impact upon different populations of Europe. The authors proposed a dual model of Neolithic spread: acculturation in Central Europe, in agreement with the results of Haak et al. (2015) and demic diffusion in southern Europe in agreement with Sampietro et al. (2007), as initially suggested >20 years ago (Simoni et al. 2000).

Non-genetic evidence suggests that the causes of this complex pattern are also complex. Factors that may have had some relevance include: (1) zones of geographical heterogeneity are known to cause resistance to demic diffusion; (2) archaeological data suggest multiple waves of Neolithic expansion; (3) the impact on the Neolithic expansion had to be very different depending on the life style of late Paleolithic and Mesolithic populations, mainly between the Mediterranean and

Atlantic; (4) the exact weight of cultural and demic factors in causing the early spread of agriculture must have varied from place to place; and, (5) rates of farming expansions appear inversely correlated with the intensity of preexisting agricultural activities. This complexity allowed Fort (2012) to integrate the demic and cultural models in a unified framework and show that cultural diffusion explains ~40% of the spread rate of the Neolithic transition in Europe, as implied by archaeological data; he concludes that cultural diffusion cannot be neglected, but, that demic diffusion was the most important mechanism in this major historical process at the continental scale.

The open question, now, is what exactly happened, and in which areas of Europe the transition to farming was not caused by immigration (as was probably the rule) but by a technological shift in the previously hunting and gathering population.

8.6 Post-Neolithic Events in the European Gene Pool

The complex events affecting the gene pools of the diverse European populations after the Neolithic expansion is a matter open to future research. In fact, there is evidence that present populations in given locations are rather different from those having lived in the past, be it a few centuries ago, or in older times, Neolithic or Paleolithic. Sharp genomic discontinuities have been described, in the British isles (Brace et al. 2019), Central Italy (Antonio et al. 2019) and Iberia (González-Fortes et al. 2019), to name just a few, and will be found when better data from ancient samples from different time periods and geographic locations will become available. Some of them will have a general explanation that the archaeological record will be able to describe in terms of expansion of given cultures and others may have a more local meaning, with populations migrating and vanishing in a more dynamic way than usually assumed.

As the more ancient events, occurring at lower population densities, are more likely to have produced changes with stronger genetic influence, some scholars have focused on whether specific cultural changes may explain the genetic patterns. No doubt one of the most influential has been the proposal by Colin Renfrew (1987 and many later papers) of a common expansion of the Neolithic and the Indo-European languages leading to what is called the farming/language dispersal hypothesis. Under this hypothesis, a single major population expansion would account, at the same time, for the diffusion from the Levant into Europe through Anatolia of agricultural technologies, genes, and Indo-European languages. A more recent view (Bouckaert et al. 2012) recently gave support to the Anatolian origin of Indo-European by extensive analyses of linguistic diversity using Bayesian phylogeographic approaches; in this study, both the inferred timing and root location of the Indo-European language trees fit with an agricultural expansion from Anatolia beginning 8000 to 9500 years ago, that would be at the basis of the genetic influence of the Neolithic expansion into Europe.

Batini et al. (2015) expanded the ancient DNA work in the Y chromosome. Previous to that, analysis of male-specific region of the Y chromosome had been widely applied to the question of the Paleolithic/Neolithic origin of the European gene pool, with different results, some of them strongly supporting the Neolithic origin of European paternal lineages (Balaresque et al. 2010). In this new study, by resequencing 3.7 Mb of Y-chromosome DNA in 334 males, comprising 17 European and Middle Eastern populations, they showed that European patrilineages underwent a recent continent-wide expansion. Dating indicates that three major lineages (I1, R1a and R1b), accounting for 64% of their sample, had very recent coalescent times, ranging between 3500 and 7300 years ago. Their results indicate a widespread male-specific phenomenon that focuses interest on the social structure of Bronze Age Europe. A recent, low-coverage study of the genomes of 101 Eurasian individuals, mostly from the Bronze Age, but including some Late Neolithic and Iron Age individuals, actually showed that large scale gene flow and even episodes of demographic replacement took place in the Bronze age, despite population sizes having reached substantial levels by that time (Allentoft et al. 2015). Whether or not these phenomena also led to major linguistic changes was claimed, but not substantiated with convincing data, in the largest, so far, study of ancient DNA, involving almost 400,000 SNPs in 69 Europeans who lived between 8000 and 3000 years ago, who were compared with individuals from over 200 modern populations (Haak et al. 2015). What this study did show, however, was that some 8000 to 7000 years ago, the Neolithic transition was accompanied by the occurrence in Germany, Hungary, and Spain, of genetically-related people, who could be distinguished from earlier hunting-gathering people of the same regions. On the contrary, the hunting-gathering populations of the Russian area were more closely related with a 24,000 year old individual from Siberia (Raghavan et al. 2014). Genomic data document a contact between these two groups some 4500 years ago, leading to the European spread of alleles of Eastern origin, which are still widespread among current Europeans (Haak et al. 2015).

8.7 Comparing Models

One of the recurrent discussions in evolutionary genetics is to which extent the models used are powerful enough to discriminate between competing hypotheses. While some scientists insist on having more and better genetic data and others develop sophisticated analysis toolkits to exploit the information available, formal tests of the competing hypotheses based on explicit demographic modeling have been surprisingly rare so far.

Using forward simulations of five evolutionary models in which only Indo-European speakers were considered, all incorporating isolation by distance, Barbujani et al. (1995) showed that European clines are best accounted for by two models where dispersal of Neolithic farmers from the Near East depends only on population growth. Models of greater complexity, where archaeological time data constrained the timing of the farmers' expansion failed to explain a larger fraction of the observed genetic variation, and hence appear unnecessary. However, the same study showed that gradients can arise in two ways, i.e. not only by incomplete admixture between dispersing farmers and preexisting hunters and gatherers (as expected under Neolithic demic diffusion), but also by founder effects during a population expansion at low densities, hence not necessarily in Neolithic times (Barbujani et al. 1995; Barbujani 2013).

Currat and Excoffier (2005) studied the expected distribution of coalescence times in Europe through a backward simulation based on the core algorithm of Barbujani et al. (1995), and compared it with the distribution of mitochondrial coalescence times over Europe. They concluded that: (a) a very small Paleolithic contribution to the original founder populations would be sufficient for most modern people to be descended from Paleolithic ancestors in the sense that minute amounts of gene flow between Paleolithic and Neolithic populations should lead to a massive Paleolithic contribution to the current gene pool of Europeans, and (b) the clinal distribution of allele frequencies might just reflect a bias in favor of highly variable SNPs. Such contrasting results in two studies essentially based on very similar methodologies show how much minor assumptions on generation times, admixture rates, rates of local gene flow, and other parameters for which we have no plausible information may have strong effects on the conclusions we may draw.

There is a need to fit the growing bulk of genetic date, both of present and ancient populations, within a more complex framework of models than usually used and this should be a research priority. This means asking in the first place whether two main models (say Paleolithic versus Neolithic expansion) are different enough to be discriminated by analyses of modern DNA diversity. But, as both expansions occurred largely along the same Southeast to Northwest axis and samples are scattered over a broad space, it is difficult to recognize whether the available methods and datasets can safely tell us which of the main models is better.

8.8 Conclusion

The incorporation of the genetic analysis to the human endeavor of reconstructing our past has been of extraordinary impact, both at the large scale (say, the evolution of our species) and at short time depths (the detailed history of specific populations). Under a strong push of all biomedical sciences, evolutionary biologists, and geneticists have been able to read the information of the genome at a high speed and increasingly affordable prices, producing large amounts of data. Improvements of techniques and methods have allowed analyses of ancient DNA at an unthinkable scale just a handful of years ago. Thus, the amount of data is huge, increases day by day and will doubtless keep increasing in the future.

Furthermore, new computational and statistical methods are being developed to allow the genetic data to be better interpreted in terms of population past and history. Of course, these results must then be compared with those of other disciplines that also study human past, including archaeology, paleontology, and historical linguistics. These disciplines provide the framework on which individuals live and reproduce and thus should provide the basis for demographic models on which it is possible to test the genetic consequences of past events. In the case of Europe, it seems about time for geneticists, archaeologists, and paleodemographers to develop a mixed model, incorporating the possibility of different population processes affecting different geographic locations in Paleolithic, Neolithic, and successive times. It will then be a matter of years until enough genomic data are assembled, and one can move to testing whether this mixed model is able to account for the existing patterns better than alternative models. The original Neolithic and Paleolithic models have been very useful for many years to interpret the data and plan research, but it seems now both necessary and possible to refine them. At the very end, the more we know, the more we need to know.

References

Achilli A, Rengo C, Magri C, Battaglia V, Olivieri A, Scozzari R, Cruciani F, Zeviani M, Briem E, Carelli V, Moral P, Dugoujon JM, Roostalu U, Loogväli EL, Kivisild T, Bandelt HJ, Richards M, Villems R, Santachiara-Benerecetti AS, Semino O, Torroni A (2004) The molecular dissection of mtDNA haplogroup H confirms that the Franco-Cantabrian glacial refuge was a major source for the European gene pool. Am J Hum Genet 75:910–918

Allentoft ME, Sikora M, Sjögren KG, Rasmussen S, Rasmussen M, Stenderup J, Damgaard PB, Schroeder H, Ahlström T, Vinner L, Malaspinas AS, Margaryan A, Higham T, Chivall D, Lynnerup N, Harvig L, Baron J, Della Casa P, Dąbrowski P, Duffy PR, Ebel AV, Epimakhov A, Frei K, Furmanek M, Gralak T, Gromov A, Gronkiewicz S, Grupe G, Hajdu T, Jarysz R, Khartanovich V, Khokhlov A, Kiss V, Kolář J, Kriiska A, Lasak I, Longhi C, McGlynn G, Merkevicius A, Merkyte I, Metspalu M, Mkrtchyan R, Moiseyev V, Paja L, Pálfi G, Pokutta D, Pospieszny L, Price TD, Saag L, Sablin M, Shishlina N, Smrčka V, Soenov VI, Szeverényi V, Tóth G, Trifanova SV, Varul L, Vicze M, Yepiskoposyan L, Zhitenev V, Orlando L, Sicheritz-Pontén T, Brunak S, Nielsen R, Kristiansen K, Willerslev E (2015) Population genomics of bronze age Eurasia. Nature 522(7555):167–172

Antonio ML, Gao Z, Moots HM, Lucci M, Candilio F, Sawyer S, Oberreiter V, Calderon D, Devitofranceschi K, Aikens RC, Aneli S, Bartoli F, Bedini A, Cheronet O, Cotter DJ, Fernandes DM, Gasperetti G, Grifoni R, Guidi A, La Pastina F, Loreti E, Manacorda D, Matullo G, Morretta S, Nava A, Fiocchi Nicolai V, Nomi F, Pavolini C, Pentiricci M, Pergola P, Piranomonte M, Schmidt R, Spinola G, Sperduti A, Rubini M, Bondioli L, Coppa A, Pinhasi R, Pritchard JK (2019) Ancient Rome: a genetic crossroads of Europe and the Mediterranean. Science 366(6466):708–714. https://doi.org/10.1126/science.aay6826

Balaresque P, Bowden GR, Adams SM, Leung H–Y, King TE, Rosser ZH, Goodwin J, Moisan J–P, Richard C, Millward A, Demaine AG, Barbujani G, Previderè C, Wilson IJ, Tyler-Smith C, Jobling MA (2010) A predominantly Neolithic origin for European paternal lineages. PLoS Biol 19:e1000285

Barbujani G (2013) Genetic evidence for prehistoric demographic changes in Europe. Hum Hered 76(3–4):133–141

Barbujani G, Sokal RR, Oden NL (1995) Indo–European origins: a computer–simulation test of five hypotheses. Am J Phys Anthropol 96:109–132

Batini C, Hallast P, Zadik D, Delser PM, Benazzo A, Ghirotto S, Arroyo-Pardo E, Cavalleri GL, de Knijff P, Dupuy BM, Eriksen HA, King TE, López de Munain A, López-Parra AM, Loutradis A, Milasin J, Novelletto A, Pamjav H, Sajantila A, Tolun A, Winney B, Jobling MA (2015) Large-scale recent expansion of European patrilineages shown by population resequencing. Nat Commun 19(6):7152

Bouckaert R, Lemey P, Dunn M, Greenhill SJ, Alekseyenko AV, Drummond AJ, Gray RD, Suchard MA, Atkinson QD (2012) Mapping the origins and expansion of the indo-European language family. Science 337:957–960

Brace S, Diekmann Y, Booth TJ, van Dorp L, Faltyskova Z, Rohland N, Mallick S, Olalde I, Ferry M, Michel M, Oppenheimer J, Broomandkhoshbacht N, Stewardson K, Martiniano R, Walsh S, Kayser M, Charlton S, Hellenthal G, Armit I, Schulting R, Craig OE, Sheridan A, Parker Pearson M, Stringer C, Reich D, Thomas MG, Barnes I (2019) Ancient genomes indicate population replacement in Early Neolithic Britain. Nat Ecol Evol 3(5):765–771. https://doi.org/10.1038/s41559-019-0871-9. Epub 2019 Apr 15

Brandt G, Haak W, Adler CJ, Roth C, Szécsényi-Nagy A, Karimnia S, Möller-Rieker S, Meller H, Ganslmeier R, Friederich S, Dresely V, Nicklisch N, Pickrell JK, Sirocko F, Reich D, Cooper A, Alt KW, Genographic Consortium (2013) Ancient DNA reveals key stages in the formation of central European mitochondrial genetic diversity. Science 342(6155):257–261

Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The history and geography of human genes. Princeton University Press, Princeton

Currat M, Excoffier L (2005) The effect of the Neolithic expansion on European molecular diversity. Proc R Soc B 272:679–688

Dupanloup I, Bertorelle G, Chikhi L, Barbujani G (2004) Estimating the impact of prehistoric admixture on the Europeans' genome. Mol Biol Evol 21:1361–1372

Fernández E, Pérez-Pérez A, Gamba C, Prats E, Cuesta P, Anfruns J, Molist M, Arroyo-Pardo E, Turbón D (2014) Ancient DNA analysis of 8000 B.C. near eastern farmers supports an early neolithic pioneer maritime colonization of Mainland Europe through Cyprus and the Aegean Islands. PLoS Genet 10(6):e1004401

Fort J (2012) Synthesis between demic and cultural diffusion in the Neolithic transition in Europe. Proc Natl Acad Sci U S A 109:18669–18673

Fu Q, Rudan P, Pääbo S, Krause J (2012) Complete mitochondrial genomes reveal Neolithic expansion into Europe. PLoS One 7:e32473

Fu Q, Li H, Moorjani P, Jay F, Slepchenko SM, Bondarev AA, Johnson PL, Aximu-Petri A, Prüfer K, de Filippo C, Meyer M, Zwyns N, Salazar-García DC, Kuzmin YV, Keates SG, Kosintsev PA, Razhev DI, Richards MP, Peristov NV, Lachmann M, Douka K, Higham TF, Slatkin M, Hublin JJ, Reich D, Kelso J, Viola TB, Pääbo S (2014) Genome sequence of a 45,000-year-old modern human from western Siberia. Nature 514(7523):445–449

Fu Q, Hajdinjak M, Moldovan OT, Constantin S, Mallick S, Skoglund P, Patterson N, Rohland N, Lazaridis I, Nickel B, Viola B, Prüfer K, Meyer M, Kelso J, Reich D, Pääbo S (2015) An early modern human from Romania with a recent Neanderthal ancestor. Nature 524(7564):216–219

González-Fortes G, Tassi F, Trucchi E, Henneberger K, Paijmans JLA, Díez-Del-Molino D, Schroeder H, Susca RR, Barroso-Ruíz C, Bermudez FJ, Barroso-Medina C, Bettencourt AMS, Sampaio HA, Grandal-d'Anglade A, Salas A, de Lombera-Hermida A, Fabregas Valcarce R, Vaquero M, Alonso S, Lozano M, Rodríguez-Alvarez XP, Fernández-Rodríguez C, Manica A, Hofreiter M, Barbujani G (2019) A western route of prehistoric human migration from Africa into the Iberian Peninsula. Proc Biol Sci 286(1895):20182288. https://doi.org/10.1098/rspb.2018.2288

Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, Brandt G, Nordenfelt S, Harney E, Stewardson K, Fu Q, Mittnik A, Bánffy E, Economou C, Francken M, Friederich S, Pena RG, Hallgren F, Khartanovich V, Khokhlov A, Kunst M, Kuznetsov P, Meller H, Mochalov O, Moiseyev V, Nicklisch N, Pichler SL, Risch R, Rojo Guerra MA, Roth C, Szécsényi-Nagy A, Wahl J, Meyer M, Krause J, Brown D, Anthony D, Cooper A, Alt KW, Reich D (2015) Massive migration from the steppe was a source for indo-European languages in Europe. Nature 522(7555):207–211

Hewitt GM (2011) Quaternary phylogeography: the roots of hybrid zones. Genetica 139 (5):617–638

Higham T, Douka K, Wood R, Ramsey CB, Brock F, Basell L, Camps M, Arrizabalaga A, Baena J, Barroso-Ruíz C, Bergman C, Boitard C, Boscato P, Caparrós M, Conard NJ, Draily C, Froment A, Galván B, Gambassini P, Garcia-Moreno A, Grimaldi S, Haesaerts P, Holt B, Iriarte-Chiapusso MJ, Jelinek A, Jordá Pardo JF, Maíllo-Fernández JM, Marom A, Maroto J, Menéndez M, Metz L, Morin E, Moroni A, Negrino F, Panagopoulou E, Peresani M, Pirson S, de la Rasilla M, Riel-Salvatore J, Ronchitelli A, Santamaria D, Semal P, Slimak L, Soler J, Soler N, Villaluenga A, Pinhasi R, Jacobi R (2014) The timing and spatiotemporal patterning of Neanderthal disappearance. Nature 512(7514):306–309. https://doi.org/10.1038/nature13621

Keller A, Graefen A, Ball M, Matzas M, Boisguerin V, Maixner F, Leidinger P, Backes C, Khairat R, Forster M, Stade B, Franke A, Mayer J, Spangler J, McLaughlin S, Shah M, Lee C, Harkins TT, Sartori A, Moreno-Estrada A, Henn B, Sikora M, Semino O, Chiaroni J, Rootsi S, Myres NM, Cabrera VM, Underhill PA, Bustamante CD, Vigl EE, Samadelli M, Cipollini G, Haas J, Katus H, O'Connor BD, Carlson MR, Meder B, Blin N, Meese E, Pusch CM, Zink A (2012) New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. Nat Commun 3:698

Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Sudmant PH, Schraiber JG, Castellano S, Kirsanow K, Economou C, Bollongino R, Fu Q, Bos K, Nordenfelt S, de Filippo C, Prüfer K, Sawyer S, Posth C, Haak W, Hallgren F, Fornander E, Ayodo G, Babiker HA, Balanovska E, Balanovsky O, Ben-Ami H, Bene J, Berrada F, Brisighelli F, Busby GBJ, Cali F, Churnosov M, Cole DEC, Damba L, Delsate D, van Driem G, Dryomov S, Fedorova SA, Francken M, Gallego Romero I, Gubina M, Guinet JM, Hammer M, Henn B, Helvig T, Hodoglugil U, Jha AR, Kittles R, Khusnutdinova E, Kivisild T, Kucinskas V, Khusainova R, Kushniarevich A, Laredj L, Litvinov S, Mahley RW, Melegh B, Metspalu E, Mountain J, Nyambo T, Osipova L, Parik J, Platonov F, Posukh OL, Romano V, Rudan I, Ruizbakiev R, Sahakyan H, Salas A, Starikovskaya EB, Tarekegn A, Toncheva D, Turdikulova S, Uktveryte I, Utevska O, Voevoda M, Wahl J, Zalloua P, Yepiskoposyan L, Zemunik T, Cooper A, Capelli C, Thomas MG, Tishkoff SA, Singh L, Thangaraj K, Villems R, Comas D, Sukernik R, Metspalu M, Meyer M, Eichler EE, Burger J, Slatkin M, Kelso J, Reich D, Krause J (2014) Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature 513:409–413

Lordkipanidze D, Ponce de León MS, Margvelashvili A, Rak Y, Rightmire GP, Vekua A, Zollikofer CP (2013) A complete skull from dmanisi, georgia, and the evolutionary biology of early homo. Science 342(6156):326–331

Mathieson I et al (2018) The genomic history of southeastern Europe. Nature 555:197

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

Pääbo S (2015) The diverse origins of the human gene pool. Nat Rev Genet 16(6):313–314

Pinhasi R, Thomas MG, Hofreiter M, Currat M, Burger J (2012) The genetic history of Europeans. Trends Genet 28(10):496–505

Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, Moltke I, Rasmussen S, Stafford TW Jr, Orlando L, Metspalu E, Karmin M, Tambets K, Rootsi S, Mägi R, Campos PF, Balanovska E, Balanovsky O, Khusnutdinova E, Litvinov S, Osipova LP, Fedorova SA, Voevoda MI, DeGiorgio M, Sicheritz-Ponten T, Brunak S, Demeshchenko S, Kivisild T, Villems R, Nielsen R, Jakobsson M, Willerslev E (2014) Upper Palaeolithic Siberian genome reveals dual ancestry of native Americans. Nature 505(7481):87–91

Renfrew C (1987) Archaeology and language: the puzzle of Indoeuropean origins. Jonathan cape, London

Richards M, Macaulay VA, Hickey E, Vega E, Sykes BC, Guida V, Rengo C, Sellitto D, Cruciani F, Kivisild T, Villems R, Thomas M, Rychkov S, Rychkov O, Rychkov Y, Gölge M, Dimitrov D, Hill E, Bradley D, Romano V, Calì F, Vona G, Demaine A, Papiha S, Triantaphyllidis C, Stefanescu G, Hatina J, Belledi M, Di Rienzo A, Novelletto A, Oppenheim A, Nørby S, Al-Zaheri N, Santachiara-Benerecetti S, Scozari R, Torroni A, Bandelt

HJ (2000) Tracing European founder lineages in the near east mtDNA pool. Am J Hum Genet 67:1251–1276

Sampietro ML, Lao O, Caramelli D, Lari M, Pou R, Martí M, Bertranpetit J, Lalueza-Fox C (2007) Palaeogenetic evidence supports a dual model of Neolithic spreading into Europe. Proc Biol Sci 274:2161–2167

Sánchez-Quinto F, Schroeder H, Ramirez O, Ávila-Arcos MC, Pybus M, Olalde I, Velazquez AM, Prada Marcos ME, Vidal Encinas JM, Bertranpetit J, Orlando L, Gilbert MT, Lalueza-Fox C (2012) Genomic affinities of two 7,000-year-old Iberian hunter-gatherers. Curr Biol 22:1494–1499

Simoni L, Calafell F, Pettener D, Bertranpetit J, Barbujani G (2000) Geographic patterns of mtDNA diversity in Europe. Am J Hum Genet 66:262–278

Skoglund P, Malmström H, Raghavan M, Stora J, Hall P, Willerslev E, Gilbert MTP, Götherström A, Jakobsson M (2012) Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. Science 336:466–469

Soares P, Achilli A, Semino O, Davies W, Macaulay V, Bandelt HJ, Torroni A, Richards MB (2010) The archaeogenetics of Europe. Curr Biol 20(4):R174–R183

Torroni A, Bandelt HJ, Macaulay V, Richards M, Cruciani F, Rengo C, Martinez-Cabrera V, Villems R, Kivisild T, Metspalu E, Parik J, Tolk HV, Tambets K, Forster P, Karger B, Francalacci P, Rudan P, Janicijevic B, Rickards O, Savontaus ML, Huoponen K, Laitinen V, Koivumäki S, Sykes B, Hickey E, Novelletto A, Moral P, Sellitto D, Coppa A, Al-Zaheri N, Santachiara-Benerecetti AS, Semino O, Scozzari R (2001) A signal, from human mtDNA, of postglacial recolonization in Europe. Am J Hum Genet 69(4):844–852

Villanea FA, Schraiber JG (2019) Multiple episodes of interbreeding between Neanderthal and modern humans. Nat Ecol Evol 3(1):39–44. https://doi.org/10.1038/s41559-018-0735-8. Epub 2018 Nov 26

Zilhâo J (2001) Radiocarbon evidence for maritime pioneer colonization at the origins of farming in West Mediterranean Europe. Proc Natl Acad Sci USA 98:14180–14185

Chapter 9
Southeast Asia

Timothy Jinam

Abstract The Southeast Asian region has experienced many episodes of human migration. It may also have been witness to contact between early *Homo sapiens* with archaic humans like Denisovans. Archeological data indicates the presence of humans in the region as far back as 40,000 years ago. These early humans are thought to be the ancestors of Australian Aborigines and several Negrito groups. Migration driven by agriculture and associated with the Austronesian language also had a major impact on the human diversity in the region. This chapter will discuss those and other migration models from the standpoint of genetic data.

Keywords Southeast Asia · Population genetics · Human migration · Single nucleotide polymorphism · Negrito

9.1 Introduction

Southeast Asia (SEA) is known for its rich biodiversity, and that includes human beings. Throughout this chapter, the SEA region is defined to encompass the geographical boundaries of mainland SEA countries (Thailand, Vietnam, Laos, Cambodia, Myanmar, West Malaysia), and island SEA countries (Indonesia, East Malaysia, Brunei, Singapore, Philippines, East Timor). Island SEA was connected with the Asian mainland up to the Last Glacial Maximum (LGM) approximately 20,000 years ago (kya), forming a landmass known as Sundaland. The Wallace line (Fig. 9.1) separated Sundaland from another landmass called Sahul that made up of what is currently Australia and New Guinea. After the LGM, sea levels began to increase and flooded the coastal areas of Sundaland, gradually creating the coastlines seen today (Bellwood 2007; Glover and Bellwood 2004; Higham 1996; Pelejero et al. 1999).

T. Jinam (✉)
Division of Population Genetics, National Institute of Genetics, Mishima, Japan
e-mail: tjinam@nig.ac.jp

N. Saitou (ed.), *Evolution of the Human Genome II*, Evolutionary Studies,
https://doi.org/10.1007/978-4-431-56904-6_9

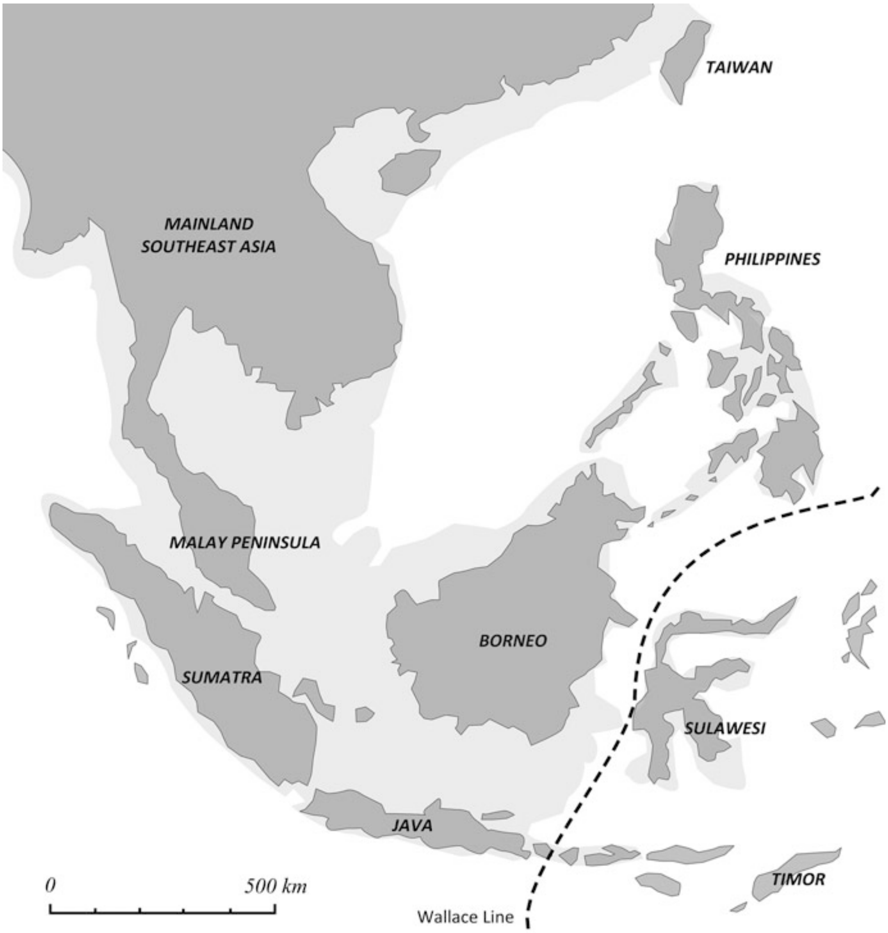


Fig. 9.1 Map of Southeast Asia. Light gray areas represent the coastline up to 12,000 years ago

These geographical changes have contributed to the vast diversity of human populations that currently reside in SEA. Although there may be more than a hundred recognized ethnic groups, they can be broadly grouped according to their language. The main language families spoken in SEA today are Tai-Kadai, Sino-Tibetan, Austro-Asiatic, and Austronesian. Of those, the latter two are the most widely spoken today. Ethnic groups that speak Tai-Kadai, Sino-Tibetan, and Austro-Asiatic languages are mostly restricted to mainland SEA, and in certain parts in India. With the exception of some groups in Eastern Indonesia that speak Papuan languages, almost all of the ethnic groups in island SEA speak Austronesian languages. The usage of Austronesian languages stretches from Madagascar off the coast of Africa to the multitude of islands in the tropical Pacific Ocean (Oceania), making it one of the most widespread linguistic group in the pre-modern era. There are 1257 Austronesian languages spoken today, and ones spoken in island SEA generally belong to the Malayo-Polynesian branch (Lewis et al. 2015).

The SEA region has experienced many episodes of human migrations, from the earliest migrants out of Africa to the fairly recent migrations driven by agriculture and trade. The earliest traces of humans found in Australia was dated to be approximately 50 kya (Roberts et al. 2005). In SEA one of the oldest human remains (approximately 40 kya) were found in Niah Cave in Borneo (Barker et al. 2007). These remains were linked to be the earliest exodus of anatomically modern humans from Africa approximately 60 kya. A coastal route via India and SEA was proposed before they eventually settled in Australia (Macaulay et al. 2005). Several indigenous groups such as the Australian Aborigines and Negritos (Sect. 9.3) are thought to be descendants of these early migrants. Another major migration event occurred during the Neolithic period, approximately 4–7 kya and coincided the rise of agriculture and the spread of the Austronesian language (Sect. 9.4). The more recent migrations were largely driven by trade and other socio-economic activities. These activities introduced both culture and humans from Arabia, India, China, and Europe to the region and influenced the landscape of human population diversity in the region.

9.2 Contact with Archaic Humans in Southeast Asia

Before there were anatomically modern humans in SEA, there were already traces of archaic humans in the region. From the genus *Homo*, *H. erectus* fossils were discovered in Indonesia and dubbed the Java man. The fossils were dated to be approximately one million years old, making it the oldest hominin remains found in SEA (Swisher et al. 1994). No fossils from other archaic humans such as *Homo neanderthalensis* or *Homo heidelbergensis* were found in SEA.

However, that does not exclude the possibility of contact and genetic exchange between archaic and anatomically modern humans. Although the first analysis of mitochondrial DNA (mtDNA) from a Neanderthal sample showed that they were an outgroup to all other modern humans (Green et al. 2008), the advent of Next Generation Sequencing (NGS) technology has allowed us to find genetic loci which are shared with these archaic humans. Introgression between archaic and modern humans was reported by Green et al. (2010) who generated whole genome sequence of a Neanderthal and found more shared alleles with Eurasians than Africans. This implies that any admixture event(s) between Neanderthals and modern humans occurred between the ancestors of all non-Africans. Given the distribution of Neanderthal remains found so far, the introgression probably occurred in what is currently Middle East, before the divergence of Europeans and East Asians.

Neanderthals are not the only archaic humans to leave their genetic imprints in modern humans. Reich et al. (2010) extracted ancient DNA from a finger bone excavated from the Denisova cave in Siberia. The resulting nuclear genome sequence from the Denisovan sample revealed higher allele sharing with Papuans than with other Eurasians and Africans. A follow-up study using genome-wide single nucleotide polymorphisms (SNP) (Reich et al. 2011) in SEA populations showed the presence of Denisovan introgression in Philippine Negritos, Australian Aborigines, East Indonesians, and Oceanians. Interestingly, populations from the west side of island SEA which included Malaysian Negritos did not show any significant Denisovan genetic components. This observation suggests that the Denisovans may have a broader distribution within Asia and that the introgression event could have occurred before the split between Papuans and Australian Aborigines with other South and East Asians. However, the exact geographical location where the introgression event could have happened remain unclear. It is clear however, that modern humans were not alone as they migrated out of Africa and into SEA.

9.3 Ancient Migrants and Hunter Gatherers

The earliest migration of modern humans out of Africa is believed to have occurred at least 50 kya, going along the coastal route via SEA and into Australia. These early migrants are thought to be the ancestors of the Australian Aborigines and several Negrito groups in SEA. One reason for that is due to the shared physical characteristics of these groups. They include darker skin, frizzy or curly hair and generally shorter stature. In this section, I will be discussing about the Negrito groups in SEA, which include Malaysian and Philippine Negritos as well as the Andamanese and Papuans.

The Negritos in Malaysia are mainly restricted to the northern part of the Malay Peninsula, close to the border with Thailand. Also referred to as *Semang* in some literature, they are grouped together with other indigenous groups in the rural and coastal regions of the Malay Peninsula as *Orang Asli*, meaning original peoples in the Malay language. There are six known Negrito subgroups in Malaysia, and all of them currently speak the Aslian branch of Austro-Asiatic languages (Lewis et al. 2015). In contrast, the Negrito groups in the Philippines speak the Malayo-Polynesian branch of the Austronesian language. They are scattered throughout the Luzon, Mindanao, and Palawan islands. There are other indigenous groups in Asia that are sometimes grouped as Negritos, and they include the Andamanese, Papuans, and Melanesians. The term "Australoid" has also been used to refer to these Negritos and the Australian Aborigines. Other than some shared phenotypes, these Negritos have long been hunter gatherers (and some still remain to this day). Various physical anthropological studies of these groups have found links between them, suggesting a common origin that traces back to the first migrants out of Africa.

The genetic studies of the Negritos have until recently been piecemeal. One of the earliest genetic surveys were done by Omoto et al. (1978) on the Philippine Negritos. A later paper that summarized data from 120 autosomal loci showed that the Philippine Negritos are closer to other Southeast Asians than they are to Papuans, Australian Aborigines, and Africans (Omoto 1984). An early study of mtDNA in the

Aeta, a Philippine Negrito group, showed close affinity with East Asians (Hanihara et al. 1988). These early studies suggest that the Negritos have experienced substantial admixture with their non-Negrito neighbors. As more Negrito groups were sampled, the picture that arose from mtDNA analyses appears to support a long history of these Negrito groups, possibly dating back to the early Out of Africa dispersal. Each Negrito group appears to have basal mtDNA haplogroups. For example, haplogroups M31 and M32 are restricted to Andamanese (Thangaraj et al. 2003, 2005), M21a in the Malaysian Negritos (Hill et al. 2006; Jinam et al. 2012), and N11 in the Philippine Negritos (Gunnarsdottir et al. 2010). These haplogroups were estimated to date around 50–30 kya, based on the molecular clock method. The basal position as well as the old age of these Negrito mtDNA haplogroups suggests that the Negritos were part of the early, southern dispersal from Africa after which they became isolated in various points in SEA (Macaulay et al. 2005). While the mtDNA may be useful to infer demographic events, it is still considered as a single genetic locus. Hence the information that we can infer from it is smaller than what we could achieve from the nuclear genome.

Analysis of autosomal loci in the past has been hampered by low throughput, whereby only a few polymorphic loci were genotyped. That has changed with the advancement of the SNP microarray technologies which allowed for typing of millions of biallelic SNP in a single sample. The first large-scale genetic survey of SEA populations was undertaken by the HUGO Pan-Asian SNP Consortium (PASNP). Using approximately 50 k SNPs, the results showed that the Malaysian, Philippine Negritos, and Papuans were genetically distinct from one another. This and other follow-up studies (Jinam et al. 2013) further showed that these Negrito groups have experienced fairly recent admixture with their non-Negrito neighbors, as opposed to being in continued state of isolation.

These genetic analyses on the various Negrito groups in SEA indicate that they have been ubiquitous in the region and that they still retain the hallmarks of the ancient migration, as shown by their mtDNA haplogroup ages. While they may have been isolated groups in the past, it is not the case now since SNP studies have shown that the Negritos have experienced gene flow from their neighbors. Contact with their neighbors not only resulted in genetic exchange, but also possibly linguistic change as seen in the Philippine Negrito groups who now speak Austronesian languages like their neighbors.

9.4 Human Diversity in Mainland Southeast Asia

Mainland Southeast Asia (mSEA) is an important waypoint for human migration, connecting South Asia (India, Bangladesh, Pakistan, etc.), East Asia (China, etc.) with the rest of island Southeast Asia. It is established that the Australian Aboriginals and Negritos are linked with the earliest migration wave out of Africa, but the migration histories of mainland SEA populations have been rather ambiguous. A survey of approximately 50,000 SNPs in 73 Asian populations (HUGO Pan-Asian

SNP Consortium 2009) suggested that East Asian populations were derived from a single wave of migration from Southeast Asia, based on phylogenetic and simulation-based analyses. Another study suggested a two-wave migration model in to SEA, based on whole genome sequence analysis of Aboriginal Australians, Han Chinese, and Europeans (Rasmussen et al. 2011). This issue can be resolved by analyzing whole genome sequences in a larger diversity of samples from Southeast Asia.

In terms of linguistic diversity, the languages in mSEA belong to the Austro-Asiatic, Tai-Kadai, Sino-Tibetan, and Hmong-Mien language families (Lewis et al. 2015). Sino-Tibetan is largely spoken in Myanmar, Tai-Kadai in Thailand and Laos, Austro-Asiatic in Cambodia and southern Vietnam, and Hmong-Mien in northern Vietnam (Glover and Bellowood 2004; Lewis et al. 2015). It has been proposed that Austro-Asiatic and Austronesian languages share a common ancestor, called Austric (Schmidt 1906) and are spoken by the indigenous populations in mSEA. Subsequently, other language groups were introduced from South China and Tibet.

The analysis of the maternal mtDNA in several mSEA populations (Peng et al. 2010; Zhang et al. 2013; Summerer et al. 2014) demonstrates the presence of basal haplogroups that have very old coalescent ages, suggesting long term presence of human populations in that region. Analysis of genome-wide autosomal markers in mSEA populations has been quite limited, but some recent studies have shown presence of substructure within Thai populations (Listman et al. 2011; Wangkumhang et al. 2013). Similar kind of studies need to be undertaken for other populations in mSEA in order to have a better understanding on the human migration events in that region.

9.5 Impact of Human Expansions Driven by Agriculture

As briefly mentioned in Sect. 9.1, almost all island SEA populations speak the Austronesian language. Taiwan has been proposed to be the origin of the Austronesian language, based on linguistic phylogenies. The so-called Out of Taiwan or Austronesian expansion took a southern route from Taiwan via the Philippine islands and from there spread westwards all the way to Madagascar and eastwards to the Pacific islands. This migration model has been well supported not just by linguistics, but also by archeological records. The Austronesian expansion has also been suggested to introduce rice agriculture to the SEA region.

Genetic support for this model came from mtDNA analysis, particularly the dispersal pattern of haplogroup B4a1. The B4a haplogroup is defined by a 9 bp deletion, sometimes referred to as the “Polynesian motif.” Phylogenetic analysis of the B4a haplogroup indicated that the basal type is found in Taiwanese aboriginal groups, whereas the derived types are mostly found in Polynesians. The ages of the haplogroups were estimated to be approximately 4–6 kya, consistent with the estimates from archeological data. Several other haplogroups were proposed to be

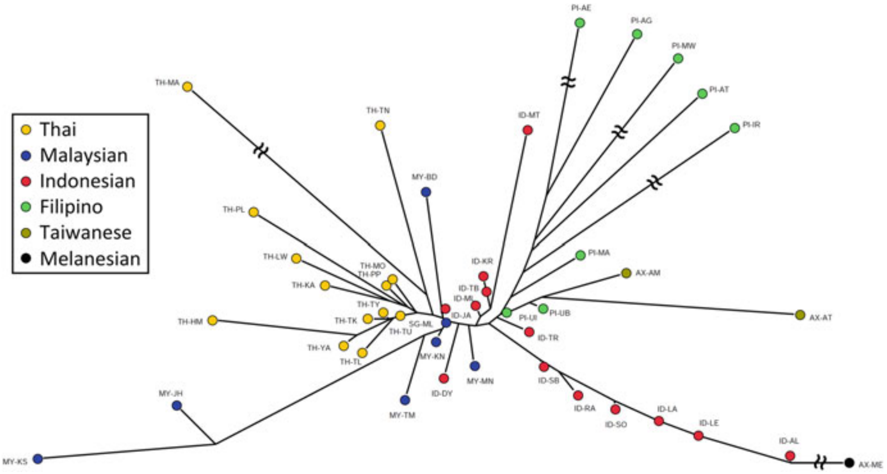


Fig. 9.2 Unrooted Neighbor-Joining tree of Southeast Asian populations from the HUGO Pan-Asian SNP Consortium. The node labels are following the original used in The HUGO Pan-Asian SNP Consortium (2009). Long branch lengths are truncated with the "≈" symbol

markers for the Austronesian expansion, due to the ancestral state found in Taiwanese aborigines, such as M7c3c, D6, Y2, and F3.

Other schools of thought argued for earlier episodes of migration from mainland SEA that happened before the Austronesian expansion (Hill et al. 2006, 2007; Karafet et al. 2010), based on mtDNA and Y-chromosomal analyses. Jinam et al. (2012) proposed an "early train" migration from mainland SEA via the Malay peninsula and into Borneo. This was based on the phylogeny of mtDNA haplogroups found in the Malay peninsula and Borneo that had ancestral types in mainland SEA. The ages of those haplogroups were estimated to be approximately 10–30 kya, which was before the Austronesian expansion. Analysis of autosomal SNP further supplemented the idea of a West-East divide between island SEA populations. The unrooted Neighbor-Joining tree constructed using pairwise Fst distances between Southeast Asian populations from the HUGO Pan-Asian SNP Consortium (2009) in Fig. 9.2 demonstrates this. On the left side of the tree are mostly Thai and Malaysian populations whereas indigenous Taiwanese and Filipinos are mostly on the right side, suggesting a dichotomy that roughly corresponds to the geographical location of these populations. Some Indonesian populations from the east side of Java island (ID-SB, ID-RA, ID-SO, ID-LA, ID-LE, and ID-AL in Fig. 9.2) seem to form a "chain" linking to Melanesians. This pattern suggests a gradient of Melanesian admixture within these Eastern Indonesian populations. A study by Lipson et al. (2014) further supports the idea of dichotomy within island SEA populations by showing mSEA genetic components in Indonesian and Malaysian populations, but not in Filipino groups. These findings have shown that the Austronesian expansion from Taiwan did not account for all the genetic diversity seen in Austronesian groups in SEA.

The impact of the Austronesian expansion however has not been diminished by these other findings, given the ubiquity of the language group in SEA. Languages are also more readily transmissible than genes. The impact of the Austronesian expansion is the introduction of rice agriculture, but also led to the suppression of the extant Negrito groups that were predominantly hunter gatherers. The Austronesian speaking populations are therefore as genetically diverse as the language they speak.

9.6 Concluding Remarks and Perspectives

The genetic studies described above mostly relied on the sequencing and genotyping of uniparental markers that include mtDNA and the Y-chromosome. Although these markers remain attractive to those studying population genetics, ultimately they are just considered a single loci and thus the information that can be garnered will be limited. The use of genome-wide SNP genotyping has been on the rise in recent years and has provided new insight into views of population substructure and admixture. However, SNP typing may soon be supplanted by whole genome sequencing technology. As the price for NGS goes down, the age where we could have the whole genome information of an individual could soon be upon us. Another advantage of NGS over chip-based SNP typing is that it is free from ascertainment bias, whereby polymorphic loci were selected (ascertained) from certain populations. This kind of sampling bias may lead to inaccurate statements about the actual population genetic diversity. The 1000 Genomes Project (The 1000 Genomes Project Consortium 2012) is an example of a large-scale sequencing project that will benefit those studying human population genetics. As more sequence data from populations become available, we can add more pieces to the jigsaw puzzle known as the origin of man.

References

Barker G, Barton H, Bird M et al (2007) The “human revolution” in lowland tropical Southeast Asia: the antiquity and behavior of anatomically modern humans at Niah Cave (Sarawak, Borneo). J Hum Evol 52:243–261

Bellwood P (2007) Prehistory of the Indo-Malaysian Archipelago. ANU E Press, Canberra

Glover I, Bellwood PS (2004) Southeast Asia: from prehistory to history. Routledge, New York

Green RE, Malaspinas A-S, Krause J et al (2008) A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. Cell 134:416–426

Green RE, Krause J, Briggs AW et al (2010) A draft sequence of the Neandertal genome. Science 328:710–722

Gunnarsdottir ED, Li M, Bauchet M et al (2010) High-throughput sequencing of complete human mtDNA genomes from the Philippines. Genome Res 21:1–11

Harihara S, Saitou N, Hirai M et al (1988) Mitochondrial DNA polymorphism among five Asian populations. Am J Hum Genet 43:134–143

Higham CFW (1996) A review of archaeology in mainland Southeast Asia. J Archaeol Res 4:3–49

Hill C, Soares P, Mormina M et al (2006) Phylogeography and ethnogenesis of aboriginal southeast Asians. Mol Biol Evol 23:2480–2491

Hill C, Soares P, Mormina M et al (2007) A mitochondrial stratigraphy for island southeast Asia. Am J Hum Genet 80:29–43

Jinam TA, Hong LC, Phipps ME et al (2012) Evolutionary history of continental southeast asians: Early train hypothesis based on genetic analysis of mitochondrial and autosomal DNA data. Mol Biol Evol 29:3513–3527

Jinam TA, Phipps ME, Saitou N (2013) Admixture patterns and genetic differentiation in negrito groups from West Malaysia estimated from genome-wide SNP data. Hum Biol 85:173–188

Karafet TM, Hallmark B, Cox MP et al (2010) Major east–west division underlies Y chromosome stratification across Indonesia. Mol Biol Evol 27:1833–1844

Lewis PM, Simons GF, Fennig CD (eds) (2015) Ethnologue: languages of the World, 18th edn. SIL International, Texas

Lipson M, Loh P-R, Patterson N et al (2014) Reconstructing Austronesian population history in Island Southeast Asia. Nat Commun 5:4689

Listman JB, Malison RT, Sanichwankul K et al (2011) Southeast asian origins of five hill tribe populations and correlation of genetic to linguistic relationships inferred with genome-wide SNP data. Am J Phys Anthropol 308:300–308

Macaulay V, Hill C, Achilli A et al (2005) Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. Science 308:1034–1036

Omoto K (1984) The Negritos: genetic origins and microevolution. Acta Anthropogenet 8:137–147

Omoto K, Misawa S, Harada S et al (1978) Population genetic studies of the Philippine Negritos. I. A pilot survey of red cell enzyme and serum protein groups. Am J Hum Genet 30:190–201

Pelejero C, Kienast M, Wang L, Grimalt JO (1999) The flooding of Sundaland during the last deglaciation: imprints in hemipelagic sediments from the southern South China Sea. Earth Planet Sci Lett 171:661–671

Peng M-S, Quang HH, Dang KP et al (2010) Tracing the Austronesian footprint in mainland southeast Asia: a perspective from mitochondrial DNA. Mol Biol Evol 27:2417–2430

Rasmussen M, Guo X, Wang Y et al (2011) An Aboriginal Australian genome reveals separate human dispersals into Asia. Science 334:94–98

Reich D, Green RE, Kircher M et al (2010) Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468:1053–1060

Reich D, Patterson N, Kircher M et al (2011) Denisova admixture and the first modern human dispersals into southeast Asia and Oceania. Am J Hum Genet 89:516–528

Roberts RG (Richard G, Morwood MJ (Mike J., Westaway KE (2005) Illuminating southeast asian prehistory: new archaeological and paleoanthropological frontiers for luminescence dating. Asian Perspect 44:293–319.

Schmidt PW (1906) Die Mon-Khmer-Vilker, ein Bindeglied zwischen V61kern Zentralasiens und Austronesiens. Archiv der Anthropologie (Braunschweig) 5:59–109

Summerer M, Horst J, Erhart G et al (2014) Large-scale mitochondrial DNA analysis in Southeast Asia reveals evolutionary effects of cultural isolation in the multi-ethnic population of Myanmar. BMC Evol Biol 14:17

Swisher CC, Curtis GH, Jacob T et al (1994) Age of the earliest known hominids in Java, Indonesia. Science 263:1118–1121

Thangaraj K, Singh L, Reddy AG et al (2003) Genetic affinities of the Andaman Islanders, a vanishing human population. Curr Biol 13:86–93

Thangaraj K, Chaubey G, Kivisild T et al (2005) Reconstructing the origin of Andaman Islanders. Science 308:996

The 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1092 human genomes. Nature 491:56–65

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

Wangkumhang P, Shaw PJ, Chaichoompu K et al (2013) Insight into the peopling of Mainland Southeast Asia from Thai population genetic structure. PLoS One 8:1–12

Zhang X, Qi X, Yang Z et al (2013) Analysis of mitochondrial genome diversity identifies new and ancient maternal lineages in Cambodian aborigines. Nat Commun 4:1–11

Chapter 10
Australia and Oceania

Ana T. Duggan and Mark Stoneking

Abstract We review the evidence for the settlement of Australia and Oceania focusing on population history prior to the arrival of Europeans. Topics discussed include archaic admixture in the ancestors of Aboriginal Australians and New Guinea Highlanders and gene flow from India to Australia. We review in detail the dual-migration history of Near Oceania and the subsequent colonisation of Remote Oceania, the evidence for contact between South America and Polynesia, and make special reference to the population of Santa Cruz.

Keywords Population genetics · Demographic history · Oceania · Australia · Austronesian expansion · Archaic admixture

10.1 Geography and Ecology of Australia and Oceania

During periods of lower seas levels, such as during glacial maxima, much of Island Southeast Asia was contiguous with mainland Asia forming a continent known as Sunda. Across the Wallace Line, the island of New Guinea, Australia and Tasmania were similarly connected forming a continent known as Sahul. It is during this period of lower sea levels that the first anatomically modern humans are believed to have reached Australia and New Guinea. Though the presence of Sunda and Sahul would have reduced water crossings, it is clear that some form of boating or rafting would have been necessary as the Wallace Line was never closed. This in turn implies that the settlement of Sahul was intentional and directed, not passive (Fig. 10.1).

Despite their historical connection, Australia has a very distinct geology and ecology compared to New Guinea. Many of these differences are driven by their difference in latitude. While New Guinea is largely tropical, Australia spans more than 30° of latitude, is comparatively flat and has large areas of desert and rock

A. T. Duggan · M. Stoneking (✉)
Department of Evolutionary Genetics, Max Plank Institute for Evolutionary Anthropology, Leipzig, Germany
e-mail: stoneking@eva.mpg.de

N. Saitou (ed.), *Evolution of the Human Genome II*, Evolutionary Studies,
https://doi.org/10.1007/978-4-431-56904-6_10

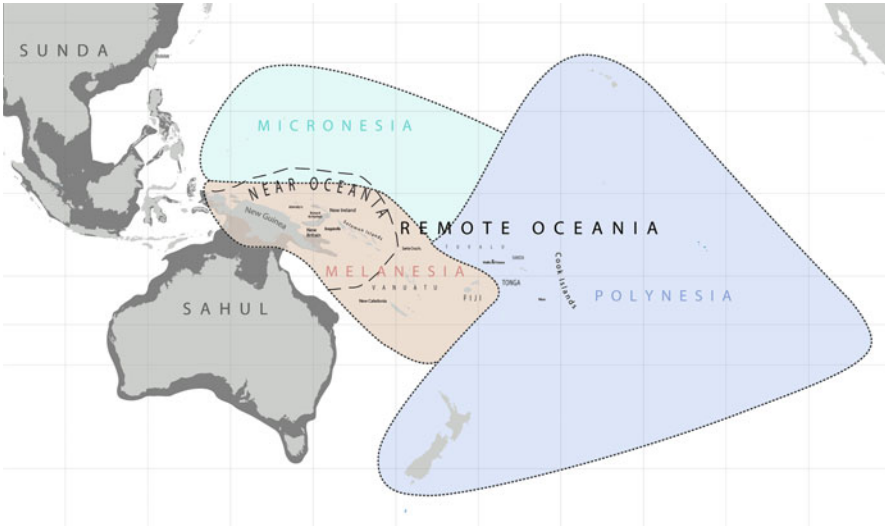


Fig. 10.1 Map indicating relative positions of Melanesia, Micronesian and Polynesia with respect to Near and Remote Oceania. Last glacial maximum boundaries of the former continents of Sunda and Sahul are indicated in dark grey. Used with permission from Duggan (2014)

(White and O'Connell 1982). There are large areas covered in grasslands and forest, though rainforests are minimal and geographically restricted, and overall rainfall is quite low and highly seasonal (White and O'Connell 1982). From the perspective of biodiversity, Australia is best known for its unique assortment of faunal species; the sharp contrast and diversity were first noted by Alfred Russell Wallace, with marsupial species, such as the kangaroo, and monotremes found almost exclusively in Australia and New Guinea.

While boundaries such as Melanesia, Polynesia and Micronesia may be more familiar, they fail to capture cohesive groups biologically, archaeologically or linguistically (Kirch 2000). These terms are increasingly replaced with Near Oceania and Remote Oceania which do reflect a shared history, both archaeologically and biologically (Green 1991) (Fig. 10.1). Near Oceania includes the island of New Guinea and surrounding islands such as those of the Bismarck Archipelago and the Solomon Islands Archipelago, extending to the islands of Makira (San Cristobal) and Santa Ana. In practical terms, this boundary represents the points beyond which islands are no longer inter-visible and it becomes necessary to complete longer crossings of open water. This likely proved to be a technological impediment as superior water craft and navigational skills would be crucial to cross beyond Near Oceania into Remote Oceania. Many of the islands in Near Oceania are quite large and topographically varied, consisting of distinct coastal regions, interior mountain ranges and rain forests. During the periods of lower sea levels these islands remained separate from Sahul; however, much of the Solomon Islands Archipelago was connected as a single larger island called Greater Bougainville or Greater Bukida

(Kirch 2000; Spriggs 1997; Walter and Sheppard 2009). With their larger size, heterogeneous landscapes and plentiful microhabitats, Near Oceanian islands generally support a larger and more varied fauna, including several endemic mammal species, and also present more arable land providing opportunities for multiple subsistence patterns (White and O'Connell 1982). Remote Oceania, in contrast, is composed of clusters of smaller islands mostly of volcanic origin or coral atolls. Beyond those previously included in the grouping of Polynesia and Micronesia, Remote Oceania also includes the islands of Santa Cruz and the Reef Islands, Vanuatu, New Caledonia and Fiji. The geology and topography of Remote Oceania result in less varied habitats and the large expanses of open water necessary to travel between islands leads to a sharp reduction in faunal diversity (Kirch 2000). There are no endemic mammals in Remote Oceania, though some have colonised the islands either as passive stowaways or as intentional introductions by humans; these islands tend to have few crops and their populations are much more dependent on the sea for sustenance.

10.2 Australia

Archaeological evidence suggests that Sahul was settled by about 45,000 years ago (kya) with sites in both Australia and New Guinea dating to this time frame and suggesting the possibility of a single migration into Sahul (O'Connell and Allen 2004; Summerhayes et al. 2010). Genetic evidence suggests that the initial inhabitants of Sahul were descended from an early Out-of-Africa migration which likely followed a southern route to Sahul and that this migration was separate from that which gave rise to modern Asian populations (Rasmussen et al. 2011; Pugach et al. 2013; Reich et al. 2011; Wollstein et al. 2010). Furthermore, with a notable exception detailed below, the populations descended from these early settlers of Sahul probably remained isolated from all other populations for 15,000–30,000 years (ky) (Rasmussen et al. 2011).

Genetic research into the history of contemporary Aboriginal Australian populations has been hampered, in part, by a deep distrust of medical and academic researchers due to a historical sense of exploitation (van Holst Pellekaan 2000). In the DNA era of molecular anthropology, there have been several studies involving mitochondrial DNA (mtDNA) and some Y-chromosome polymorphisms; however, samples sizes have often been small and population coverage unfortunately low. These studies have shown that there are limited mtDNA lineages present amongst Aboriginal Australians; some of these appear to be unique to Australia (such as haplogroups O and S), whereas others (such as haplogroups P and Q) form deep branches that are shared with New Guinea Highlanders (Ingman and Gyllensten 2003; van Holst Pellekaan et al. 1998, 2006; Huoponen et al. 2001; Redd and Stoneking 1999; Hudjashov et al. 2007). Notably, Aboriginal Australian mtDNA and Y-chromosome lineages comprise some of the deepest rooting lineages found outside of Africa, once more reinforcing the early Southern route hypothesis, and suggesting that these are among the oldest continuous populations found outside of Africa (Hudjashov et al. 2007; van Holst Pellekaan et al. 2006; Kayser et al. 2001, 2006). However, within the past few years, there have been several significant contributions to the study of Aboriginal Australian population history employing whole-genome methods. Rasmussen et al. (2011) published the first whole-genome sequence of an Aboriginal Australian, analysing a museum specimen of hair collected approximately 100 years previously; this hair was clearly collected well after the first European settlement of Australia in 1788 but pre-dates most recent admixture with Europeans. Despite its relatively young age, this hair sample suffered from many problems associated with ancient DNA, such as short fragment lengths and deamination at the ends of each fragment (Rasmussen et al. 2011; Pääbo et al. 2004); however, genome coverage of 6.4X was achieved nevertheless. The mtDNA and Y-chromosome haplogroups for the sample were consistent with lineages already found in Australia, suggesting that there were no issues with contamination despite being a heavily handled museum specimen, and also providing some preliminary indication that there was no recent European admixture in this individual (Rasmussen et al. 2011).

In a Principle Components Analysis (PCA) comparison with populations from across the globe, the ancient Aboriginal Australian genome clustered with samples from the Highlands of Papua New Guinea, further supporting their shared ancestry, and was in next closest proximity with a population from Bougainville (an island politically part of Papua New Guinea but geographically part of the Solomon Islands archipelago), and the Aeta, a ‘Negrito’ population from the Philippines (Rasmussen et al. 2011). These findings are further supported by an ADMIXTURE analysis in which the Aboriginal Australian genome is best represented as having primarily near equal Highland New Guinea and Bougainville-like ancestries (Rasmussen et al. 2011).

With respect to population history, one of the more interesting results from Rasmussen et al. (2011) is the estimation of population split times calculated from linkage disequilibrium and allele frequencies between pairs of populations. With their method of estimation, Rasmussen et al. (2011) concluded that the ancestors of Aboriginal Australians and New Guinea Highlanders have been largely separated from Eurasian populations since 62–75 kya (with some cases of admixture described below), while European and Asian populations split only 25–38 kya.

The Aboriginal Australian genome paper was further augmented by the publication of genotyping results from 12 modern Aboriginal Australians (Pugach et al. 2013). Apart from a greater sample size, the genotyping study also benefitted from a greater array of Asian and Southeast Asian populations for comparison. An earlier study had a similar design but was more limited in terms of insights into Aboriginal Australian history, as almost all of the Aboriginal Australians included were known to have recent, mostly European, admixture and in fact had on average only 64% Aboriginal Australian ancestry (McEvoy et al. 2010). Pugach et al. (2013) confirmed the shared ancestry of Aboriginal Australians and New Guinea Highlanders, and moreover showed that the Mamanwa, a ‘Negrito’ group from the Philippines—similar to the Aeta mentioned above—were equally related to both Aboriginal

Australians and New Guinea Highlanders, suggesting they had once been part of the same early Southern route population and that the Mamanwa split from this group before the separation of Australian Aboriginals and New Guinea Highlanders (Pugach et al. 2013). Curiously, in an examination of linkage disequilibrium patterns, it was observed that the patterns of linkage disequilibrium in the Mamanwa suggested that they had undergone a rather recent bottleneck and the New Guinea Highlanders a more ancient bottleneck (Pugach et al. 2013). The presence of an ancient bottleneck in New Guinea Highlanders but absent in the Mamanwa suggested that this bottleneck occurred after the split of the Mamanwa lineage from New Guinea Highlanders and Aboriginal Australians. The absence of evidence for such a bottleneck in Aboriginal Australians suggests that either they were not affected by the same event as the New Guinea Highlanders, i.e. the bottleneck in New Guinea Highlanders occurred after the split from Aboriginal Australians, or that there was less isolation in Aboriginal Australian populations compared to New Guinea Highlanders (Pugach et al. 2013).

Indeed, it seems Aboriginal Australians were not as isolated as New Guinea Highlanders and instead, around 4230 years ago, received an influx of genes from a population most closely related to present-day Dravidian speakers in Southern India (Pugach et al. 2013). This finding was confirmed through several different analyses, including PCA, ADMIXTURE and TreeMix analyses, which are all calculated under different algorithms and methods and yet all demonstrated clear signs of admixture between Aboriginal Australians and southern Indians (Pugach et al. 2013). The TreeMix analysis estimated the amount of gene flow from India to represent approximately 11% of the Aboriginal Australian genome (Pugach et al. 2013) (Fig. 10.2). This finding, while surprising, has further support. At approximately the same time as this admixture is estimated to have occurred, several changes take place in the Australian archaeological record: the stone tool technology changes, as does the manner in which plants are processed, and the dingo makes its first appearance (Pugach et al. 2013 and references therewithin). Beyond archaeological evidence, there had been genetic hints of this admixture event as well; the authors of the ancient Aboriginal Australian genome study (Rasmussen et al. 2011) noted a tiny component of Indian ancestry in their ADMIXTURE analyses as well but dismissed it as an artefact where no appropriate ancestral population was provided to the algorithm (understandably so when working with a single representative sample). And previous studies on both mtDNA and Y-chromosome variation in Australia found closer than expected affinities between Aboriginal Australians and the Indian subcontinent (Redd and Stoneking 1999; Redd et al. 2002).

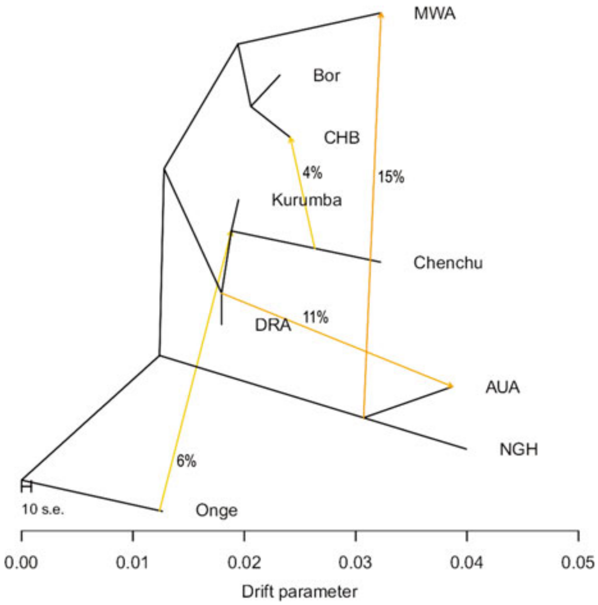


Fig. 10.2 TreeMix analysis of population relationships. Percentages indicate the contribution from source to recipient populations. Modified with permission from Pugach et al. (2013)

10.3 Oceania

10.3.1 Near Oceania in the Pleistocene

Many of the early archaeological sites in New Guinea are found in the southeast of the island. The oldest sites date to approximately 45 kya; however, there are sites of a similar age found on the islands of New Britain and New Ireland in the Bismarck Archipelago, suggesting that there was further rapid and intentional movement across the strait which separates the islands from New Guinea (Summerhayes et al. 2010; Leavesley et al. 2002; Leavesley and Chappell 2004; Groube et al. 1986). These populations also managed to spread at least as far as Buka, at the northernmost tip of the Solomon Islands Archipelago, and Manus in the Admiralty Islands by 28 and 12 kya, respectively (Wickler and Spriggs 1988; Fredericksen et al. 1993) (Fig. 10.3). Curiously, despite ample evidence of settlement in Buka and given the contiguous nature of Greater Bougainville including Buka, there are no other archaeological sites of similar antiquity in the Solomon Islands. The only other evidence of human habitation in the Solomon Islands prior to the arrival of the Austronesian expansion, to be discussed in Sect. 10.3.2, is a site on the island of Guadalcanal dated to approximately 6 kya (Roe 1993).

Much of the existing genetic evidence from Near Oceania comes from uniparental markers which seem to indicate that, despite a common ancestry, the early settlers formed populations which were largely isolated. From the maternal perspective, these older populations are characterised by autochthonous haplogroups such as Q, P, M27, M28 and M29, some of which show extreme geographic specificity

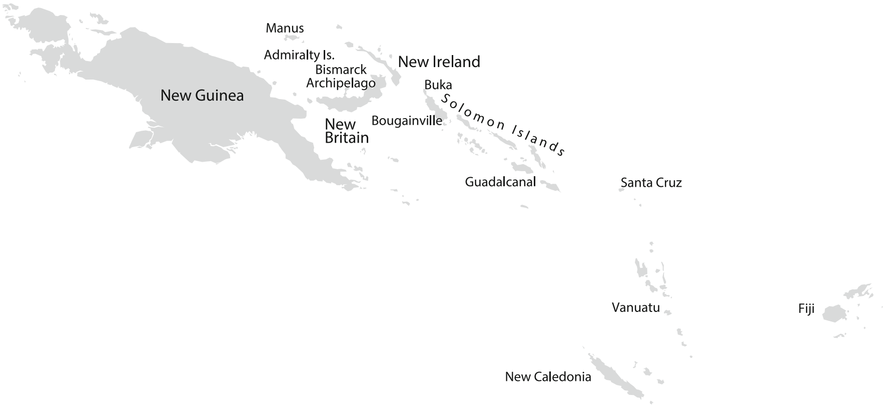


Fig. 10.3 Map indicating relative location of populations referred to in the text

(Friedlaender et al. 2005, 2007; Merriwether et al. 2005; Duggan et al. 2014). Studies of mtDNA indicate that autochthonous haplogroups are often exclusive to particular populations or regions. Moreover, specific sequences (haplotypes) within these haplogroups are rarely shared between populations and are frequently defined by multiple mutations, suggesting prolonged periods of separation have allowed drift to occur (Friedlaender et al. 2007; Duggan et al. 2014). Frequencies of these haplogroups are higher in New Guinea, particularly in the Highlands which appear to have been largely unaffected by the Austronesian expansion, as well as the Bismarck Archipelago and Bougainville, in keeping with archaeological data which suggests that these areas have deeper settlement histories (Friedlaender et al. 2005, 2007; Merriwether et al. 2005; van Oven et al. 2014; Duggan et al. 2014). Even within the island of New Guinea there are large differences observed between populations in the highlands and the coast, attributable to the effect of the Austronesian expansion on coastal regions (Stoneking et al. 1990; Redd et al. 1995; Redd and Stoneking 1999; Kayser et al. 2006; van Oven et al. 2014; Ingman and Gyllensten 2003; Tommaseo-Ponzetta et al. 2002).

While many populations show a strong impact of the Austronesian migrants in the maternal line, discussed further in Sect. 10.3.2, Y-chromosome haplogroups throughout Near and Remote Oceania are generally dominated by those of Near Oceanian origin (Delfin et al. 2012; Scheinfeldt et al. 2006; Kayser et al. 2000, 2001, 2003, 2006, 2008a; van Oven et al. 2014; Hagelberg et al. 1999). Y-chromosome studies to date have been primarily based on SNP or STR typing and thus cannot offer the same level of resolution as most whole mtDNA genome sequence studies. However, it has been observed that Y-chromosome diversity is particularly low in the highlands of New Guinea, while in Remote Oceania, where a single mtDNA haplogroup of Asian origin reaches near fixation, the Y-chromosome diversity is considerably higher and Near Oceanian haplogroups are found at, on average, 66% frequency (Kayser et al. 2003, 2006; Mona et al. 2007; Hagelberg et al. 1999). These differences have been largely attributed to cultural practices and the customs of patrilocality and potentially polygyny in Papuan societies, versus a probably matrilocal residence pattern in Austronesian societies (Feil 1987; Jordan et al. 2009).

To date, studies of genome-wide data in Near and Remote Oceania are limited. One study of autosomal STR variation suggested that populations are genetically distinct and that population differentiation is driven less by linguistic classification than by island size and topographical complexity (Friedlaender et al. 2008). However, differentiation can also be observed between groups in the same geographic regions which are separated by languages, as has been observed in the Southern Highlands of Papua New Guinea between Huli and Mendi-Kewa speaking populations (Wollstein et al. 2010). We will discuss genome-wide data in more detail in Sect. 10.3.3.

10.3.2 The Austronesian Expansion and the Colonisation of Remote Oceania

Around 3.5 kya, more than 40 ky after the initial Papuan settlement, a new cultural complex appears in the Near Oceanian archaeological record (Specht and Gosden 1997; Summerhayes 2001). First identified in the Bismarck Archipelago and synonymous with a particular style of pottery, the Lapita Cultural Complex is believed to have arrived with migrants whose ancestors left Taiwan approximately 5 kya and travelled throughout ISEA before reaching Near Oceania (Kirch 2010; Ko et al. 2014; Gray et al. 2009). The migrants who arrived in the Bismarck Archipelago are believed to have spoken Proto-Oceanic, a language which soon after began to diversify and spread and is the ancestral language of all Austronesian languages spoken in Near and Remote Oceania (with some exceptions in Micronesia) (Lewis et al. 2013; Gray et al. 2009). Compared to the Pleistocene settlers of Near Oceania, the Austronesian migrants are thought to have had better sailing technology which allowed them to spread more rapidly and also cross larger distances over water. Indeed, the Austronesian expansion colonised Remote Oceania, and potentially much of the Solomon Islands, within approximately 2.5 kya (Kirch 2010; Wilmshurst et al. 2011).

The Austronesian expansion is closely associated with a particular mtDNA lineage—the B4a1a1 lineage—which reaches near fixation in Remote Oceania (Melton et al. 1995; Redd et al. 1995; Kayser et al. 2006; Hagelberg et al. 1999; Duggan et al. 2014). The B4a1a1 lineage is found in high frequencies in almost all Near Oceanian populations, even Papuan speaking populations, with the exception of New Guinea Highland populations and some populations in the Bismarck Archipelago (Kayser et al. 2006, 2008a; Delfin et al. 2012; van Oven et al. 2014; Duggan et al. 2014; Friedlaender et al. 2007). A particular mutation on the B4a1a1 lineages, an A→G transition at position 16,247, became known as the 'Polynesian motif' due to its extreme frequency in Remote Oceania and was synonymous with the Austronesian spread in Oceania (Melton et al. 1995). Reconstructed time estimates for the first appearance of this mutation have generated the suggestion that it was in fact present in Near Oceania prior to the arrival of the Lapita cultural complex (Soares et al. 2011). However, the estimates for haplogroup ages have large confidence intervals and these intervals do include the archaeological time estimate for the arrival of Lapita in Near Oceania (Soares et al. 2011; Specht and Gosden 1997). Also, the transition known as the 'Polynesian motif' has been found to be unstable; the derived allele appears to have arisen only a single time but has undergone multiple independent back mutations on multiple lineages and was found to have elevated levels of heteroplasmy, which likely confounds any attempt to date the lineage (Duggan and Stoneking 2013). Following the documentation of the instability of the derived 16,247 allele, the phylogeny of the B4a1a1 lineage underwent major revisions and the diagnostic mutations for many haplogroups were modified (PhyloTree Build 16, van Oven and Kayser 2009). The rapid expansion of Austronesians throughout Oceania is evident in the B4a1a1 lineage, where there is remarkably little diversity in haplogroups between populations. Identical whole mtDNA haplotypes can be observed between populations separated by thousands of kilometres (Duggan et al. 2014). This extensive sharing and the low observed haplogroup diversity leads to two conclusions: first, Austronesian women admixed extensively with existing Near Oceanian populations and second, the people involved in the Austronesian expansion through Oceania were small in number and/or very closely related. Their movement throughout Oceania occurred so rapidly and recently that mutations did not have time to develop on their lineage during the expansion nor have the populations been settled long enough for drift to occur.

While Austronesian mtDNA is clearly dominated by the B4a1a1 lineage, there is no single dominant Austronesian Y-chromosome lineage, though there may be population- or island-specific dominance of a given lineage (Hurles et al. 2002; Kayser et al. 2000, 2006; Scheinfeldt et al. 2006; Hagelberg et al. 1999; Delfin et al. 2012). In fact, the presence of Austronesian Y-chromosomes in Oceania is not nearly as extreme as is the case with Austronesian mtDNA. Approximately 94% of Remote Oceanian mtDNAs belong to the single B4a1a1 lineage while approximately 65% of Remote Oceanian Y-chromosomes are of Near Oceanian origin and belong to multiple lineages (Kayser et al. 2006). Thus, while the settlement of Remote Oceania was clearly achieved by Austronesian peoples with Lapita cultural practices and Oceanic languages, Near Oceanian men were undoubtedly involved as well. As methods for the study and phylogenetic resolution of Y-chromosomes become more advanced, it will become possible to conduct studies of male-mediated diversity and origin similar to those which are currently achieved with whole mtDNA genomes (Karafet et al. 2014; Lippold et al. 2014). These developments should shed further light on the paternal history of Oceania and may illuminate additional patterns of diversity or gene flow.

10.3.3 Duality of Oceanian Heritage

Uniparental markers provide ample evidence for dual-ancestry in Oceanian populations, however, given their unique inheritance patterns and lack of recombination, it can prove difficult to estimate time since admixture and accurate individual ancestry proportions from mtDNA and Y-chromosome data. Genome-wide data are a much richer source of information on ancestry and admixture and while data is still sparse, the existing genome-wide studies of Oceanian populations have already provided great insights.

Studying populations primarily from Northern Island Melanesia and analysing insertion/deletion polymorphisms and STR variation, Friedlaender et al. (2008) concluded that Oceanian populations did have heritage from both Papuan and Asian sources but that the Austronesian signal was very low and present almost exclusively in Austronesian-speaking populations. This work is notable for its extensive sampling in the Bismarck Archipelago and Bougainville. The findings of higher genome-wide Papuan ancestry validate earlier mtDNA and Y-chromosome work in the area and corroborate with archaeological evidence which decisively points to early and prolonged Papuan settlement in the area prior to the arrival of the Austronesian expansion (Scheinfeldt et al. 2006; Friedlaender et al. 2007, 2008; Summerhayes 2007).

The other two genome-wide studies focused on the ancestry of Polynesian populations. In a study of STR variation, it was concluded that Polynesians carry, on average, 79% Austronesian and 21% Papuan ancestry (Kayser et al. 2008b). A later study working with genome-wide SNP data and accounting for ascertainment bias in the SNP selection, adjusted this figure to 87% Austronesian and 13% Papuan ancestry (Wollstein et al. 2010). Both of these estimates suggest a greater proportion of Austronesian than Papuan ancestry but not to the same extreme as observed in mtDNA data, further validating the value of genome-wide ancestry assessments (Kayser et al. 2006). In addition to investigating Polynesian ancestry, Wollstein et al. (2010) concluded that Fiji represented a population which had undergone an additional admixture event with Near Oceania. Fijian ancestry was estimated to be approximately 65% Polynesian and 35% Papuan (Wollstein et al. 2010). The initial admixture event between Austronesians and Papuans to produce the Polynesian ancestry was estimated to have occurred approximately 3 kya, which is very close to the estimate of 3.4 kya for the arrival of Austronesians in Near Oceania from archaeological records, while the second admixture between the ancestors of Fijians and Near Oceanians to produce a modern Fijian ancestry occurred sometime after the ancestors of Polynesians had already left Fiji (Wollstein et al. 2010).

10.3.4 Evidence for Contact Between South America and Polynesia

The two ancestral components, Near Oceanian and Austronesian, are well supported archaeologically, linguistically and biologically; however, there has been considerable speculation of contact between Remote Oceanian populations and South American populations. Such hypotheses originated to account for the presence of sweet potato (Roullier et al. 2013) and bottle gourds (Green 2000) from South America in Oceania. New research has demonstrated that the population of Rapa Nui (Easter Island), while predominantly Polynesian, have signatures of ancestry from both Native Americans and Europeans in their genomes (Moreno-Mayar et al. 2014). These signals were absent from other Polynesian samples examined and while the Native American proportion was, on average, less than the European proportion of the Rapa Nui genome (8% and 16%, respectively), the Native American component was composed of shorter tracts of inheritance and was more evenly distributed across the population suggesting that it reflected an older admixture event that the European proportion (Moreno-Mayar et al. 2014). The Native American admixture was estimated to have occurred between AD 1280–1425, well before the first European contact with Rapa Nui in AD 1722, and the European admixture was estimated to have occurred between AD 1850–1870, contemporary with European mediated slave trade in the Pacific (Moreno-Mayar et al. 2014 and references therewithin). This genomic signature has two possible explanations: South Americans sailed to Rapa Nui and intermixed with the local population; or Polynesians sailed to South America and then returned to Rapa Nui either carrying with them individuals of South American ancestry or the products of gene flow in South America. While the second scenario involves sailing greater distances, the feat of settling all the Pacific islands is formidable and there is no reason to believe that the Austronesian Expansion should have stopped there; the coast of South America is closer to Rapa Nui than either New Zealand or the Hawai'ian Islands, the other two points of the Polynesian Triangle. Further evidence for Polynesian contact with South America may come by way of Brazil. Two skulls associated with the Botocudo population from the Museu Nacional of Brazil were found to be of Polynesian origin (Malaspinas et al. 2014). While radio isotopes were unable to unequivocally confirm that the skulls pre-dated European contact with Polynesia, they do suggest that these two Polynesian individuals made their way not only to South America, but also either across the Andes or around the Cape to eastern Brazil (Malaspinas et al. 2014).

10.3.5 The Unique History of Santa Cruz

Santa Cruz occupies a distinctive position in the history of Oceania. It is part of a group of islands located beyond the boundary of Remote Oceania, approximately 400 km past the main chain of the Solomon Islands. Archaeological evidence suggests that Santa Cruz was colonised 3 kya by Lapita people who engaged in obsidian trade with populations from New Britain in the Bismarck Archipelago over a period of several hundred years (Sheppard et al. 2015). Curiously, it appears that the people voyaging between New Britain and Santa Cruz did so directly, 'leapfrogging' over the rest of the Solomon Islands, which lacks evidence of any contemporary Lapita settlements (Sheppard and Walter 2006). Linguistically, the languages of the Temotu family found in Santa Cruz and the nearby Reef Islands also have a unique history. Once believed to be the easternmost Papuan languages, they are now recognised as part of the Oceanic family of Austronesian though they are more distantly related to other Oceanic languages and are thought to have separated from the rest of the Oceanic family soon after the diversification of Proto-Oceanic (Ross and Næss 2007).

Curiously, the mtDNA and Y-chromosome lineages found in Santa Cruz contrast greatly with the rest of Remote Oceania. Available genetic data on Santa Cruz is restricted to mtDNA and the Y-chromosome, but nonetheless these exhibit interesting patterns. Despite an archaeological record which suggests no pre-Lapita settlement, the mtDNA and Y-chromosome composition of Santa Cruz has a greater proportion of Near Oceanian ancestry than any other population in Remote Oceania (Delfin et al. 2012; Duggan et al. 2014; Friedlaender et al. 2002). In fact, the proportion of Near Oceanian ancestry in Santa Cruz is greater than that of many Near Oceanian populations, particularly with regard to mtDNA ancestry (Fig. 10.4) (Delfin et al. 2012; Duggan et al. 2014). Furthermore, this Near Oceanian ancestry is present as multiple haplotypes within diverse haplogroups, which suggests that this elevated Near Oceanian component is not the result of a recent bottleneck or population replacement (Duggan et al. 2014). There are several possible explanations for this observation, all detailed in Delfin et al. (2012). One of the possibilities is that there was in fact a pre-Austronesian settlement of Santa Cruz; this scenario was recently tested by comparing empirical mtDNA data from Santa Cruz and New Britain to simulated data for a variety of settlement histories. The results suggested that a time depth of greater than 3 kya for the settlement of Santa Cruz was possible, but did not completely exclude the possibility of a Lapita-age colonisation (Duggan et al. 2014).

The apparent disconnect between archaeology, genetics and linguistics in Santa Cruz also calls into question the progression of settlement into Remote Oceania. The mtDNA genetic profile of Santa Cruz is strikingly different from other Remote Oceanian populations (Duggan et al. 2014); this makes it unlikely that it was a seed population for the colonisation of islands further to the east. Archaeology suggests that Vanuatu and New Caledonia were settled by Lapita people at approximately the same time as Santa Cruz (Walter and Sheppard 2009; Sheppard and Walter 2006). Comparable datasets from these regions could clarify whether their composition is similar to Santa Cruz or, if they are different in comparison with Santa Cruz, could they possibly be the seed populations for Remote Oceanian populations. If Vanuatu and New Caledonia do appear more similar to other Remote Oceanian populations, it could further proposed model for the settlement of Santa Cruz whereby it was first colonised by Lapita people but incurred extensive gene

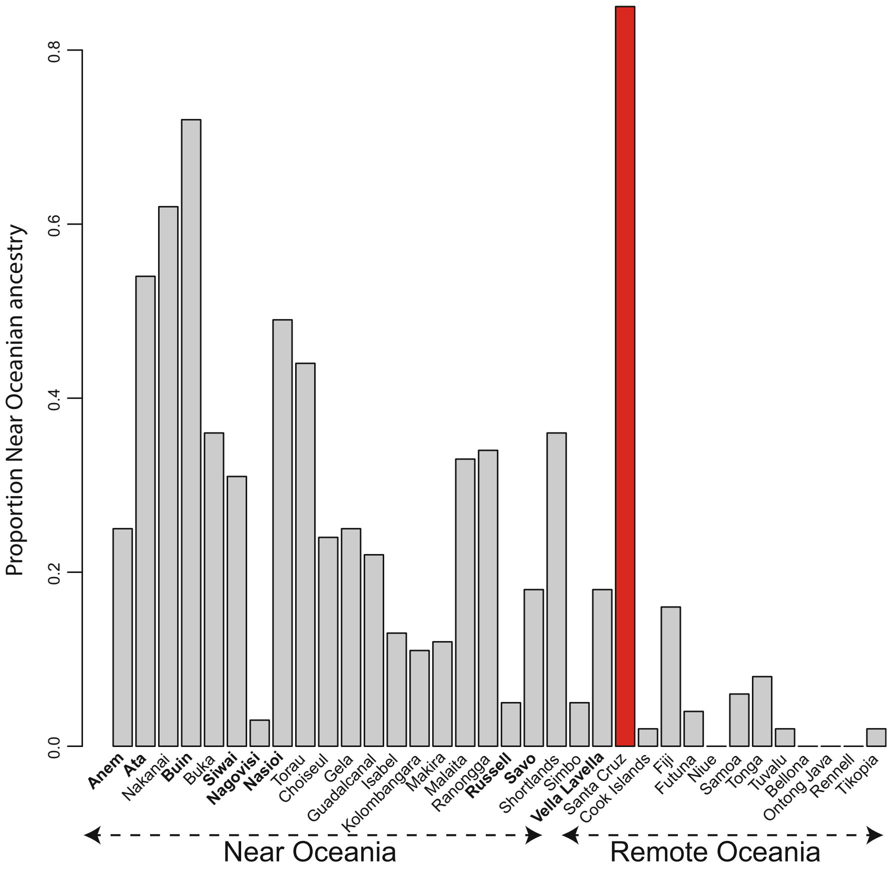


Fig. 10.4 Relative proportion of Near Oceanian ancestry, based on mtDNA data. Papuan speaking populations are identified in bold, the extreme Near Oceanian ancestry of Santa Cruz is highlighted in red. Data source: Duggan et al. (2014), supplemental Table 2

flow and possible secondary settlement during the obsidian trade (Delfin et al. 2012). Additionally, genome-wide data could elucidate the genetic history of Santa Cruz with greater precision in determining ancestry composition and dating the time of admixture events(s).

10.4 Archaic Admixture in Australia and Oceania

Thus far, we have only discussed the ancestry of Australia and Oceania as they relate to modern human populations. Ancient DNA—in particular, admixture signals from archaic hominins—is another source of ancestry information. All modern human populations outside of Africa received a small portion of their genome, 1–2% on average, from Neanderthals (Green et al. 2010; Prüfer et al. 2014). However, when a second archaic hominin more closely related to Neanderthals than humans, 'Denisovans', was identified, evidence of introgression between modern humans and Denisovans was found only in New Guinea Highlanders and Bougainvilleans (Reich et al. 2010). In a follow-up study, 33 Asian and Oceanian populations were screened for evidence of Denisovan admixture with genome-wide SNP data (Reich et al. 2011). This work concluded that there was evidence of admixture in New Guinea Highlanders, Aboriginal Australians, Fijians, Polynesians, populations from the Nusa Tenggaras and Moluccas in eastern Indonesia, and the Mamanwa 'Negrito' group and Manobo populations of the Philippines (Reich et al. 2011) (Fig. 10.5). New Guinea Highlanders and Aboriginal Australians have approximately the same amount of Denisovan ancestry, consistent again with their origin from a common ancestral population. The average Denisovan ancestry in Mamanwa is only about one half of that present in New Guinea Highlanders and Aboriginal Australians. However, this ancestry is best modelled whereby a common ancestral population of New Guinea Highlanders, Aboriginal Australians and Mamanwa incurs gene flow from a Denisovan population, and subsequent to the split of Mamanwa from the ancestral population of New Guinea Highlanders and Aboriginal Australians, the Mamanwa incur significant admixture with one or more groups not carrying any Denisovan ancestry (Reich et al. 2011). Furthermore, the Denisovan ancestry of the Fijian, Polynesian and Indonesian populations from the Nusa Tenggaras and Moluccas is derived from a more recent admixture event between incoming Austronesian populations and existing Papuan populations (Reich et al. 2011; Xu et al. 2012). Denisovan ancestry was also found in the ancient Aboriginal Australian genome sequence (Rasmussen et al. 2011), albeit at slightly lower levels than previously estimated in New Guinea Highlanders (Reich et al. 2010).

10.5 Conclusions

In summary, existing genetic data suggest a very dynamic history of population migration and admixture within Australia and Oceania. Descendants of an initial early migration out of Africa, the ancestral population of Aboriginal Australians, New Guinea Highlanders and the Philippine Mamanwa admixed with Denisovans, an archaic hominin population. This ancestral population then split and each division incurred separate and unique admixture events: the Mamanwa with one or more groups related to modern Han populations, Aboriginal Australians with a population from the Indian subcontinent, and with the exception of some populations isolated in the Highlands of New Guinea, almost all Papuan populations experienced significant gene flow with Austronesian migrants.

Through the study of modern populations, it is possible to discern how modern populations are unique but also deeply related to each other. Many of these discoveries have happened only recently due to technological improvements in obtaining genetic data and in statistical modelling and methodology. Recent work on Southeast

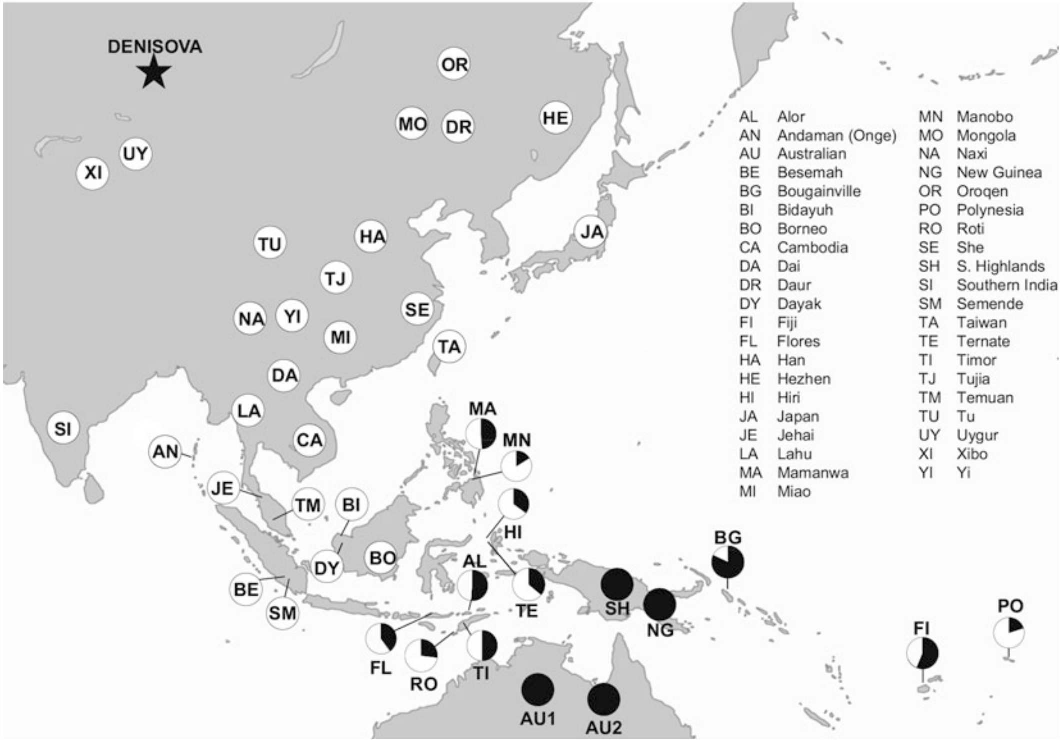


Fig. 10.5 Denisovan ancestry as a fraction of that found in New Guineans. Used with permission from Reich et al. (2011)

Asian populations identifies the potential of yet another admixture or population expansion event which has not been previously detected through archaeology or linguistics (Lipson et al. 2014). Despite sound linguistic reconstructions (Gray et al. 2009), the genetic origin of the Austronesian expansion in Taiwan was only recently confirmed thanks to the discovery of an 8000 year old skeleton ancestral to Taiwanese aboriginal populations, and sophisticated modelling which identified an entry into Taiwan at approximately 6 kya and departure of the Austronesian expansion from Taiwan around 4 kya (Ko et al. 2014). Future work in Oceania should shed light on additional population migrations and interactions as well. We are particularly hopeful that as technologies continue to improve, additional results from ancient DNA, such as the recent study of ancient Maori (Knapp et al. 2012), as well as genome-wide studies of large numbers of SNPs or even whole-genome sequencing will investigate areas of Near Oceania such as the Solomon Islands (comparatively under-studied and curiously bereft of archaeological sites), as well as non-Polynesian populations in Remote Oceania (i.e. Vanuatu and Near Caledonia). These populations will be of particular interest and may provide answers to lingering questions as to the settlement and provenance of some Oceanian populations.

References

Delfin F, Myles S, Choi Y, Hughes D, Illek R, van Oven M, Pakendorf B, Kayser M, Stoneking M (2012) Bridging near and remote Oceania: mtDNA and NRY variation in the Solomon Islands. Mol Biol Evol 29(2):545–564

Duggan AT (2014) Investigations of the settlement and demographic history of Oceania through whole mitochondrial genome sequence analysis. PhD thesis, University of Leipzig, Leipzig

Duggan AT, Evans B, Friedlaender FR, Friedlaender JS, Koki G, Merriwether DA, Kayser M, Stoneking M (2014) Maternal history of Oceania from complete mtDNA genomes: contrasting ancient diversity with recent homogenization due to the Austronesian expansion. Am J Hum Genet 94(5):721–733

Duggan AT, Stoneking M (2013) A highly unstable recent mutation in human mtDNA. Am J Hum Genet 92(2):279–284

Feil DK (1987) The evolution of Highland Papua New Guinea societies. Cambridge University Press, Cambridge

Fredericksen C, Spriggs M, Ambrose W (1993) Pamwak rockshelter: a pleistocene site on Manus Island, Papua New Guinea. In: Smith MA, Spriggs M, Fankhauser B (eds) Sahul in review: pleistocene archaeology in Australia, New Guinea and Island Melanesia. Australian National University, Canberra

Friedlaender J, Schurr T, Gentz F, Koki G, Friedlaender F, Horvat G, Babb P, Cerchio S, Kaestle F, Schanfield M, Deka R, Yanagihara R, Merriwether DA (2005) Expanding Southwest Pacific mitochondrial haplogroups P and Q. Mol Biol Evol 22(6):1506–1517

Friedlaender JS, Friedlaender FR, Hodgson JA, Stoltz M, Koki G, Horvat G, Zhadanov S, Schurr TG, Merriwether DA (2007) Melanesian mtDNA complexity. PLoS One 2(2):e248. https://doi.org/10.1371/journal.pone.0000248

Friedlaender JS, Friedlaender FR, Reed FA, Kidd KK, Kidd JR, Chambers GK, Lea RA, Loo J-H, Koki G, Hodgson JA, Merriwether DA, Weber JL (2008) The genetic structure of Pacific Islanders. PLoS Genet 4(1):e19. https://doi.org/10.1371/journal.pgen.0040019

Friedlaender JS, Gentz F, Green K, Merriwether DA (2002) A cautionary tale on ancient migration detection: mitochondrial DNA variation in Santa Cruz Islands, Solomon Islands. Hum Biol 74 (3):453–471

Gray R, Drummond A, Greenhill S (2009) Language phylogenies reveal expansion pulses and pauses in Pacific settlement. Science 323(5913):479–483

Green R, Krause J, Briggs A, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz M, Hansen N, Durand E, Malaspinas A-S, Jensen J, Marques-Bonet T, Alkan C, Prüfer K, Meyer M, Burbano H, Good J, Schultz R, Aximu-Petri A, Butthof A, Höber B, Höffner B, Siegemund M, Weihmann A, Nusbaum C, Lander E, Russ C, Novod N, Affourtit J, Egholm M, Verna C, Rudan P, Brajkovic D, Kucan Z, Gusic I, Doronichev V, Golovanova L, Lalueza-Fox C, de la Rasilla M, Fortea J, Rosas A, Schmitz R, Johnson P, Eichler E, Falush D, Birney E, Mullikin J, Slatkin M, Nielsen R, Kelso J, Lachmann M, Reich D, Pääbo S (2010) A draft sequence of the Neandertal genome. Science 328(5979):710–722

Green RC (1991) Near and remote Oceania: disestablishing "Melanesia" in culture history. In: Pawley A (ed) Man and a half: essays in Pacific anthropology and Ethnobiology in honour of Ralph Bulmer. Polynesian Society, Auckland, pp 491–502

Green RC (2000) A range of disciplines support a dual origin for the bottle gourd in the Pacific. J Polynesian Soc 109(2):191–198

Groube L, Chappell J, Muke J, Price D (1986) A 40,000 year-old human occupation site at Huon Peninsula, Papua New Guinea. Nature 324(6096):453–455

Hagelberg E, Kayser M, Nagy M, Roewer L, Zimdahl H, Krawczak M, Lió P, Schiefenhövel W (1999) Molecular genetic evidence for the human settlement of the Pacific: analysis of mitochondrial DNA, Y chromosome and HLA markers. Philos Trans R Soc Lond B Biol Sci 354 (1379):141–152

Hudjashov G, Kivisild T, Underhill P, Endicott P, Sanchez J, Lin A, Shen P, Oefner P, Renfrew C, Villems R, Forster P (2007) Revealing the prehistoric settlement of Australia by Y chromosome and mtDNA analysis. Proc Natl Acad Sci U S A 104(21):8726–8730

Huoponen K, Schurr T, Chen Y, Wallace D (2001) Mitochondrial DNA variation in an aboriginal Australian population: evidence for genetic isolation and regional differentiation. Hum Immunol 62(9):954–969

Hurles M, Nicholson J, Bosch E, Renfrew C, Sykes B, Jobling M (2002) Y chromosomal evidence for the origins of oceanic-speaking peoples. Genetics 160(1):289–303

Ingman M, Gyllensten U (2003) Mitochondrial genome variation and evolutionary history of Australian and new Guinean aborigines. Genome Res 13(7):1600–1606

Jordan F, Gray R, Greenhill S, Mace R (2009) Matrilocal residence is ancestral in Austronesian Societies. Proc Biol Sci R Soc 276(1664):1957–1964

Karafet T, Mendez F, Sudoyo H, Lansing J, Hammer M (2014) Improved phylogenetic resolution and rapid diversification of Y-chromosome haplogroup K-M526 in Southeast Asia. Eur J Hum Genet 23(3):369–373

Kayser M, Brauer S, Cordaux R, Casto A, Lao O, Zhivotovsky L, Moyse-Faurie C, Rutledge R, Schiefenhoevel W, Gil D, Lin A, Underhill P, Oefner P, Trent R, Stoneking M (2006) Melanesian and Asian origins of Polynesians: mtDNA and Y chromosome gradients across the Pacific. Mol Biol Evol 23(11):2234–2244

Kayser M, Brauer S, Weiss G, Schiefenhövel W, Underhill P, Shen P, Oefner P, Tommaseo-Ponzetta M, Stoneking M (2003) Reduced Y-chromosome, but not mitochondrial DNA, diversity in human populations from West New Guinea. Am J Hum Genet 72(2):281–302

Kayser M, Brauer S, Weiss G, Schiefenhövel W, Underhill P, Stoneking M (2001) Independent histories of human Y chromosomes from Melanesia and Australia. Am J Hum Genet 68 (1):173–190

Kayser M, Brauer S, Weiss G, Underhill P, Roewer L, Schiefenhövel W, Stoneking M (2000) Melanesian origin of Polynesian Y chromosomes. Curr Biol 10(20):1237–1246

Kayser M, Choi Y, van Oven M, Mona S, Brauer S, Trent R, Suarkia D, Schiefenhövel W, Stoneking M (2008a) The impact of the Austronesian expansion: evidence from mtDNA and

Y chromosome diversity in the Admiralty Islands of Melanesia. Mol Biol Evol 25 (7):1362–1374

Kayser M, Lao O, Saar K, Brauer S, Wang X, Nürnberg P, Trent R, Stoneking M (2008b) Genome-wide analysis indicates more Asian than Melanesian ancestry of Polynesians. Am J Hum Genet 82(1):194–198

Kirch P (2010) Peopling of the Pacific: a holistic anthropological perspective. Ann Rev Anthropol 39:131–148

Kirch PV (2000) On the road of the winds—an archaeological history of the Pacific Islands before European contact. University of California Press, Berkeley

Knapp M, Horsburgh K, Prost S, Stanton J-A, Buckley H, Walter R, Matisoo-Smith E (2012) Complete mitochondrial DNA genome sequences from the first New Zealanders. Proc Natl Acad Sci U S A 109(45):18350–18354

Ko A, Chen C-Y, Fu Q, Delfin F, Li M, Chiu H-L, Stoneking M, Ko Y-C (2014) Early Austronesians: into and out of Taiwan. Am J Hum Genet 94(3):426–436

Leavesley MG, Bird MI, Fifield LK, Hausladen P, Santos G, Di Tada M (2002) Buang Merabak: early evidence for human occupation in the Bismarck Archipelago, Papua New Guinea. Aust Archaeol 54:55–57

Leavesley MG, Chappell J (2004) Buang Merabak: additional early radiocarbon evidence of the colonisation of the Bismarck Archipelago, Papua New Guinea. Antiquity 78, online project gallery. http://www.antiquity.ac.uk/Projgall/leavesley/

Lewis MP, Simons GF, Fennig CD (2013) Ethnologue: languages of the world, Seventeenth edition. SIL International. http://www.ethnologue.com

Lippold S, Xu H, Ko A, Li M, Renaud G, Butthof A, Schroeder R, Stoneking M (2014) Human paternal and maternal demographic histories: insights from high-resolution Y chromosome and mtDNA sequences. Investigative Genet 5:13. https://doi.org/10.1186/2041-2223-5-13

Lipson M, Loh P-R, Patterson N, Moorjani P, Ko Y-C, Stoneking M, Berger B, Reich D (2014) Reconstructing Austronesian population history in Island Southeast Asia. Nat Commun 5:4689. https://doi.org/10.1038/ncomms5689

Malaspinas A-S, Lao O, Schroeder H, Rasmussen M, Raghavan M, Moltke I, Campos PF, Sagredo FS, Rasmussen S, Gonçalves VF, Albrechtsen A, Allentoft ME, Johnson PLF, Li M, Reis S, Bernardo DV, DeGiorgio M, Duggan AT, Bastos M, Wang Y, Stenderup J, Moreno-Mayar JV, Brunak S, Sicheritz-Ponten T, Hodges E, Hannon GJ, Orlando L, Price TD, Jensen JD, Nielsen R, Heinemeier J, Olsen J, Rodrigues-Carvalho C, Lahr MM, Neves WA, Kayser M, Higham T, Stoneking M, Pena SDJ, Willerslev E (2014) Teo ancient human genomes reveal Polynesian ancestry among the indigenous Botocudos of Brazil. Curr Biol 24:R1035–R1037

McEvoy B, Lind J, Wang E, Moyzis R, Visscher P, van Holst Pellekaan S, Wilton A (2010) Whole-genome genetic diversity in a sample of Australians with deep aboriginal ancestry. Am J Hum Genet 87(2):297–305

Melton T, Peterson R, Redd A, Saha N, Sofro A, Martinson J, Stoneking M (1995) Polynesian genetic affinities with southeast Asian populations as identified by mtDNA analysis. Am J Hum Genet 57(2):403–414

Merriwether DA, Hodgson JA, Friedlaender FR, Allaby R, Cerchio S, Koki G, Friedlaender JS (2005) Ancient mitochondrial M haplogroups identified in the Southwest Pacific. Proc Natl Acad Sci U S A 102(37):13034–13039

Mona S, Tommaseo-Ponzetta M, Brauer S, Sudoyo H, Marzuki S, Kayser M (2007) Patterns of Y-chromosome diversity intersect with the trans-New Guinea hypothesis. Mol Biol Evol 24 (11):2546–2555

Moreno-Mayar JV, Rasmussen S, Seguin-Orlando A, Rasmussen M, Liang M, Flåm ST, Lie BA, Gilfillan GD, Nielsen R, Thorsby E, Willserslev E, Malaspinas A-S (2014) Genome-wide ancestry patterns in Rapanui suggest pre-European admixture with Native Americans. Curr Biol 24(21):2518–2525

O'Connell JF, Allen J (2004) Dating the colonization of Sahul (Pleistocene Australia–New Guinea): a review of recent research. J Archaeol Sci 31:835–853

Pääbo S, Poinar H, Serre D, Jaenicke-Despres V, Hebler J, Rohland N, Kuch M, Krause J, Vigilant L, Hofreiter M (2004) Genetic analyses from ancient DNA. Annu Rev Genet 38:645–679

Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant P, de Filippo C, Li H, Mallick S, Dannemann M, Fu Q, Kircher M, Kuhlwilm M, Lachmann M, Meyer M, Ongyerth M, Siebauer M, Theunert C, Tandon A, Moorjani P, Pickrell J, Mullikin J, Vohr S, Green R, Hellmann I, Johnson P, Blanche H, Cann H, Kitzman J, Shendure J, Eichler E, Lein E, Bakken T, Golovanova L, Doronichev V, Shunkov M, Derevianko A, Viola B, Slatkin M, Reich D, Kelso J, Pääbo S (2014) The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505(7481):43–49

Pugach I, Delfin F, Gunnarsdóttir E, Kayser M, Stoneking M (2013) Genome-wide data substantiate Holocene gene flow from India to Australia. Proc Natl Acad Sci U S A 110(5):1803–1808

Rasmussen M, Guo X, Wang Y, Lohmueller K, Rasmussen S, Albrechtsen A, Skotte L, Lindgreen S, Metspalu M, Jombart T, Kivisild T, Zhai W, Eriksson A, Manica A, Orlando L, De La Vega F, Tridico S, Metspalu E, Nielsen K, Ávila-Arcos M, Moreno-Mayar J, Muller C, Dortch J, Gilbert M, Lund O, Wesolowska A, Karmin M, Weinert L, Wang B, Li J, Tai S, Xiao F, Hanihara T, van Driem G, Jha A, Ricaut F-X, de Knijff P, Migliano A, Gallego Romero I, Kristiansen K, Lambert D, Brunak S, Forster P, Brinkmann B, Nehlich O, Bunce M, Richards M, Gupta R, Bustamante C, Krogh A, Foley R, Lahr M, Balloux F, Sicheritz-Pontén T, Villems R, Nielsen R, Wang J, Willerslev E (2011) An aboriginal Australian genome reveals separate human dispersals into Asia. Science 334(6052):94–98

Redd A, Roberts-Thomson J, Karafet T, Bamshad M, Jorde L, Naidu J, Walsh B, Hammer M (2002) Gene flow from the Indian subcontinent to Australia: evidence from the Y chromosome. Curr Biol 12(8):673–677

Redd A, Stoneking M (1999) Peopling of Sahul: mtDNA variation in aboriginal Australian and Papua new Guinean populations. Am J Hum Genet 65(3):808–828

Redd A, Takezaki N, Sherry S, McGarvey S, Sofro A, Stoneking M (1995) Evolutionary history of the COII/tRNALys intergenic 9 base pair deletion in human mitochondrial DNAs from the Pacific. Mol Biol Evol 12(4):604–615

Reich D, Green R, Kircher M, Krause J, Patterson N, Durand E, Viola B, Briggs A, Stenzel U, Johnson P, Maricic T, Good J, Marques-Bonet T, Alkan C, Fu Q, Mallick S, Li H, Meyer M, Eichler E, Stoneking M, Richards M, Talamo S, Shunkov M, Derevianko A, Hublin J-J, Kelso J, Slatkin M, Pääbo S (2010) Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468(7327):1053–1060

Reich D, Patterson N, Kircher M, Delfin F, Nandineni M, Pugach I, Ko A, Ko Y-C, Jinam T, Phipps M, Saitou N, Wollstein A, Kayser M, Pääbo S, Stoneking M (2011) Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. Am J Hum Genet 89 (4):516–528

Roe D (1993) Prehistory without pots: Prehistoric settlement and economy of North-west Guadalcanal, Solomon Islands. PhD thesis. Australian National University, Canberra

Ross M, Næss Å (2007) An Oceanic origin for Äiwoo, the language of the Reef Islands? Oceanic Linguistics 46:456–498

Roullier C, Benoit L, McKey DB, Lebot V (2013) Historical collections reveal patterns of diffusion of sweet potato in Oceania obscured by modern plant movements and recombination. Proc Natl Acad Sci U S A 110(6):2205–2210

Scheinfeldt L, Friedlaender F, Friedlaender J, Latham K, Koki G, Karafet T, Hammer M, Lorenz J (2006) Unexpected NRY chromosome variation in Northern Island Melanesia. Mol Biol Evol 23(8):1628–1641

Sheppard P, Chiu S, Walter R (2015) Re-dating Lapita movement into Remote Oceania. J Pacific Archaeol 6(1):26–36

Sheppard P, Walter R (2006) A revised model of Solomon Islands culture history. J Polynesian Soc 115(1):47–76

Soares P, Rito T, Trejaut J, Mormina M, Hill C, Tinkler-Hundal E, Braid M, Clarke D, Loo J-H, Thomson N, Denham T, Donohue M, Macaulay V, Lin M, Oppenheimer S, Richards M (2011) Ancient voyaging and Polynesian origins. Am J Hum Genet 88(2):239–247

Specht J, Gosden C (1997) Dating Lapita Pottery in the Bismark Archipelago, Papua New Guinea. University of Hawai'i Press (Honolulu)

Spriggs M (1997) The Island Melanesians. Blackwell Publishers, Cornwall

Stoneking M, Jorde L, Bhatia K, Wilson A (1990) Geographic variation in human mitochondrial DNA from Papua New Guinea. Genetics 124(3):717–733

Summerhayes G, Leavesley M, Fairbairn A, Mandui H, Field J, Ford A, Fullagar R (2010) Human adaptation and plant use in highland New Guinea 49,000 to 44,000 years ago. Science 330 (6000):78–81

Summerhayes GR (2001) Lapita in the far west: recent developments. Archaeol Ocean 36(2):53–64

Summerhayes GR (2007) Island Melanesian pasts: a view from archaeology. In: Friedlaender JS (ed) Genes, language and culture history in the Southwest Pacific. Oxford University Press, New York

Tommaseo-Ponzetta M, Attimonelli M, De Robertis M, Tanzariello F, Saccone C (2002) Mitochondrial DNA variability of West New Guinea populations. Am J Phys Anthropol 117 (1):49–67

van Holst Pellekaan S, Frommer M, Sved J, Boettcher B (1998) Mitochondrial control-region sequence variation in aboriginal Australians. Am J Hum Genet 62(2):435–449

van Holst Pellekaan S, Ingman M, Roberts-Thomson J, Harding R (2006) Mitochondrial genomics identifies major haplogroups in aboriginal Australians. Am J Phys Anthropol 131(2):282–294

van Holst Pellekaan SM (2000) Genetic research: what does this mean for indigenous Australian communities? Aust Aborig Stud 1(2):65–75

van Oven M, Brauer S, Choi Y, Ensing J, Schiefenhövel W, Stoneking M, Kayser M (2014) Human genetics of the Kula Ring: Y-chromosome and mitochondrial DNA variation in the Massim of Papua New Guinea. Eur J Hum Genet 22(12):1393–1403

van Oven M, Kayser M (2009) Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat 30(2):E386–E394

Walter R, Sheppard P (2009) A review of Solomon Islands archaeology. In: Sheppard P, Thomas T, Summerhayes G (eds) Lapita: ancestors and descendants. New Zealand Archaeological Association, Auckland

White JP, O'Connell JF (1982) A prehistory of Australia, New Guinea and Sahul. Academic Press Australia, North Ryde

Wickler S, Spriggs M (1988) Pleistocene human occupation of the Solomon Islands, Melanesia. Antiquity 62(237):703–706

Wilmshurst J, Hunt T, Lipo C, Anderson A (2011) High-precision radiocarbon dating shows recent and rapid initial human colonization of East Polynesia. Proc Natl Acad Sci U S A 108 (5):1815–1820

Wollstein A, Lao O, Becker C, Brauer S, Trent R, Nürnberg P, Stoneking M, Kayser M (2010) Demographic history of Oceania inferred from genome-wide data. Curr Biol 20(22):1983–1992

Xu S, Pugach I, Stoneking M, Kayser M, Jin L, Consortium HP-AS (2012) Genetic dating indicates that the Asian-Papuan admixture through eastern Indonesia corresponds to the Austronesian expansion. Proc Natl Acad Sci U S A 109(12):4574–4579

Chapter 11
America

Inaho Danjoh, Hideaki Kanzawa-Kiriyama, and Naruya Saitou

Abstract It is estimated that settlement of North and South American continents by ancestors of current indigenous peoples occurred 12 to 15,000 years ago. It is unlikely that evolutionary change occurred in these continents within this short period of time. Therefore, in the field of human genetics, research on indigenous peoples in North and South America is focused on: (1) Eurasian continents of origin; (2) the longitudinal spread of people across the American continents; and (3) how the indigenous people established their lives within the American continents.

The release of sequence data for the euchromatin region of the human genome in 2003, in addition to SNP (*S*ingle *N*ucleotide *P*olymorphism) data obtained from the International Hap Map Project, provided access to a tremendous amount of information. This dramatically altered the approach to research, and involved development of complicated statistical methods for analysis and the entry of Bioinformatists and Molecular Biologists into the field with little knowledge of human genetics. Interpretation of these genome analyses requires deep insight and information about history, linguistics, culture, customs, and archeology. Moreover, as a special case in American continents, incursion of Europeans during the early modern period caused destruction of the native civilization, and led to racial mingling followed by an era of discrimination against natives. This is one reason why it is so difficult to trace the migration paths of the ancestors of current natives entering from the Eurasian continent into the American continents.

In this chapter, we summarize the recent progress of genetic analysis of indigenous peoples in the American continents, and discuss the ethical issues associated with analysis of genome data.

I. Danjoh
Tohoku Medical Megabank Organization, Tohoku University, Sendai, Japan
e-mail: inaho.danjoh@megabank.tohoku.ac.jp

H. Kanzawa-Kiriyama
National Museum of Science and Nature, Tsukuba, Japan
e-mail: hkanzawa@kahaku.go.jp

N. Saitou (✉)
National Institute of Genetics, Mishima, Japan
e-mail: saitounr@nig.ac.jp

N. Saitou (ed.), *Evolution of the Human Genome II*, Evolutionary Studies,
https://doi.org/10.1007/978-4-431-56904-6_11

Keywords Amerind · North America · Central America · South America · American continents · Genome analysis · Indigenous · Ethical issues

11.1 Introduction

The American continents have a unique history when considering human dispersal all over the world. There is no evidence that any other hominids lived or evolved in the American continents prior to the arrival of the first Americans during the glacial age. It is thought that American continents received several migration waves of Native American ancestors from the Eurasian continent; entry and settlement were not a single event (Reich et al. 2012). According to the most accepted theory, the first American entered the American continent at the end of the glacial age approximately 15,000 years ago (15 KYA), and then dispersed across both continents within 2000 years, which is an abnormally short period of time for human migration during the prehistoric age. Due to the recent and rapid migration, there was not enough time to undergo and establish evolutionary alterations to the genomes of Native Americans within the American continents. As such, research is focused on: (1) the Eurasian continent of origin; (2) migration routes for dispersal across the American continents; and (3) how Native Americans lived and adapted to their new environment.

After settlement in the American Continents, and through the Stone Age, Native Americans produced some highly developed civilizations by the Age of Exploration, around the fifteenth to the seventeenth century. Facing expeditions from European countries of the age, some of these civilizations lost their independence and identity. In addition, an enormous number of people were transported from Africa to the Americas through the slave trade. Recent globalization at the end of the twentieth century has integrated indigenous peoples and cultures into the global economy. Because of these complex histories, American continents are racially and culturally more diverse than other areas in the world, and ethical and human rights problems related to unresolved racial issues still persist. It is, in other words, difficult for Native Americans to maintain their original cultures and their pedigrees as indigenous populations.

Release of the human genome draft sequence in 2001 (Lander ES et al. 2001) and progress of the International Hap Map Project (The International HapMap Consortium 2005; The International HapMap Consortium 2007; The International HapMap 3 Consortium 2010) that has identified SNPs within the human genome, dramatically changed the approaches and strategies for human genome analysis. We are now conducting this research in the midst of the era of Post-Human Genome Sequencing. Before the completion of whole genome sequencing, it was difficult to identify genomic loci to be analyzed because we did not know precisely which loci retained polymorphisms, or how diverse the loci were among individuals or tribes or races. Highly polymorphic regions of the mitochondrial genome, the HLA region, or genomic region coding cytochrome gene family (CYPs), have been traditionally used for comparative analyses. Entering into the Post-Human Genome Sequencing era, we already have information about polymorphic loci, and the linkage pattern of each region within the human genome passed down for generations. As analysis platforms, DNA microarrays, microsatellite probes, or even genomic sequences obtained by Next Generation Sequencers (NGSs) are supplied by companies and public databases, researchers can do whole genome analysis, either by themselves or through outsourcing. With this increased access to substantial genomic resources, many scientists without a specialization in human genetics have entered into this field of research. The influx of new perspectives and techniques has helped to further progress in the field, but there are limitations with respect to interpretation of these data. Rigorous interpretation requires deep insight into the history and relationships among the tribes to get an overview of genetic admixture at sites that is crucial for deciphering results of genome analyses. These limitations should be considered when reviewing the literature. Moreover, because of the complex history, serious ethical issues related to unresolved racial issues still persist as described above. When we analyze genomes of Native Americans, we should consider these ethical issues carefully.

11.2 Peopling of the American Continents During the Prehistoric Age

The American continents are the farthest destination of the Great Journey of *Homo sapiens* since leaving Africa. All human skeletal remains found in the Americas are anatomically modern *Homo sapiens*, and there is no evidence that other hominids had lived here before the modern humans entered the continents. The entry of modern *Homo sapiens* onto the North American continent was estimated at 15 KYA, a timeframe too recent for evolutionary change to have occurred locally. Therefore the main research objectives are:

From what part of the Eurasian continent did Native Americans originate?
What routes did they travel to disperse over American continents?
How did they establish their lives and adapt to their new environment?

11.2.1 Environment

The first Americans entered Alaska during the last ice age. At that period, the Bering Strait between Alaska and Eastern Siberia was land-bridged because the sea level was far lower than at present. This land bridge is called Beringia and it joined the two continents between 65 and 36 KYA. The presence of Beringia was affected by sea level, and the Bering Strait opened again between 30 and 13 KYA. It is thought that Beringia was not covered with ice but was tundra, allowing for the presence of many animals that could be hunted, including mammoth, bison, horses, and camels, enabling humans to adapt and survive in this environment by hunting and walking across Beringia from the Eurasian continent to the North American continent (Reviewed in Human Evolutionary Genetics).

The other factor affecting human migration was the ice sheets covering the North American continent. About 40 KYA, the Cordilleran (western area) and Laurentide (eastern area) ice sheets covered most areas of North Canada. The height and areas occupied by the ice sheets, like the sea level, also changed during the last ice age. Animals, including humans, could not pass through the ice sheets when they grew. In the warmer period, the volume of the ice sheets decreased enough to open ice-free corridors along the Pacific coast and Plains east of the Canadian Rockies, and many species spread from Alaska to the south, or vice versa. Precise calculation of timing for the opening of the coastal and interior corridors is still difficult because various Cordilleran glaciers behaved differently. Figure 11.1a indicates the fluctuation of sea levels during the past 24 KY since the last glacial period. At the "last glacial maximum," sea level was lower than present by approximately 120 m. Sea level did not rise constantly with the passage of time, but instead rose rapidly every 400–500 years during a "Meltwater Pulse." Sea level rose between 16 and 25 m during each Meltwater Pulse. Meltwater Pulse 1A was one of the highest rates of post-glacial sea level rise. It is thought that the coastal corridor opened by at least 15 KYA, whereas the interior corridor opened between 14 and 13.5 KYA (Mandryk et al. 2001; Dyke 2004) (Fig. 11.1b).

11.2.2 First Humans of the American Continents

Before migration into Beringia, humans had to adapt and establish their lives in the severe environments of the Arctic area (Siberia in the Eurasian continent). By 32 KYA, people survived using some stone artifacts. The Yana Rhinoceros Horn site, which is located along the Yana River in the northwest of Beringia, provides some historical information of this time (Pitulko et al. 2004). The earliest well-verified archeological evidence was found in central Alaska at Swan Point, from eastern Beringia approximately 14 KYA (Holmes and Crass 2003). The features of their artifacts are similar to the ones found in the late Upper Paleolithic remains in central Siberia.

Representative examples for sites containing Paleo-Indian remains are shown in Fig. 11.2. Most assessed remains on the south side of the Canadian ice sheets were Clovis dated 13.2-13.1 KYA, indicating their presence in the North American continent as mobile hunter-gatherers (Haynes 2002). Simultaneous existence of Clovis cultures across the North American continent indicates that they expanded around the continent quite rapidly, within a few hundred years. They hunted mammoth and mastodon regularly, and it is thought that they were associated with the extinction of these big animals in the North American continent. The Clovis culture, however, was soon taken over by other styles of artifacts. In Central and

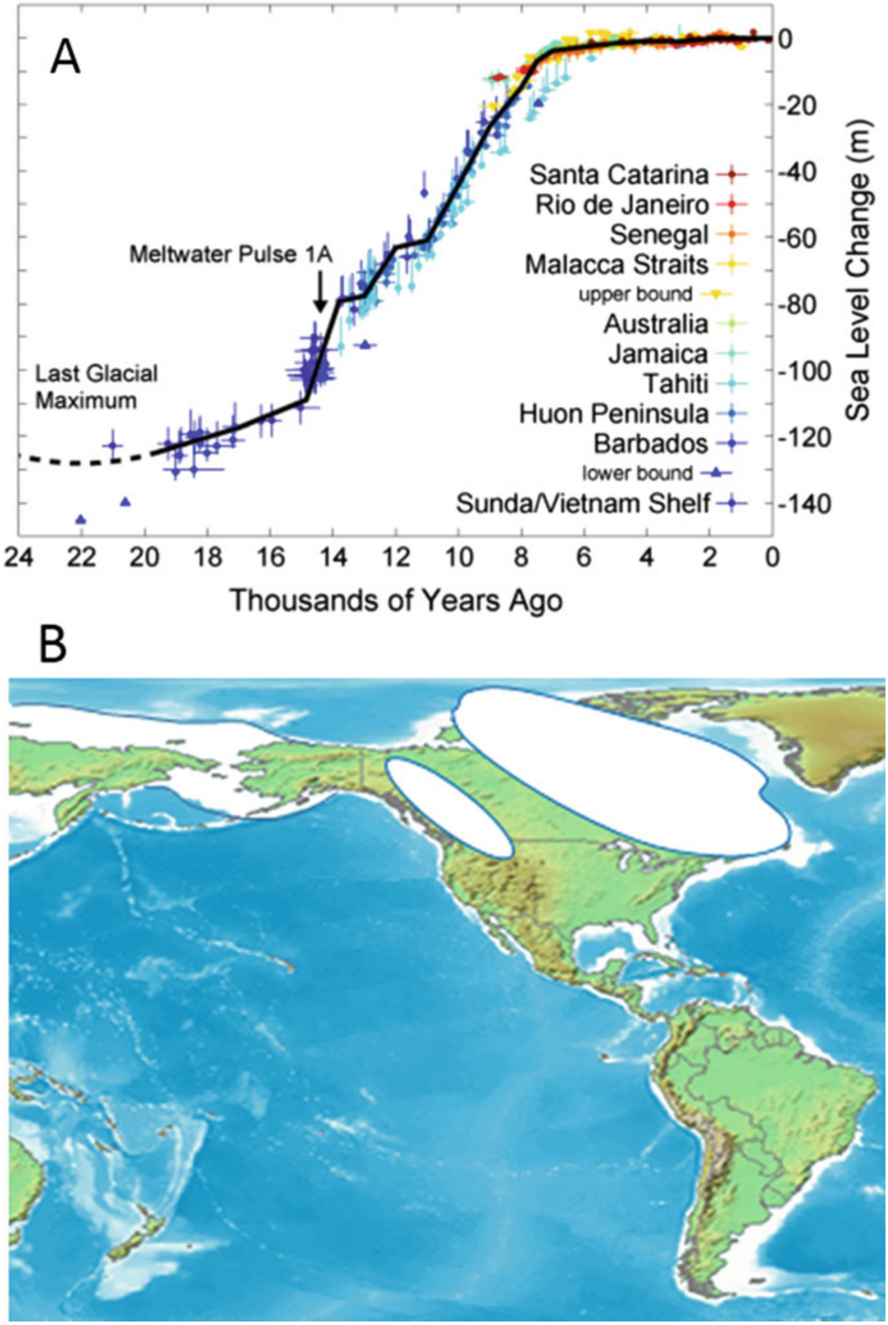


Fig. 11.1 Fluctuation of sea levels, and formation of the Beringia land bridge and ice-free corridors. (**a**) Sea level change since the last glacial age. Image created by Robert A. Rohde/Global Warming Art (https://commons.wikimedia.org/wiki/File:Post-Glacial_Sea_Level.png). (**b**) Formation of the Beringia and ice-free corridors. White areas indicate glaciers. On the North American continent, the inland part between the two large glaciers and the pacific coast were ice-free during warmer periods as described in the text

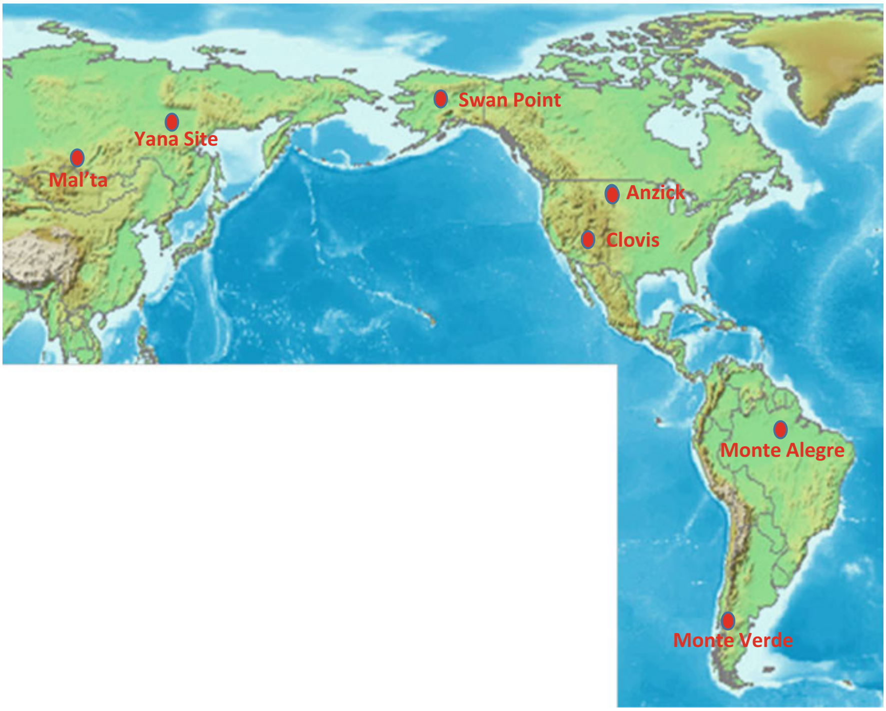


Fig. 11.2 Location of archeological remains mentioned in the text

South America, in contrast, few Clovis remains have been found (Morrow 2006). Many researchers now accept that Clovis and other people(s) lived in the American continents by 13 KYA. Because of the abundance of their remains, the “Clovis-first” hypothesis has persisted, which suggests that the Clovis people were the first humans to colonize the American continent, and then migrated south through an ice-free corridor. Was Clovis really the first American? Although some remains have been found in both American continents, the reliability is low for most of them. Among them, the Monte Verde site in Chile, from 14.6 KYA, is widely accepted as a pre-Clovis site (Dillehay 1997) and supports a model of pre-Clovis migration along the Pacific coast (Reviewed by Dixon 2001). In addition, a variety of artifacts have been found in these areas, both on the Pacific coastal side and in the Amazonian area (Fig. 11.2). For example, the remains from the cave “Caverna da Pedra Pintada,” near Monte Alegre in Amazonian Brazil, contain rock paintings and biological remains estimated from 13.2 to 12.5 KYA (Reviewed in Human Evolutionary Genetics 2nd Edition). Styles of artifacts in the Central and South Americas are distinct from Clovis, and they took place at roughly the same time.

Another question related to Paleo-Indian evolutionary history is whether they are ancestors of recent Native Americans or not. This will be discussed later in Sect. 11.3.

11.3 Genetic Analysis of Contemporary Native Americans

"The history of the earth is recorded in the layers of its crust; the history of all organisms is inscribed in the chromosomes" (Reviewed by Crow JF 1994). This is the word of world famous plant geneticist Dr. Hitoshi Kihara, who established the concept of "genome" as the minimum set containing all the essential genes. In the former section, we discussed the remains of civilizations recessed in layers of earth. In this section, we will consider the history of humans *inscribed* in our genome.

There are three different genomic regions used to analyze the history of humans: mitochondrial DNA (mtDNA), Y-chromosome markers, and autosomal chromosomes. As paternal mitochondria are selectively eliminated after fertilization in most animals including humans (Reviewed by Song et al. 2014), maternal lineage of a person can be tracked with mtDNA. Paternal lineage can be traced with the Y-chromosome because females do not have a Y-chromosome. Both mtDNA and Y-chromosome are small in size; therefore, polymorphic regions had been identified long before the completion of human whole genome sequencing and have been extensively used for genetic analysis. Some regions on the autosomal chromosomes were also highly polymorphic, and were utilized in addition to mtDNA and the Y-chromosome for investigations of the human genome. Those regions include HLA, cytochrome P450 genes (CYP), and repeat numbers in microsatellites or minisatellites. After whole genome sequencing was complete, a number of additional polymorphic markers were identified and accuracy of genome analysis was improved. Since then, along with the International Hap Map data for SNPs (The International HapMap 3 Consortium 2010), differences inscribed within the entire genome are potential targets for genetic analysis.

All three genome resources indicate that contemporary Native Americans came from Asia (Merriwether 2006; Karafet et al. 2006). The genome analysis of Native Americans, however, retains some serious problems. The first one originates as a regional issue. Native Americans living near a big city or civilized areas, with constant contact with other cultures and tribes and races, tend to admix and lose their genetic purity. Those who are pure Native American live in isolated regions that are hard to access, making it difficult to collect biomaterial as a source of DNA samples. Globalization in recent years has opened up isolated areas, resulting in the movement of young people to big cities and, consequently, old villages are gradually disappearing. Admixture with outside world people is inevitable in globalization. The second problem is associated with historical issues specific to American continents. Contrary to racial clashes and ethnic feuds among people living in neighboring areas of the Eurasian continent, completely different groups of people were coming to the American continents from other continents carrying far advanced weapons, around the fifteenth to the seventeenth century. These movements caused destruction of some Native Tribes and promoted genetic admixture of the Native Americans with those people. When we analyze the genomes of Native Americans, we should take care to remove the samples that have undergone such "artificial" admixture and select the "pure genome" as much as possible. Based on this perspective, genome resources of Native Americans obtained from the USA are not commonly utilized for genome analysis.

11.3.1 Genome Analysis with mtDNA and the Y Chromosome

Recent analysis showed that mtDNA of all humans can be tracked back to only one sequence. This sequence is referred to as "mitochondrial Eve," which is the maternal most recent common ancestor (MRCA) for all modern humans (van Oven and Kayser 2009; Behar et al. 2012; http://www.phylotree.org/tree/main.htm). Similar to mtDNA, the origin of the Y chromosome also can be traced back to a common ancestor (International Society of Genetic Genealogy, http://www.isogg.org/). During a migration that covered tens of thousands of years after leaving Africa, mutations have been accumulating within genomic DNA including mtDNA and the Y chromosome. The route of migration can be visualized as a phylogenetic tree of mtDNA and Y chromosome lineage as shown in Figs. 11.3a and 11.4a (http://www.phylotree.org/tree/index.htm, http://www.scs.illinois.edu/~mcdonald/WorldHaplogroupsMaps.pdf for mtDNAs; http://www.isogg.org/tree/ISOGG_YDNATreeTrunk.html, http://www.scs.illinois.edu/~mcdonald/WorldHaplogroupsMaps.pdf for Y chromosomes).

A group of people sharing the same series of mutations on their mitochondrial genome or Y chromosome is called as haplogroup. Recent research revealed precise world distribution of haplogroups of mtDNA and the Y chromosome (Figs. 11.3b, 11.4b). For mtDNA, haplogroup L is only observed in Africa, whereas the distribution pattern of haplogroups in the Eurasian continent is complex. Modern Native Americans retain simple distribution patterns compared to the Eurasian complexity; these include mtDNA haplogroups A, B, C, D, and X, all of which are found among indigenous peoples in southern Siberia, from the Altai to Amur region (Derenko et al. 2001; Starikovskaya, et al. 2005; Zegura et al. 2004a, b). Within these haplogroups, three subclades of C1 sub-haplogroup are widely distributed among North, Central, and South America, whereas they are absent in Asia. This suggests that the subclades were established after settlement of the American continents (Tamm et al. 2007). Distribution of the Y chromosome shows patterns similar to mtDNA; haplogroups A, B, E are mainly observed in Africa, and Europe shows a complex pattern of haplogroups. The proportion of Native Americans retaining haplogroups C and Q is large; the genetic variation of the Y chromosome is extremely small even when compared to the diversity in Asia. Haplogroup Q is not observed in other areas except northeast areas of the Eurasian continent. Moreover, as a result of admixture with European colonists and African slaves, persons retaining haplogroups R and E, which are dominant in Europeans and Africans, respectively, are commonly observed (Zegura et al. 2004a, b; Stefflove et al. 2009).

Analysis of remains and skulls of Paleo-Indians helped develop a hypothesis that they came to the American continent(s) first, and was then replaced by ancestors of modern Native Americans (Swedlund and Anderson 1999; Owsley and Jantz 2001).

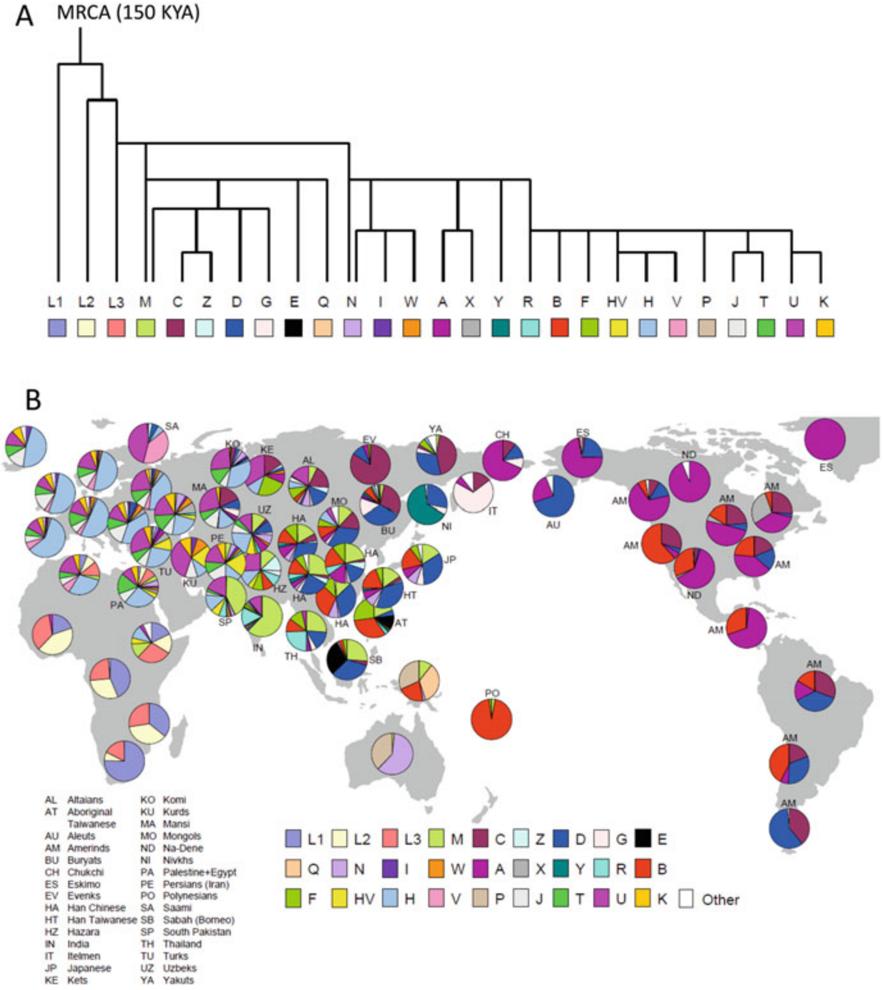


Fig. 11.3 Variation of mtDNA. (**a**) mtDNA lineage from MRCA. (**b**) World distribution of mtDNA haplogroup. These figures are reprinted with permission from the copyright holder. The original figures are in the "WorldHaplogroupsMaps" (http://www.scs.illinois.edu/~mcdonald/WorldHaplogroupsMaps.pdf)

It has long been considered that contemporary Native Americans were established from several waves of prehistoric migrations (Greenberg et al. 1986a, b). Genetic data, however, do not support these hypotheses. As all major haplogroups of mtDNA and Y-chromosome in Native Americans originated from central Asia and all haplotypes share a coalescent date, modern Native Americans might have spread from a single ancestral gene pool (Merriwether 2006; Zegura et al. 2004a, b; Wang et al. 2007). Recent analysis of mtDNA further suggests that current Native Americans were from a founding population of less than 5000 individuals (Kitchen et al. 2008).

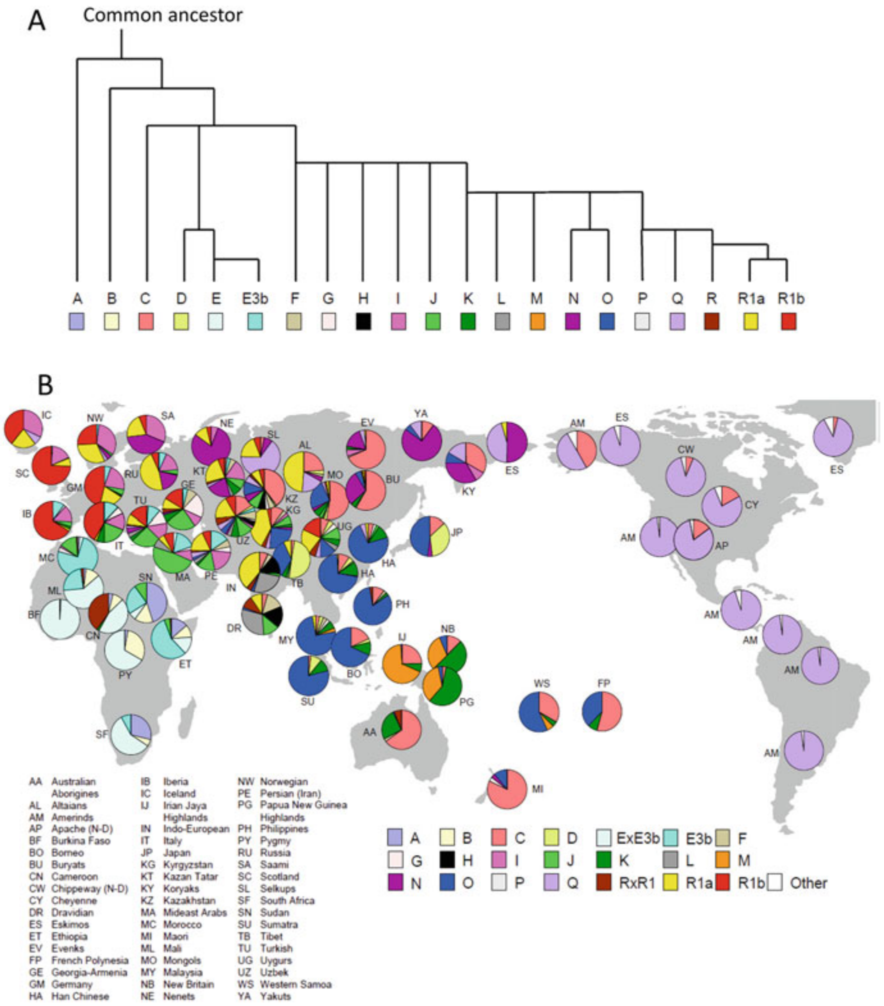


Fig. 11.4 Variation of Y chromosome markers. (**a**) Y chromosome lineage from common ancestor. (**b**) World distribution of Y chromosome haplogroup. These figures are reprinted with permission from the copyright holder. The original figures are in the "WorldHaplogroupsMaps" (http://www.scs.illinois.edu/~mcdonald/WorldHaplogroupsMaps.pdf)

11.3.2 Genome Analysis with Nuclear DNA

During the last 10 years, genome-wide analysis of nuclear DNA has been feasible due to the release of whole genome sequence of humans and accumulation of SNP data by the effort of Hap Map projects. The first genome-wide analysis of Native Americans was reported by Wang in 2007, using 678 microsatellite loci (Wang et al.

2007). Following the microsatellite analysis, genome-wide analysis with approximately 365,000 SNPs was also reported (Reich et al. 2012).

The results of genome-wide SNP analysis showed that all Native Americans and northeast Siberians (Chukche, Naukan, and Koryak) diverged from the Asian population (Fig. 11.5). This model is consistent with mtDNA analysis (Kitchen et al. 2008). They also revealed that the Arctic population first separated from an Asian population, then northern North American, North American, Southern Mexican, Lower Central American, and finally South American populations branched off one by one. This suggests that prehistoric migration occurred in a direction from north to south. Genetic variation data also support this migration pathway because the variation is reduced relative to the distance from the Bering Strait. Contrary to the generally accepted hypothesis in which the ancestors of Native Americans passed through ice-free corridors toward the south, as mentioned in Sect. 11.2.2, genome-wide SNP data showed that the correlation between genetic diversity and the distance from the Bering Strait is most parsimonious if ancestors migrated south along the coastline.

Do genome-wide analyses provide information on the number of prehistoric immigration waves to the American continents? Contemporary Native Americans have a wide variety of languages as shown in Fig. 11.6, including over 200 dialects. These languages are divided into three major independent language groups: Eskimo-Aleut (the Arctic area including Alaska, Canada and Greenland), Na-Dene (mainly in Canada), and Amerind (mainly in South America and the USA). Based on this classification, many linguists consider three waves of migrations (Greenberg et al. 1986a, b). Analysis of mtDNA or the Y chromosome has not yet clarified whether there was only a single or multiple migration waves. Genome-wide SNP data described above supported a model of at least three episodes of gene flow into the American continents from the Asian population; nearly 90% of Amerind tribes analyzed originated from only “the First American,” whereas 50% of the ancestors of Eskimo-Aleut were identified as a second wave of gene flow from Asia. Interestingly, 10% of the genome of Chipewyan living in Canada using Na-Dane language consists of a third wave. These admixture patterns are consistent with the geographic location and the distribution of the language groups. The higher admixture rate at the Arctic area may be the result of a backward migration from the North American continent to northeast Siberia.

Compared to the dynamic genetic admixture around the Arctic area and northern part of the North American continent, gene flow among Central and South American tribes is very low. Of special note in these areas are Chibchan, a Meso-American population, and the genetic flow from other continents. Chibchan is one of the language families in South America and people that speak this language live on both sides of the Isthmus of Panama. Their genomes retain both Northern and Southern genetic characteristics, suggesting admixture of both populations by a “back migration” after establishment of subpopulations. In the case of Meso-American populations, they have an extremely low level of genetic drift, indicating a larger effective population size since the settlement by ancestors of Native Americans. In contrast to the Meso-American population, South American tribes in the

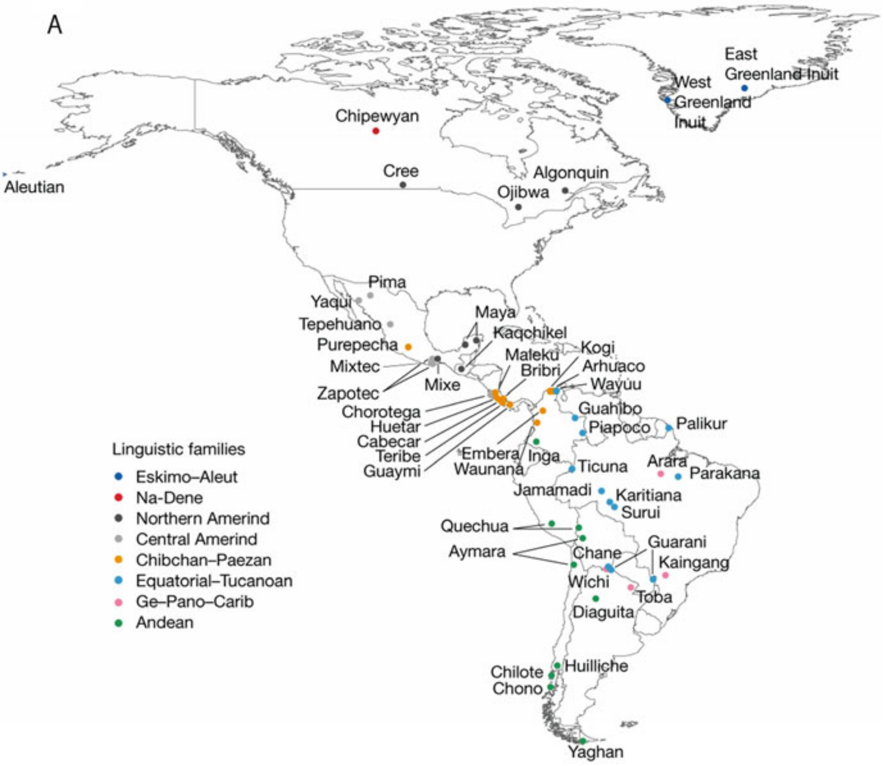


Fig. 11.5 Phylogenetic tree to search ancestors of Native Americans. (**a**) The locations from where samples were collected. (**b**) Phylogenetic tree to search ancestors of Native Americans. These figures are reprinted with permission from the copyright holder. The original figures are in the article, Reich et al. Nature, 488, 370, 2012 (doi: https://doi.org/10.1038/nature11258)

coastal region retain genetic features of both Europeans and Africans due to admixture with those people after the Age of Exploration around the fifteenth century.

The progressive advances in NGSs in recent years enable us to sequence genomes of ancient humans in large scale. NGSs are a powerful tool to analyze genetic alteration directly in chronological order. In fact, whole genome sequencing of a child (MA-1), who lived 24 KYA in Mal'ta, south-central Siberia, was recently completed (Raghavan et al. 2014). The sequence data showed that the MA-1 genome did not originate from eastern Asia but was closer to western Eurasians and Native Americans. These results strongly suggest that the origins of Native Americans are not only eastern Asia but also south-central Siberia. Present calculations estimate that 14–38% of the genome of Native Americans has derived from south-central Siberians, and that the gene flow occurred after Native Americans branched from Asian populations. In addition to this discovery, other exciting results were obtained from another ancient human genome. An adult male (named Anzick-1), who lived in

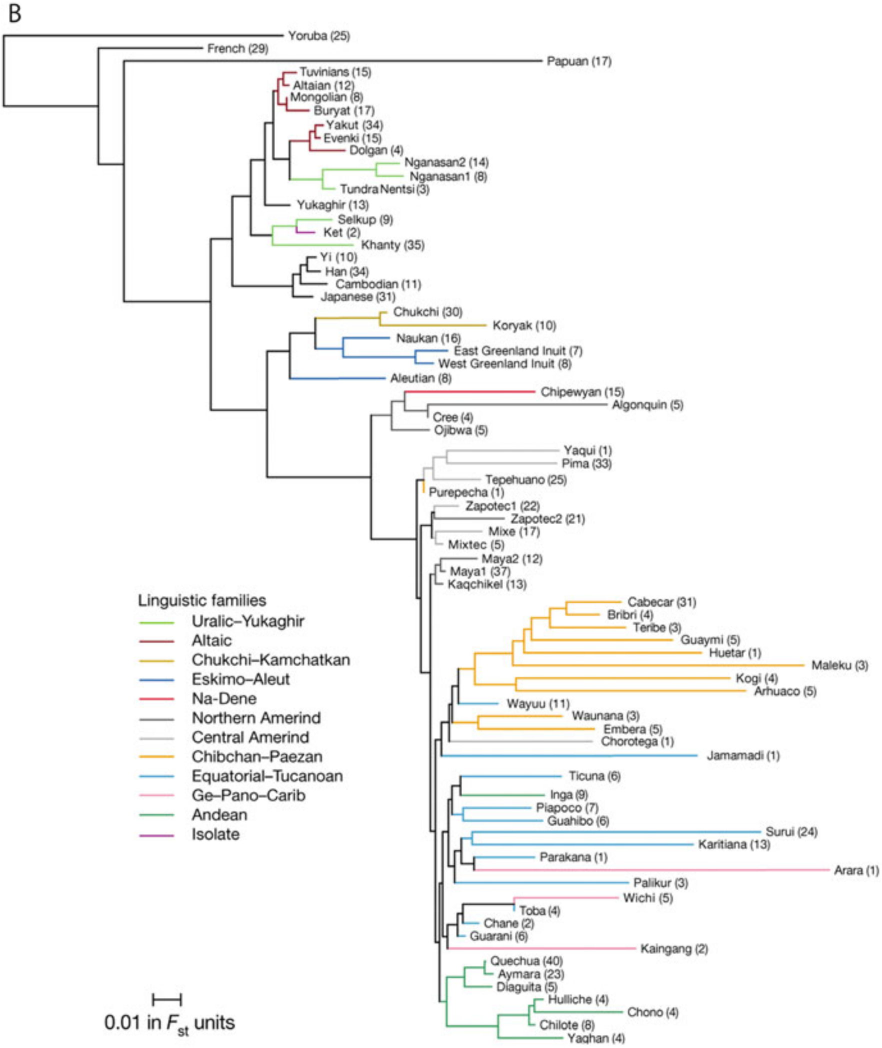


Fig. 11.5 (continued)

Anzick around 12,707–12,556 calendar years BP (*B*efore *P*resent) using Clovis tools, demonstrated the same gene flow patterns as MA-1 even though the gene flow from south-central Siberia was 12.6 KYA (Rasmussen et al. 2014). Interestingly, Anzick-1 was genetically closest to Native Americans than other modern people. The results of lineage analysis among Northern Amerind, Southern Amerind, and Anzick-1 revealed that divergence between Northern and Southern Amerind was much older than the divergence of Anzick-1. How much alteration have the

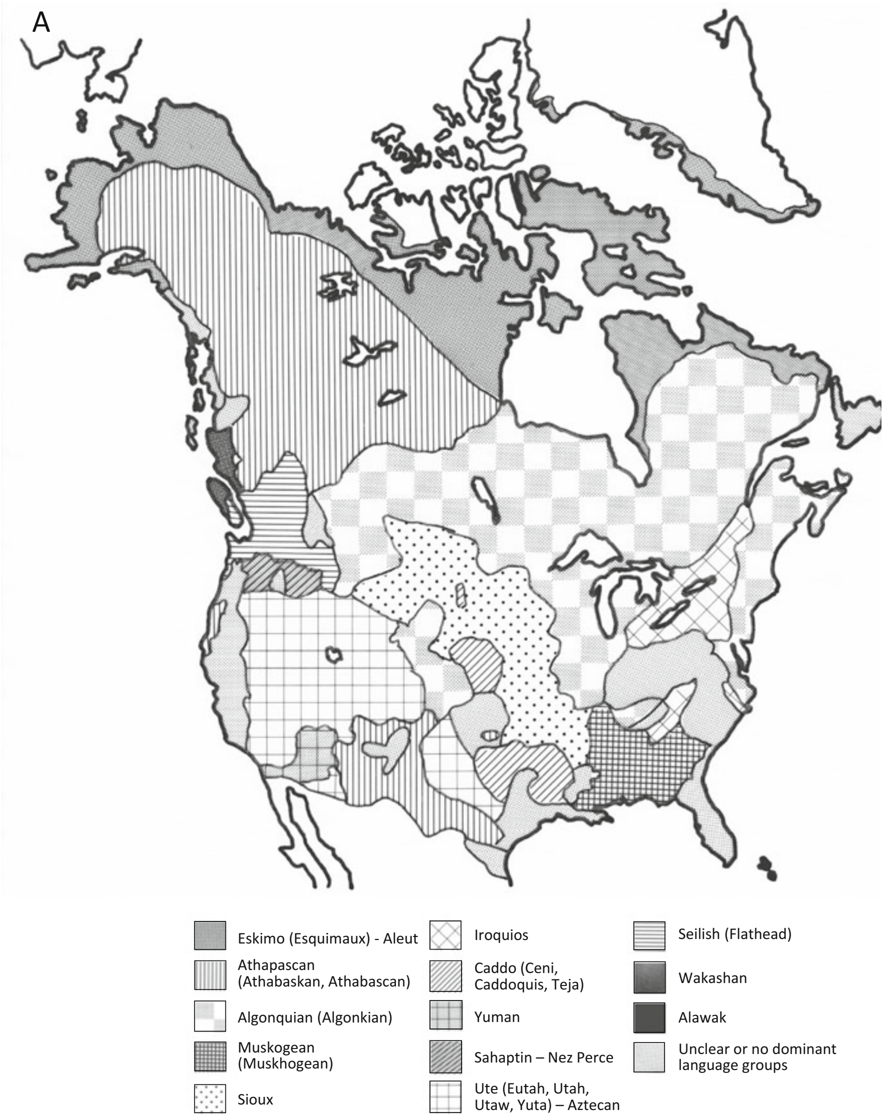

A
Eskimo (Esquimaux) - Aleut
Athapascan (Athabaskan, Athabascan)
Algonquian (Algonkian)
Muskogean (Muskhogean)
Sioux
Iroquios
Caddo (Ceni, Caddoquis, Teja)
Yuman
Sahaptin – Nez Perce
Ute (Eutah, Utah, Utaw, Yuta) – Aztecan
Seilish (Flathead)
Wakashan
Alawak
Unclear or no dominant language groups

Fig. 11.6 Language groups in the North and the South American continents. (**a**) Language groups in the North American continent. (**b**) The South American continent. These figures are reprinted with permission from the copyright holders. The original figures are in "Sekai minzoku jiten," Edited by Tsuneo Ayabe, Kobundo, Tokyo, 2000; written in Japanese. A contains a slight modification to the original figure

genomes of Native Americans experienced since they branched from Asian populations? We do not have the exact answer for this question at present.

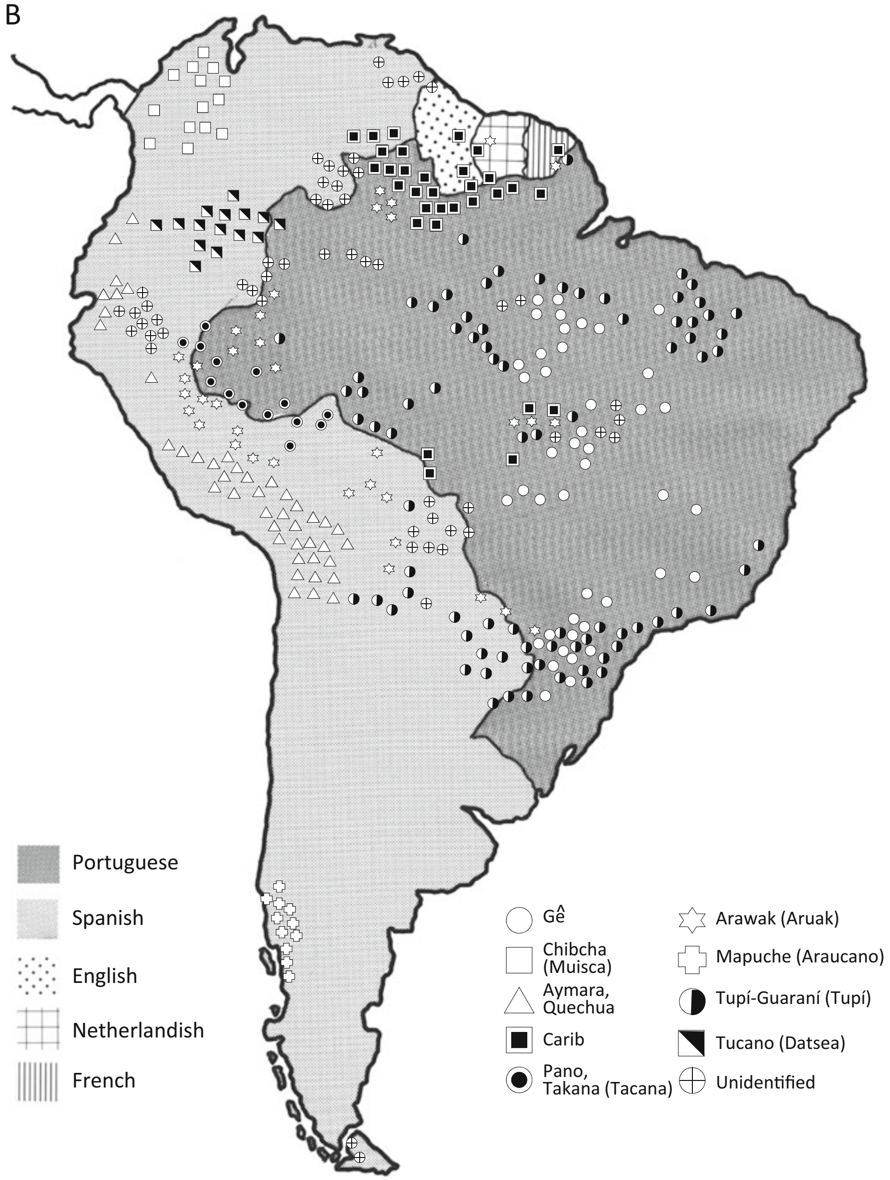


Fig. 11.6 (continued)

When we focus on adaptation events, there is no evidence of evolutionary change specific to Native Americans. It is likely that the time period since their ancestors migrated to the American continents is too short for any genetic alterations to have occurred and persisted in the population. The current progress of NGS analysis will unveil many events that we have not been able to address so far.

The field of Human Genetic Research is now entering a new phase: mining large datasets. Consequently, many scientists are entering into this research field with backgrounds far from Human Genetics, as mentioned in Sect. 11.1. However, deep historical insight in combination with a holistic vision of the relationships among tribes is essential to decipher the results of genome analyses. We should not forget this important point or the data would mislead us.

11.4 Genetic Complexities and Ethical Issues on Genome Analysis of Native Americans

Genome data are necessary for analysis of population genetics. Generally, population geneticists utilize genome data deposited into public databases, or collect human biomedical materials by themselves on site. Collection of samples presents some challenges; it takes a long time, sometimes exposes the collector to dangers, requires laborious negotiation for getting informed consent, requires special handling and equipment to maintain the samples in good condition for extraction of genomic DNA, etc. To avoid these problems and facilitate access to genome data, some public resource centers release a wide variety of population samples for research in recent years, which includes genomic DNAs, cell lines, and other biomaterials. The Human Genome Diversity Project Cell Line Panel (HGDP-CEPH; http://www.cephb.fr/en/index.php), Coriell Institute (https://catalog.coriell.org/1/NHGRI), and RIKEN BioResource Center (RIKEN BRC; http://www.brc.riken.jp/lab/cell/english/), for instance, maintain a vast collection of various populations. These resource centers receive resources from the researchers who collected them, preserve the quality, and provide the resources to other researchers working for academic research institutes.

Consideration of genetic background must be integrated into genome analyses; for example, we must consider whether target individuals or tribes are racially pure or mixed-breed, and when admixture occurred if they are not pure. Genetic admixture is closely linked with history. Genetic complexity is higher in the American continents than in the Eurasian continent. Many people retain genetic background from both Native Americans and Europeans, and even from Africans. It is due to their long and complex history, and this genetic complexity makes genome analysis difficult and confusing.

Native Americans had developed several cultures in both the North and the South American continents thousands of years ago. Highly developed civilizations with large cities were built by the sixteenth century, especially in Central and South America, during the Age of Exploration when the Spanish and Lusitanian explorers intruded into American continent. Then, the Europeans conquered the American world and imported African slaves as labors, established a social hierarchy of victor and vanquished. During this time, a large number of Native Americans died from diseases brought from the Old World and that the Native Americans had never been exposed to before. Because of this history, racial mingling has progressed gradually over many generations.

Within the history of the American continents, Native Americans have often been the targets of racial discrimination (Summarized by Burger 1990). We should be very careful that the results of genome analyses will not be used to promote discrimination. Some organizations and resource centers take countermeasures against these ethical issues; HGDP-CEPH will not provide their resources to profit-oriented organizations and users, the Coriell Institute requests that users disclose the purpose of their research so that it can be evaluated by an ethical committee outside the Coriell Institute. In the case of RIKEN BRC, the office requests evidence of approval from an ethical review committee from the institutions of the users. As an individual researcher, we should disclose the purpose of each project to the donors who provide biomaterials used for genome analysis, and then obtain their written consents. Informed consent is considered one of the methods available to respect the human rights of donors.

Acknowledgement We would like to thank Drs. Rumiko Saito and Satoshi Nishikawa, and Ms. Natsumi Konno of Tohoku Medical Megabank Organization for their help to collect manuscripts and discussions.

References

Behar DM, van Oven M, Rosset S, Metspalu M, Loogväli EL, Silva NM, Kivisild T, Torroni A, Villems R (2012) A “Copernican” reassessment of the human mitochondrial DNA tree from its root. Am J Hum Genet 90(4):675–684. https://doi.org/10.1016/j.ajhg.2012.03.002. http://www.mtdnacommunity.org/human-mtdna-phylogeny.aspx

Burger J (1990) The Gaia atlas of first peoples—a future for the indigenous world. Gaia Books Ltd., London

Crow JF (1994) Hitoshi Kihara, Japan’s pioneer geneticist. Genetics 137:891–894

Derenko MV, Grzybowski T, Malylarchuk BA, Czarny J, Mescicka-Sliwka D, Zakharov IA (2001) The presence of mitochondrial haplogroup x in Altaians from South Siberia. Am J Hum Genet 69(1):237–241

Dillehay TD (1997) The archaeological context and interpretation. In: Monte Verde: a late Pleistocene settlement in Chile, vol 2. Smithsonian Institution Press, Washington, DC

Dixon EJ (2001) Human colonization of the Americas: timing, technology and process. Quat Sci Rev 20:277–299

Dyke AS (2004) In: Ehlers J, Gibbard PL (eds) Quaternary glaciations—extent and chronology, part II: North America. Elsevier, Amsterdam, pp 373–424

Greenberg JH, Turner CG II, Zegura SL (1986a) The settlement of the Americas: a comparison of the linguistic, dental, and genetic evidence. Curr Anthropol 27(5):477–497

Greenberg JH, Turner CG II, Zegura SL, Campbell L, Fox JA, Laughlin WS, Szathmary EJE, Weiss KM, Woolford E (1986b) The settlement of the Americas: a comparison of the linguistic, dental, and genetic evidence. Curr Anthropol 27:477–497

Haynes GA (2002) The early settlement of North America: the Clovis Era. Cambridge Univ. Press, Cambridge

Holmes CE, Crass BA (2003) Early cultural components in central Alaska: an update from Swan Point. In: Paper presented at the 30th annual meeting of the Alaska Anthropological Association

Karafet TM, Zegura SL, Hammer MF (2006) Environment, origins, and population. In: Ubelaker DH (ed) Handbook of North American Indians, vol 3. Smithsonian Institution Press, Washington, DC, pp 831–839

Kitchen A, Miyamoto MM, Mulligan CJ (2008) A three-stage colonization model for the peopling of the Americas. PLoS One 3(2):e15962008. https://doi.org/10.1371/journal.pone.0001596

Lander ES, Linton LM, Birren B et al (2001) Initial sequencing and analysis of the human genome. Nature 409(6822):860–921

Mandryk CAS, Hosenhans H, Fedje DW, Mathewes RW (2001) Late quaternary paleoenvironments of northwestern North America: implications for inland versus coastal migration routes. Quat Sci Rev 20:301–314

Merriwether DA (2006) Environment, origins, and population. In: Ubelaker DH (ed) Handbook of North American Indians, vol 3. Smithsonian Institution Press, Washington, DC, pp 817–830

Morrow JE (2006) In: Gnecco C (ed) Paleoindian archaeology: a hemispheric perspective. Univ. Press of Florida, Gainesville

Owsley DW, Jantz RL (2001) Archaeological politics and public interest in Paleoamerican studies: lessons from Gordon Creek woman and Kennewick man. Am Antiq 66(4):565–575

Pitulko VV, Mikolsky PA, Girya EY, Basilyan AE, Tumskoy VE, Koulakov SA, Astakhov SN, Pavlova EY, Anisimov MA (2004) The Yana RHS site: humans in the Arctic before the last glacial maximum. Science 303:52–56

Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, Moltke I, Rasmussen S, Stafford TW Jr, Orlando L, Metspalu E, Karmin M, Tambets K, Rootsi S, Mägi R, Campos PF, Balanovska E, Balanovsky O, Khusnutdinova E, Litvinov S, Osipova LP, Fedorova SA, Voevoda MI, DeGiorgio M, Sicheritz-Ponten T, Brunak S, Demeshchenko S, Kivisild T, Villems R, Nielsen R, Jakobsson M, Willerslev E (2014) Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature 505(7481):87–91. https://doi.org/10.1038/nature12736

Rasmussen M, Anzick SL, Waters MR, Skoglund P, DeGiorgio M, Stafford TW Jr, Rasmussen S, Moltke I, Albrechtsen A, Doyle SM, Poznik GD, Gudmundsdottir V, Yadav R, Malaspinas AS, White SS 5th, Allentoft ME, Comejo OE, Tambets K, Eriksson A, Heintzman PD, Karmin M, Korneliussen TS, Meltzer DJ, Pierre TL, Stenderup J, Saag L, Warmuth VM, Lopes MC, Malhi RS, Brunak S, Sicheritz-Ponten T, Barmes I, Collins M, Orland L, Balloux F, Manica A, Gupta R, Metspalu M, Bustamante CD, Jakobsson M, Nielsen R, Wellerslev E (2014) The genome of a late pleistocene human from a Clovis burial site in western Montana. Nature 506 (7487):225–229. https://doi.org/10.1038/nature13025

Reich D, Patterson N, Campbell D et al (2012) Reconstructing native American population history. Nature 488:370–375

Song WH, Ballard JW, Yi YJ, Sutovsky P (2014) Regulation of mitochondrial genome inheritance by autophagy and ubiquitin-proteasome system: implications for health, fitness, and fertility. Biomed Res Int 2014:981867

Starikovskaya EB, Sukernik RI, Derbeneva OA, Volodko NV, Ruiz-Pesini E, Torroni A, Brown MD, Lott MT, Hosseini SH, Huoponen K, Wallace DC (2005) Mitochondrial DNA diversity in indigenous populations of the southern extent of Siberia, and the origins of native American haplogroups. Ann Hum Genet 69(Pt1):67–89

Stefflove K, Dulik MC, Pai AA, Walker AH, Zeigler-Johnson CM, Gueye SM, Schurr TG, Rebbeck TR (2009) Evaluation of group genetic ancestry of populations from Philadelphia and Dakar in the context of sex-biased admixture in the Americas. PLoS One 4(11):e7842

Swedlund A, Anderson D (1999) Gordon Creek woman meets Kennewick Man: new interpretations and protocols regarding the peopling of the Americas. Am Arch 64(4):569–576

Tamm E, Kivisild T, Reidla M, Metspalu M, Smith DG, Mulligan CJ, Bravi CM, Rickards O, Martinez-Labarga C, Khusnutdinova EK, Fedorova SA, Golubenko MV, Stepanov VA, Gubina MA, Zhadanov SI, Ossipova LP, Damba L, Voevoda MI, Dipierri JE, Villems R, Malhi RS (2007) Beringian standstill and spread of Native American founders. PLoS One 2(9):e829

The International HapMap 3 Consortium (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467:52–58

The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299–1320

The International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature:851–862

Van Oven M, Kayser M (2009) Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat 30(2):E386–E394. https://doi.org/10.1002/humu.20921. http://www.phylotree.org/tree/main.htm

Wang S, Lewis CM, Jakobsson M, Ramachandran S, Ray N, Bedoya G, Rojas W, Parra MV, Molina JA, Gallo C, Mazzotti G, Poletti G, Hill K, Hurtado AM, Labuda D, Klitz W, Barrantes R, Bortolini MC, Salzano FM, Petzl-Erler ML, Tsuneto LT, Llop E, Rothhammer F, Excoffier L, Feldman MW, Rosenberg NA, Ruiz-Linares A (2007) Genetic variation and population structure in native Americans. PLoS Genet 3(11):e185

Zegura SL, Karafet TM, Zhivotovsky LA, Hammer MF (2004a) High-resolution SNPs and microsatellite haplotypes point to a single, recent entry of native American Y chromosomes into the Americas. Mol Biol Evol 21(1):164–175

Zegura SL, Karafet TM, Zhivotovsky LA, Hammer MF (2004b) High-resolution SNPs and microsatellite haplotypes point to a single, recent entry of native American Y chromosomes into the Americas. MolBiolEvol 21(1):164–175

Chapter 12
Simulations of Human Dispersal and Genetic Diversity

Mathias Currat, Claudio S. Quilodrán, and Laurent Excoffier

Abstract Humans are a highly mobile species that has colonized the entire globe in a few tens of thousands of years after it went out of Africa. There are still many unknowns about the routes followed by our ancestors during this expansion process, which has been influenced by various environmental, biological, and cultural factors, but these migrations have contributed to shape the genetic diversity of our species. A powerful approach to study the consequences of human dispersal on our genome is the modelling of complex evolutionary scenarios via computer simulation. In this chapter, we present three types of approaches used to simulate human dispersal in a geographic landscape. We focus on a spatially explicit method, simulating the demographic and migratory dynamic of populations forward in time and their resulting genetic diversity backward in time using the coalescent. We describe this approach and illustrate its interest with two important results: the process of gene surfing during population expansion and the genetic consequences of hybridization during species expansions. We show that a relatively simple scenario of global expansion of *Homo sapiens* from Africa, with rare hybridization events with archaic humans, such as Neanderthals or Denisovans, over a large geographic area reasonably explains the introgression pattern of archaic DNA in the genome of our species.

Keywords Human evolution · Computer simulations · Competition · Hybridization · Admixture · Range expansion · Neanderthal

M. Currat (✉) · C. S. Quilodrán
Anthropology, Genetics and Peopling History Lab (AGP), Department of Genetics and Evolution – Anthropology Unit, University of Geneva, Geneva, Switzerland
e-mail: mathias.currat@unige.ch

L. Excoffier (✉)
Computational and Molecular Population Genetics Lab, Institute of Ecology and Evolution, University of Berne, Berne, Switzerland

Swiss Institute of Bioinformatics, Lausanne, Switzerland
e-mail: laurent.excoffier@iee.unibe.ch

N. Saitou (ed.), *Evolution of the Human Genome II*, Evolutionary Studies,
https://doi.org/10.1007/978-4-431-56904-6_12

12.1 Introduction

Homo sapiens is a highly mobile species. From its likely place of origin in Africa, it colonized the entire planet in less than 50,000 years. This early colonization phase was constrained by geographic and environmental boundaries (coastlines, mountains, forests), but also by the presence of other archaic *Homo* species (Neanderthals, Denisovans, *H. erectus*), who were present in most of the old world (Veeramah and Hammer 2014). Given the fossil record, Africa and the Near-East have been occupied by anatomically modern humans (AMH) before Asia, Europe, and the Americas (e.g. Henn et al. 2012). The most supported scenario for the origin of our own species is often termed "Recent African Origin" (RAO) and proposes an emergence of AMH in East Africa between 200 and 100 kya (Tattersall 2009), followed by a rapid spread toward the rest of the world starting around 60 kya (Henn et al. 2012), which led to a full replacement of all archaic human forms. The initial version of RAO has been challenged by the sequencing of the Neandertal and Denisovan genomes (Green et al. 2010; Reich et al. 2010; Prufer et al. 2014) and by statistical analyses of existing genomic data sets (Wall et al. 2013). These studies support the existence of a relatively low genetic contribution of archaic forms in current non-African humans, which probably arose through hybridization after the exit out of Africa. These new analyses changed the old RAO paradigm to a scenario one could call "RAO with hybridization" (Stringer 2014). However, very little is known on the exact extant of the interactions between AMH and Neanderthals (NE) and even less with other archaic populations.

Despite this general RAO scenario, many unknowns still exist on the migration routes followed by our ancestors. The exact dates of colonization of many regions of the world are also uncertain simply because old fossils are rare or absent in many regions, even though such an absence is not a proof that these regions have not been colonized by early humans. Consequently, several alternative hypotheses have been proposed to explain the colonization of different regions of the world (Veeramah and Hammer 2014). For instance, at least three main possible routes of migration have been proposed to explain the diffusion of AMH from Africa to East Asia and Oceania: a coastal "southern route" along the coast of the Pacific Ocean (Macaulay et al. 2005), a second more continental route in the south of Himalaya (Rasmussen et al. 2011), and another migration route North of the Himalaya (Di and Sanchez-Mazas 2011).

The dispersal of AMH out of Africa has been probably triggered by a combination of various factors such as new abilities or opportunities (due to technological advances or cultural changes), climatic variation, and demographic pressure (Powell et al. 2009; Eriksson et al. 2012; Lahr and Foley 1998). In any case, this initial global dispersal has been followed by many additional migrations, at various times and geographic scales. Climate variation played an important role in promoting those migrations through several mechanisms: first, the sea level was about 120 m below the current level during cold periods, which created many land bridges between continents and previously isolated islands (e.g. British Isles, New Guinea, Beringia);

second, glaciers and ice caps extended over much larger areas during cold periods (Ray and Adams 2001) making them uninhabitable; third, vegetation cover has also significantly changed with temperature variations, displacing towards refugia (Hewitt 2000) most animals and plants that were food sources for humans (Banks et al. 2008). In addition, at the beginning of the Holocene, most human populations have passed through a major economic and cultural transition in different parts of the world, switching from a hunter-gatherer to a food production lifestyle (Bellwood 2001). This change corresponds to the appearance of plant and animal domestication, pottery, and of sedentary lifestyle and probably involved demographic increases and large-scale migrations (Zvelebil 2001). Finally, sedentarized human populations continued to exchange migrants until present time and additional waves of migration took place around the world during the historical period, but probably at a smaller scale (Sokal 1991).

Ongoing short range migrations have left traces in the genome of contemporary humans (Ray et al. 2005), which shows a strong concordance between geography and genetics in most parts of the world (e.g. Novembre et al. 2008). For these reasons, it seems necessary to consider the spatial constraints on the dispersal of modern humans when they colonized the planet. However, it is still very difficult to consider these population dynamics when analysing genomic data, as envisioned models quickly become very complex and are difficult to parameterize and test. Fortunately, modelling and computer simulations of complex evolutionary scenarios can be used to produce realistic models of human migration over time and space and to assess their impact on current genetic and genomic diversity. Thanks to a regular increase in computer power and to the development of efficient algorithms and realistic models, simulation approaches have become essential to study human evolution. In this review, we shall describe the main modelling approaches used to study the effect of human dispersal on genetic diversity at the global or continental scale, with a focus on spatially explicit simulations. We shall also present and comment their major contributions to our understanding of human evolution and discuss future developments.

12.2 Modelling and Simulating Human Dispersal

The modelling and simulation of evolutionary scenarios using computer programs consists of a combination of mathematical models defined by a series of parameters. Typically, three different approaches have been used to simulate human dispersal in a geographic landscape: Forward diffusion models, forward individual-based simulations, and forward demographic coupled with backward genetic simulations.

Diffusion approaches are based on a set of spatially explicit differential equations specifying density and diffusion (migration) parameters in a continuous space. They are extensions of Fisher's approach to model the wave of advance of advantageous genes in one dimension (Fisher 1937). Fisher's models have since been extended to model the colonization of different continents in two dimensions (e.g. Steele et al.

1998; Fort and Pujol 2008; Martino et al. 2007; Fort et al. 2004) or to the spread of cultures or technologies in already settled habitats (Ackland et al. 2007; Fort and Mendez 1999). These diffusion models have also been refined to take into account the heterogeneity of the habitat (see Steele 2009 for a review). Note that these models can mainly predict colonization times or the spread of beneficial variants in an already colonized habitat, and they have been coupled to the generation of genetic data only recently (Barton et al. 2013a, b).

Individual-based or agent-based simulations are the most flexible way to simulate genetic data, as they allow one to simulate very specific behaviour, such as sex-specific reproductive success or migration, age-specific mortality, or migration rates. They are usually performed in discrete demes positioned on a one-dimensional (e.g. Eswaran 2002; Eswaran et al. 2005; Ramachandran et al. 2005; Fix 1997) or two-dimensional lattice (Itan et al. 2009; Rasteiro et al. 2012; Deshpande et al. 2009; Rendine et al. 1986; Barbujani et al. 1995) corresponding to stepping-stone models, but they can be extended to 2D hexagonal demes (e.g. Eriksson et al. 2012) enabling a more realistic spatial diffusion process (Lavrentovich et al. 2013). Individual-based simulations also allow one to simulate loci under arbitrarily complex selection models (Itan et al. 2009; Currat et al. 2010; Fix 1996; Peischl et al. 2013) fully or partially linked with neutral loci (e.g. Chadeau-Hyam et al. 2008). However, this complexity has a cost in terms of computing power and time, implying that a small number of individuals and populations can be modelled, even though the rescaling of mutation, migration rates, and selection coefficients can allow one to model large populations with a smaller number of individuals (e.g. Chadeau-Hyam et al. 2008). Another drawback of forward approaches is that one needs to define starting conditions or to wait that some demo-genetic equilibrium has been reached before starting recording the simulations. Nevertheless, these models have been applied to very interesting problems in human evolution, ranging from the spread of Neolithic farmers into Europe and their interaction with local hunter-gatherers (Rasteiro et al. 2012; Rendine et al. 1986; Barbujani et al. 1995; Rasteiro and Chikhi 2013), the expansion out of Africa, either with (Eswaran 2002; Eswaran et al. 2005) or without (Ramachandran et al. 2005; Deshpande et al. 2009) hybridization, or to the estimation of the geographical origin of beneficial mutation in European populations (Itan et al. 2009).

Despite their great flexibility, forward individual-based simulations can sometimes be very slow and require a lot of memory, since they simulate all the individuals of many populations. A way to solve these problems is to use forward simulations to simulate the demography of the populations and then coalescent approaches to simulate their genetic diversity. These approaches have typically been implemented in the SPLATCHE program (see Currat et al. 2004; Ray et al. 2010, and details below) or in simpler 1 D serial founder effect models (DeGiorgio et al. 2011), and a similar approach has been used to simulate genetic data linked to a site under selection (Ewing and Hermisson 2010).

One usage of these modelling approaches is to generate genetic data for various alternative hypotheses (therefore various combinations of values assigned to the parameters), in order to evaluate which is the scenario most (or less) compatible with the observed genetic data and thus to perform model choice. They can also be used to describe and study evolutionary processes and to estimate demographic or genetic parameters. In the latter case, simulations can be coupled with Bayesian inference procedures such as the approximate Bayesian computation (ABC, Beaumont et al. 2002; Wegmann et al. 2010) to estimate demographic parameters under various evolutionary scenarios.

12.3 Realistic Simulation with Spatially Explicit Simulations

Here we focus on spatially explicit simulation of human dispersal and their genetic diversity. By spatially explicit, we mean that the simulation of populations or individuals considers their geographic position. The stepping-stone model (Kimura 1953), which consists of an array of cells in one or two dimensions is the most commonly used spatially explicit model. Here, the location of each population is defined by its coordinates and because migrants are only exchanged between neighbouring populations, the effective amount of gene flow between two populations is inversely related to their distance on the grid. This leads to genetic isolation by distance (IBD) (Wright 1943), a process that played an important role in human evolution (Morton 1977, 1982), and which largely explains the correspondence between geography and genetic variation in extent human populations (e.g. Novembre et al. 2008). Note that very detailed geographic (e.g. continental contours) and environmental (e.g. mountains, deserts) information can be used in spatially explicit simulations, which is facilitated using geographic information system (GIS). Note that spatially explicit simulations can also be done by specifying an arbitrary migration matrix between all pairs of populations in a model, as implemented in programs like SIMCOAL and FASTSIMCOAL (Excoffier et al. 2000; Excoffier and Foll 2011; Laval and Excoffier 2004) or in MS (Hudson 2002). However, in this case the migration matrix can become huge and difficult to set up for models with complex geographic features, and therefore we will thus not discuss these programs below. In Table 12.1, we give a list of other spatially explicit simulation models and programs that have been used or could potentially be used in the context of human dispersal. Since their exact implementation varies from one model to the other, we redirect the readers to the original papers to get details about their underlying methodology. However, we illustrate below the implementation of the main key elements of a spatially explicit simulation framework through the program SPLATCHE that we have developed.

Table 12.1 List of spatially explicit computer programs and studies (*) where human dispersal and genetic diversity can or have been modelled, respectively. If the program is available online, the URL is given, as well as the operating system (W/M/L for Windows/Mac OSX/Linux) on which the program runs

Program name (if available) and short model description.	Related references	Available version	Dimension	OS (language)
EASYPOP: Individual-based computer program for population genetics simulations. https://www.unil.ch/dee/en/home/menuinst/open-positions-and-public-resources/softwares--dataset/softwares/easypop.html	(Balloux 2001)	Ver. 2.0.1 2006	2D	W/M (C)
SPLATCHE: Simulation of demography and resulting molecular diversity for a wide range of evolutionary scenarios, taking into account environmental heterogeneity. Coalescent. http://www.splatche.com	(Currat et al. 2004; Ray et al. 2010)	Ver. 2.01 2012	2D	M/L (C++)
NEMO: A stochastic, individual based, genetically explicit, and stochastic simulation program designed to study the evolution of life history/phenotypic traits and population genetics. http://nemo2.sourceforge.net/	(Guillaume and Rougemont 2006)	Ver. 2.2.0 2011	2D	W/M/L (C ++)
QUANTINEMO: An individual-based program for the analysis of quantitative traits with explicit genetic architecture potentially under selection in a structured population. http://www2.unil.ch/popgen/softwares/quantinemo/	(Neuenschwander et al. 2008)	Ver. 1.0.4 2011	2D	W/M/L (C ++)
CDPOP: Individual-based spatially explicit simulator of gene flow in complex landscapes. https://github.com/ComputationalEcologyLab/CDPOP	(Landguth and Cushman 2010)	Ver 1.2 2014	2D	W/M/L (Python)
GINKGO: Agent-based forward-time simulation to produce gene genealogies	(Sukumaran and Holder 2011)	Ver. 3.9.0 2010	2D	W/M/L (C ++)

(continued)

Table 12.1 (continued)

Program name (if available) and short model description.	Related references	Available version	Dimension	OS (language)
for multiple populations in a spatially explicit and environmentally heterogeneous framework. http://phylo.bio.ku.edu/ginkgo/index.html				
MARLIN: Program to create, run, analyse, and visualize spatially explicit population genetic simulations. http://www.bentleydrummer.nl/software/software/Marlin.html	(Meirmans 2011)	Ver. 1.0 2009	2D	M (C)
DIM SUM: Stand-alone Java program for the simulation of population demography and individual migration, recording ancestor-descendant relationships. http://code.google.com/p/bio-dimsum	(Brown et al. 2011)	Ver. 0.91 2010	2D	W/M/L (Java)
SIMADAPT: Spatially explicit, individual-based, landscape-genetic simulation model to represent evolutionary processes and population dynamics in changing landscapes. http://www.openabm.org/model/3137/version/8/view	(Rebaudo et al. 2013)	Ver. 8 2013	2D	W/M/L (NetLogo)
PHYLOGEOSIM: Simulation of DNA sequences under a model of coalescence on a 2-dimensional grid of populations. http://ebe.ulb.ac.be/ebe/PhyloGeoSim.html	(Dellicour et al. 2014)	Ver. 1.0 (in prep)	2D	W/M/L (Java)
* Simulation of the demic diffusion model of human expansion. Based on the model of the wave of advance of advantageous genes (Fisher 1937)	(Ramachandran et al. 2005; Rendine et al. 1986; Edmonds et al. 2004; Sgaramella-Zonta and Cavalli-Sforza 1973)	–	2D	–
* Simulation of microevolution in European populations, incorporating	(Barbujani et al. 1995)	–	2D	–

(continued)

Table 12.1 (continued)

Program name (if available) and short model description.	Related references	Available version	Dimension	OS (language)
genetic drift and gene flow (isolation by distance)				
* Analytical dynamic colonization population genetics model in a one-dimensional habitat, based on model published in Austerlitz et al. (1997)	(Liu et al. 2006)	–	1D	–
* Serial founder effect model. Expansion in one-dimensional array of populations, each with the same carrying capacity.	(Deshpande et al. 2009)	–	1D	–
* Flexible computer simulation model to explore the spread of lactase persistence, dairying, other subsistence practices, and unlinked genetic markers in Europe and western Asia's geographic space	(Itan et al. 2009)	–	2D	–
* SELECTOR: Forward simulation program to study diploid individuals within a spatially explicit framework. Simulation of allele frequencies at neutral or selected loci	(Currat et al. 2010)	–	2D	–
* CISGeM: Climate informed spatial genetic models. Spatially explicit stepping-stone model divided into hexagonal cells	(Eriksson et al. 2012; Eriksson and Manica 2012)	–	2D	–

12.4 SPLATCHE: An Example of Spatially Explicit Simulation Program

The program SPLATCHE is able to translate environmental information into genetic diversity (Currat et al. 2004; Ray et al. 2010). The simulations are done in two phases: during the first forward part, the demography and dispersal of a species can be simulated from one or several origins, considering environmental information. In a second phase, the molecular genetic diversity of one or several samples drawn from the population simulated during the first phase can be generated.

12.4.1 Spatial Structure

A digital map in ASCII raster format is used for the spatial specifications of the simulations. It represents the geographical contour of the region of interest with a geographic projection adapted to the representation of surfaces (Ray 2003). Additional information, such as vegetation cover, terrain, seacoast, or rivers, may also be incorporated in the program using the same file format. The map is divided into geographic cells that are characterized by their coordinates and possibly by environmental characteristics. The dimension of the cell (usually identical for all cells) can be of arbitrary size and represents a real geographic distribution. For studying human dispersal, cells of 50 × 50 km^2 (e.g. Currat and Excoffier 2005) or 100 × 100 km^2 are usually used (e.g. Currat and Excoffier 2011) depending on the desired resolution and on the computational constraints (more cells imply more computing time). Sizes of this magnitude seem to be adequate to represent subpopulations of hunter-gatherers (Anderson and Gillam 2000; Gronenborg 1999; Cavalli-Sforza and Hewlett 1982; Hewlett et al. 1982).

The environmental characteristics of each cell (vegetation, altitude, coastal area, river, etc.) may be directly used to compute two demographic variables: the carrying capacity K and the friction rate F (Fig. 12.1). It is possible to take into account the uncertainty of these parameters by performing multiple simulations with different parameter values and by doing sensitivity analyses (Ray et al. 2008). Each cell contains one deme, which represents a sub-unit of the whole populations (or species) under study, or alternatively, two demes representing two interacting species or populations (see Box 12.1 and Fig. 12.2 for details).

Time A single simulation consists of recording the density of all demes and the number of migrants exchanged between demes during a fixed number of generations (parameter t).

Dispersal and Migration A migration probability (parameter m) for individual genes to move between (neighbouring) demes. This probability is constant in SPLATCHE but can be different during the dispersion phase and at demographic equilibrium in other programs (e.g. Eriksson et al. 2012; Deshpande et al. 2009). This basic migration scheme may be altered in different ways:

- By using friction values (parameter F), which represent the difficulty of crossing a cell depending on its specific environmental characteristics. Higher F values assigned to a given cell generally imply less migrants entering it (Ray et al. 2008). This parameter can be used, for instance, to favour movements along coastlines or rivers, or to the opposite, to use rivers, hills, deserts, or mountains as barriers to gene flow.

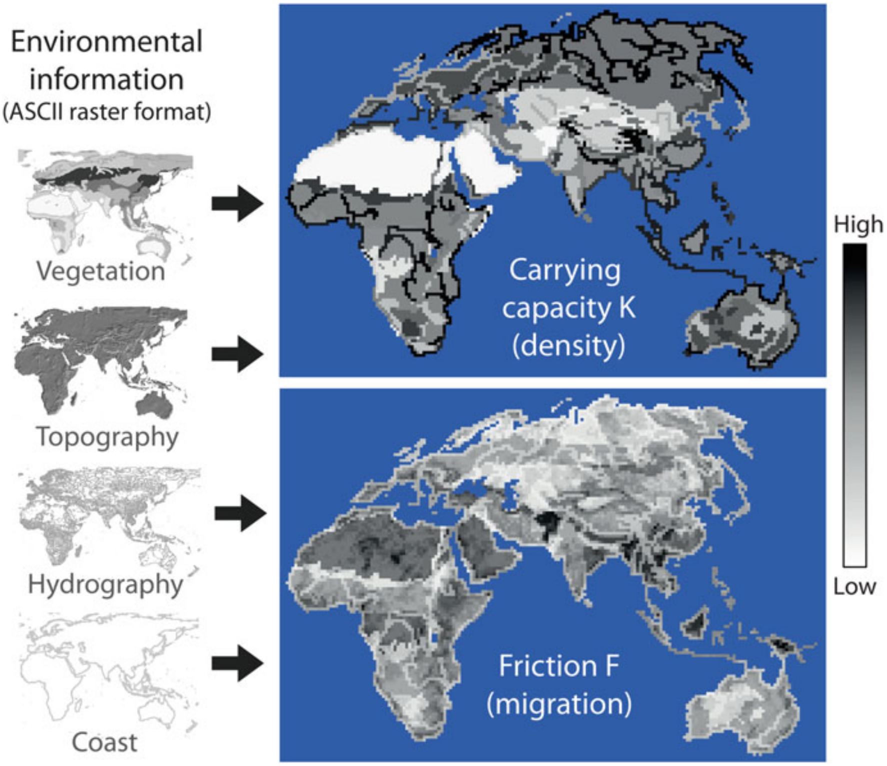


Fig. 12.1 Schematic representation of the incorporation of environmental information in SPLATCHE. The influence of various environmental factors, such as vegetation, hydrography, topography, and coastlines, is translated into two demographic parameters: carrying capacity *K* which affects population densities; Friction *F* which affects migration

- By allowing for long-distance dispersals (Ray and Excoffier 2010), occurring at a given rate (parameter λ) and at various distance (parameter d).
- By directing migration specifically into a given direction (Arenas et al. 2012), e.g. towards the South during glaciation or to the North during a post-glacial colonization.

Demographic Dynamics Deme density is usually logistically regulated, reflecting intra-deme competition for resources. The increase in density is controlled by the growth rate (parameter r) and the maximum number of individuals that may be sustained by the cell (the carrying capacity, parameter K). This last parameter may reflect the influence of the environment (e.g. different vegetation types leading to different K) or the type of culture or economy (e.g. food production *versus* hunter-gathering techniques). Heterogeneous environments can also be considered in the dispersion model (Wegmann et al. 2006).

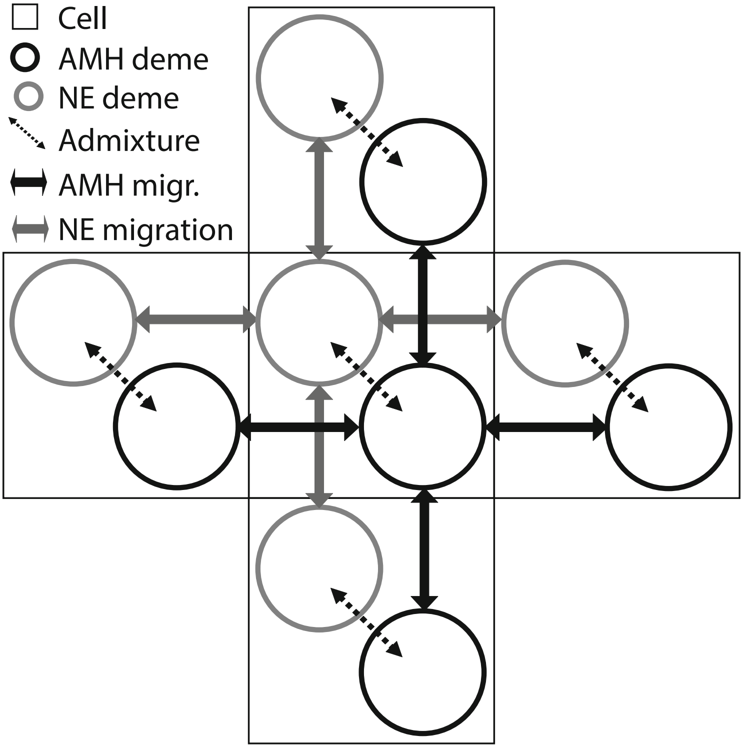


Fig. 12.2 Schematic representation of a spatially explicit modelling of two interacting populations. The whole area under study is divided into geographic cells, which each contain two demes representing two interacting populations (for example, AMH and NE). Demes of the same population belonging to neighbouring cells can exchange migrants, while admixture (gene flow) can occur between demes within the same cell

Generation of Genetic Data SPLATCHE is using a coalescent approach (Hudson 1990; Kingman 1982) to reconstruct backward in time the genealogy of a series of sampled genes. For neutral loci, the virtual genetic diversity obtained at the end of a simulation is constrained by the demography of the simulated population and possibly by a mutation and recombination model that depends on the type of simulated genetic data (e.g. allele frequencies or molecular data).

The combination of the various elements described above allows one to construct and simulate realistic scenarios of human dispersal and testing those using genetic data. For instance, the genetic consequences of a range expansion (demographic increase linked to a geographical spread), bottleneck, population contraction to refuge area(s), interactions between populations (competition and admixture), or a combination of those processes, can be investigated using this general approach.

12.5 Main Results and Discussion

The simulation of human dispersal offers a powerful tool to study the evolution of our species, and it complements other methodological approaches. Most of the early models for the simulation of human dispersal have been developed in the context of the peopling history of Europe (Rendine et al. 1986; Barbujani et al. 1995; Currat and Excoffier 2005; Arenas et al. 2013), but we focus here only on simulations of human dispersal at the global scale. The first realistic attempts (Ray et al. 2005; Liu et al. 2006) to simulate worldwide dispersal showed an excellent fit between the predictions of the models and real data, confirming that models explicitly incorporating demography and geography were powerful to make inferences on human peopling history. This has been confirmed by a study incorporating Pleistocene climatic variation (Eriksson et al. 2012), which showed that climatic change had a significant impact on human dispersal and the establishment of current genetic diversity.

12.5.1 Genes Surfing the Waves of Expansion

An important result brought by the simulation of human dispersal was to explain the mechanisms by which genetic diversity progressively decreases with distance from Africa: a series of founder effects during the range expansion of human populations (Ramachandran et al. 2005; Deshpande et al. 2009). The same mechanism has also been proposed to explain clinical genetic patterns in Europe (Barbujani et al. 1995; Currat and Excoffier 2005). The process of decreasing diversity along a colonization route was extensively described and demonstrated theoretically by a series of spatially explicit simulations performed in a 2-dimensional stepping-stone (Deshpande et al. 2009). This effect results from a phenomenon called “allele surfing” (Edmonds et al. 2004; Klopfstein et al. 2006), which describes how the frequency of a neutral allele can dramatically increase during a population expansion due to pure neutral demographic effects. Surfing is a stochastic process that only applies to a relatively small number of alleles from the source population or to new mutations appearing during the expansion, but it has important evolutionary consequences (Excoffier and Ray 2008; Petit and Excoffier 2009). Simulations have shown that the frequency of surfing varies depending on the demographic parameters of the population (Klopfstein et al. 2006). The probability of surfing increases with the growth rate (parameter *r*), while it inversely decreases with the carrying capacity (*K*) and the migration rate (*m*). This gene surfing process has been identified theoretically by spatially explicit simulations before being confirmed by empirical studies, both in yeasts and bacteria (Hallatschek et al. 2007) and by a survey of the recent human colonization of the Saguenay Lac Saint-Jean in Quebec (Moreau et al. 2011). This finding shows that neutral evolutionary processes can produce patterns identical to those expected under the action of positive selection (i.e. an allele present at very high frequency in the final population). It thus challenged the view that large differences between populations such as those observed at some loci between Africans and non-Africans were due to ongoing selection (Currat et al. 2006).

12.5.2 Hybridization During Expansion

On their road out of Africa, AMH met various archaic human forms, such as Neanderthals (NE), Denisovans (DE), and probably others (Prufer et al. 2014). Very little is known about the exact nature of the interactions between AMH and NE and virtually nothing with DE. Simulation is thus an inestimable tool to assess, in a spatially dynamic context, the effects on genetic diversity of at least two kinds of interactions: admixture and competition. The main interest of spatially explicit simulation compared to previous mathematical models is that progressive admixture in time and space can be simulated between NE and AMH, instead of instantaneous merging of two panmictic populations (Nordborg 1998; Serre et al. 2004).

Spatially explicit simulations have shown that continuous interbreeding over space and time between NE and AMH during the spread out of Africa of the latter may well explain the current patterns of genetic introgression. First, it has been shown that the absence of Neanderthal type of mitochondrial DNA in contemporary human population is compatible with an extremely low interbreeding success rate (parameter γ in Box 12.1) between AMH and Neanderthal (Currat and Excoffier 2004). It was shown (Currat and Excoffier 2011) that less than 2% of interbreeding success rate (γ) is compatible with the observed presence of 2–3% of Neanderthal DNA in the genome of contemporary non-Africans (Green et al. 2010; Reich et al. 2010; Wall et al. 2013). Such low levels of introgression could be due to only about 200–300 successful hybridization events between NE and AMH over their whole cohabitation period (at least 10 Ky in Europe, even more in the Near-East) and over all their area of overlap. New estimations using a more symmetrical hybridization model (Excoffier et al. 2014) and considering in the estimation procedure that NE introgression is higher in Asia than in Europe (Wall et al. 2013) have produced similar results with an estimate γ of less than 3% (Box 12.2). These results thus demonstrate that the reported pattern of NE introgression in extent modern humans is compatible with a strong reproductive isolation between AMH and NE. In addition, spatially explicit simulations suggest that the hybridization between AMH and NE occurred over a large geographic zone covering Western and Central Asia and probably reaching southern Siberia. The analysis was not able to precisely delineate the Neanderthal occupation zone, but it suggested that it should be as big in Asia as in Europe at the time of the spread of AMH (Currat and Excoffier 2011).

12.5.3 Limitations and Future Developments

One sensitive point of any modelling approach is the choice of the parameter values, which may be difficult for some parameters. For instance, demographic parameters for prehistoric populations (densities, growth, and migration rates) are often difficult to evaluate and the mutation rate for the studied loci might not be known precisely (Gibbons 2012). To somehow circumvent this problem, a thorough examination of the literature is necessary to establish possible intervals of parameter values. For instance, density estimates may come from ethnographic comparisons (e.g. Pennington 2001) or long-term estimation (Biraben 1979), while growth and migration rates may be derived using estimates of colonization times in different continental areas (Ray et al. 2005; Currat and Excoffier 2005, 2011, 2004). The simulation approach then allows an extensive exploration of the parameter space and the validation of plausible values.

Compared to deterministic mathematical models, computer simulations have the advantage of considering stochastic processes in the analyses, which could play an important role in the evolution of humans, as in the case of the gene surfing phenomenon described above. A benefit of the simulation approach is that models can be improved step by step by adding new elements and new information brought by scientific discoveries (Currat and Silva 2013).

Another advantage of realistic simulations of human dispersal is their ability to integrate various sources of information, such as genetics, archaeology, and environment. This feature could be very useful in the near future as new types of data are regularly produced. For instance, in the last decades, it has been increasingly possible to extract DNA from fossil remains and this technical advance has been widely used in the study of human evolution. Computer simulations should allow one to analyse the increasing number of genomic data retrieved from ancient specimen (Sanchez-Quinto et al. 2012; Skoglund et al. 2014, 2012), which should be especially powerful when combined with modern DNA. Moreover, computer simulations could be used to check if selection has favoured the introgression of some genes of archaic origin (Caspermeyer 2014; Ding et al. 2014). Finally, computer simulation could be used to understand the relation between Denisovan, Southeast Asians, and Oceanians (Reich et al. 2011), but the lack of spatial information about the exact ancestral Denisovan range makes it a challenging task.

12.6 Conclusion

Computer simulations are a powerful tool to study the effects of population dynamics on the genetic diversity of populations. It has been shown that properly considering the spatial dynamics of populations can drastically change the interpretation of empirical data. For instance, a spatial expansion does not show the same typical bell-shaped mitochondrial mismatch distribution than a pure demographic expansion in a panmictic population, but rather a multimodal mismatch distribution when migration rates between demes are small to moderate (Ray et al. 2003; Excoffier 2004).

All in all, spatially explicit simulations support a relatively simple scenario of global human dispersal out of Africa with very rare interbreeding events with archaic humans over a large geographic range. They revealed that observed low levels of Neanderthal ancestry in Eurasians are compatible with a very low rate of interbreeding ($<3\%$) and that those rare and distinct admixture events occurred also after the split of Europeans and Asian over a wide European and Asiatic range, well beyond the Middle East. A spatially explicit model of population expansion with continuous but limited interbreeding explains most observed patterns of human genomic diversity, such as: (1) a recent single origin (Stewart and Stringer 2012), (2) decreasing genetic diversity from Africa (Prugnolle et al. 2005); (3) limited and relatively uniform Neanderthal introgression in Eurasia (Green et al. 2010), larger in East Asia than in Europe (Wall et al. 2013), (4) introgression asymmetry between NE and AMH (Green et al. 2010), (5) lack of mitochondrial introgression (Reich et al. 2010; Serre et al. 2004), (6) introgression in area where Neanderthal never existed (Green et al. 2010), (7) more than one hybridization event with archaic humans (Wall et al. 2013). Spatially explicit simulation of human dispersal thus provides a simple but powerful framework for the interpretation of genomic data, with considerable room for improvements and extensions.

Box 12.1 Simulation of Interactions Between Two Populations

The program SPLATCHE offers the possibility to study the interactions between two populations (e.g. hunter-gatherers and Neolithic farmers) or two species (e.g. NE and AMH) in a spatially explicit framework (Ray et al. 2010). It simulates two demes per geographic cell, each of them representing one population (Fig. 12.2), and the simulated world can thus be seen as two superimposed layers of demes. SPLATCHE considers two kinds of interaction between interacting populations or species: competition and admixture (Fig. 12.2).

Competition is simulated using a classical Lotka–Volterra model (Lotka 1932), which is an extension of the logistic regulation model. The density of one species is directly constrained by the density of the other one, assuming that both are in competition for local resources (e.g. habitat, food). Competition coefficients (α) are used to reflect the intensity of this kind of interactions and can be different in the two populations, i.e. due to a competitive advantage of one over the other. The competition coefficients α can be fixed to a specific value, such as 1, which means that an individual of the rival population exerts as much competitive pressure as an individual belonging to the same population. Alternatively, α can be density dependent (Currat and Excoffier 2005), i.e. computed as $\alpha_{ij} = N_j/(N_i + N_j)$, where α_{ij} represents the effect of competition of an individual of population j on an individual of population i and N_i

(continued)

Box 12.1 (continued)

and N_j are the densities of both populations in the cell. In that case, the strength of competition between both populations evolves over time and the most numerous one has a competitive edge over the less numerous one. Under this model, if the carrying capacities (K) between the two populations are sufficiently different, as it is assumed for AMH over NE (Currat and Excoffier 2011, 2004), then the one with the lower K will eventually goes extinct due to competition.

Admixture is simulated by local gene flow between the two demes belonging to the same cell and it can be regulated by an interbreeding success rate (parameter ɣ). If ɣ is equal to 0, then there is no gene flow between the two populations. If ɣ is equal to 1, there is random mating between them. Lower values of ɣ imply the existence of barriers to gene flow between the two species, which can be either pre-zygotic (e.g. cultural avoidance, disassortative mating) or post-zygotic due to lower hybrid fitness or a combination of those various factors.

In previous studies, we implemented a model of density-dependent gene flow between AMH and NE (Currat and Excoffier 2011, 2004). A new admixture model has been implemented in SPLATCHE, which is fully symmetrical when both species are at demographic equilibrium and which is more accurate for the description of interspecific hybridization than the previous model (Excoffier et al. 2014). In this new model, each N_i' newborn individuals in a population i have at least one parent belonging to population i. Then, assuming random mating the probability that the second parent originated from population j is simply computed as $N_j/(N_i + N_j)$, where N_i and N_j are the densities of both populations in the previous generation. Thus, the expected number of gene flow events (introgressions) from population j to i at each generation is defined as:

$$S_{ji} = \gamma N_i' \frac{N_j}{2(N_i + N_j)}$$

This new admixture model gives results qualitatively and quantitatively very similar to the previous one (Excoffier et al. 2014).

Box 12.2 Estimation of Hybridization Between Neanderthals and AMH

To investigate hybridization between Neanderthals and AMH, we explored a series of scenarios of human dispersal out of Africa into Eurasia, with various demographic parameters, hybridization and competition zones and varying intensities of admixture. We redirect the readers to the original article for exact details on those alternative scenarios (Currat and Excoffier 2011). In all examined scenarios, Neanderthals is assumed to be at demographic equilibrium in its entire range at the beginning of the simulation, while AMH is expanding demographically and spatially (Fig. 12.3). Various sizes of Neanderthal occupation zone where hybridization with AMH occurred were tested, ranging from the Middle East only to a wider area extending to southern Siberia (Fig. 12.3). During the AMH expansion, admixture can occur, and Neanderthals disappear due to competition with AMH. At the end of a simulation, the proportion of Neanderthal ancestry is measured in modern human genomes in Europe and in East Asia and compared to the reported levels of 2–3% (Reich et al. 2010). For each of the 13 envisioned scenarios, we performed 10,000 coalescent simulations and we computed the proportion of those simulations that were compatible with the observation. In our new study, a simulation was declared compatible if it resulted in 2–3% of Neanderthal ancestry in both Europe and East Asia and if Neanderthal introgression was slightly larger in East Asia than in Europe (as shown in Wall et al. 2013). Figure 12.4 and Table 12.2 show the interbreeding rates obtained with this new series of simulations. They confirm previous results (Currat and Excoffier 2011; Excoffier et al. 2014), as very few successful hybridization events ($\gamma < 3\%$) over a wide Eurasian range are sufficient to result in 2–3% of introgression in non-Africans. Such a low hybridization rate is sufficient to explain current Neanderthal introgression because the few Neanderthal genes that are incorporated continuously at the wavefront of the AMH expansion tend to be amplified by the surfing phenomenon (Currat et al. 2008). Indeed, a few introgression events occurring in an invading deme on the wavefront can result in many more introgressed copies, because these introgressions usually occur when the invading deme is still growing and has not reached its carrying capacity. Second, AMH pioneers are recruited at the front of the expansion and consequently have a higher probability to propagate their genes (including recently introgressed Neanderthal genes) further away in the expanding population. Rare but continuous interbreeding events during the expansion of AMH over a large Neanderthal Eurasian range are thus a simple and efficient model to explain patterns of Neanderthal ancestry in current genomes.

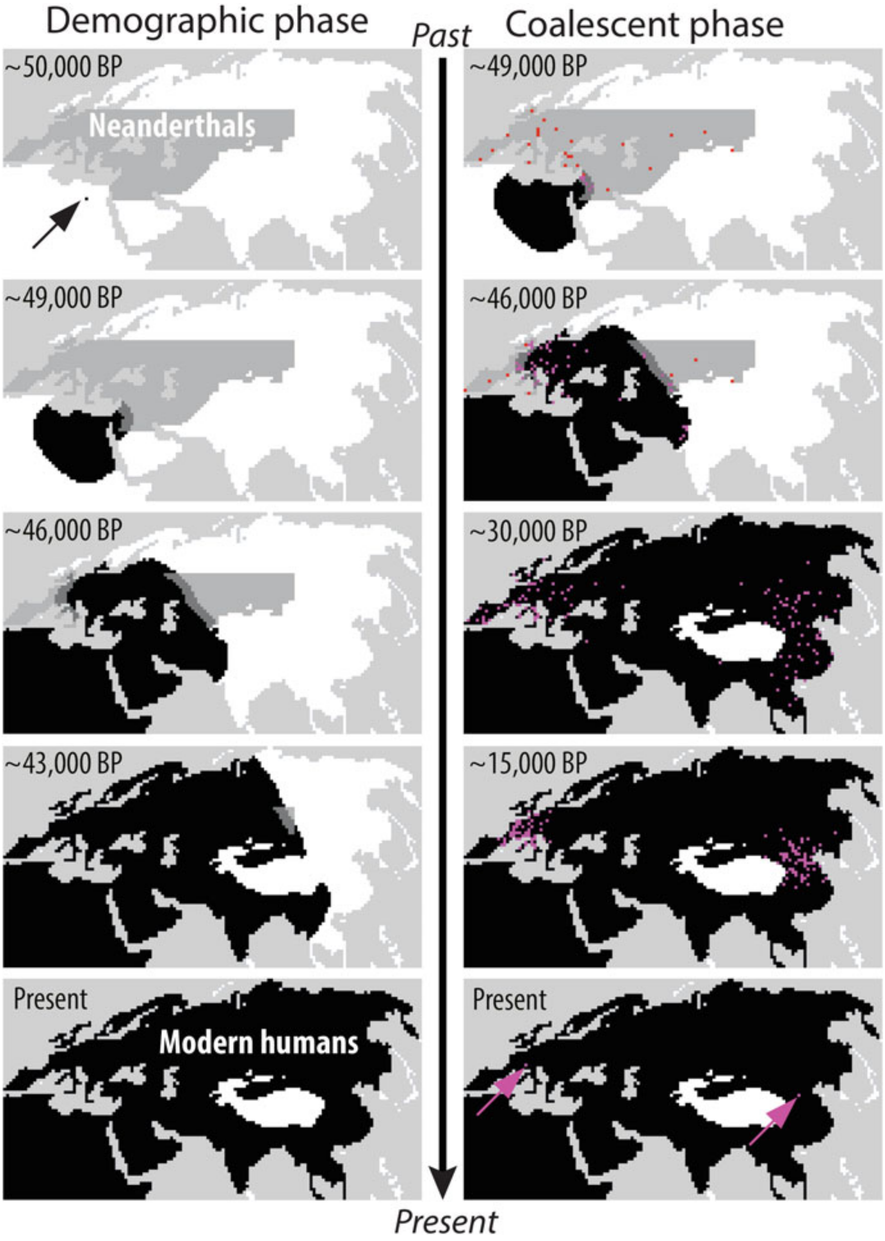


Fig. 12.3 Example of the simulation of the dispersal of AMH out of Africa and the progressive disappearance of NE that inhabited an extended area in Eurasia. Black arrow represents the origin of AMH and pink arrows the current sampled locations. White represents empty cells, medium grey cells are occupied by NE only, black cells are occupied by AMH only, and dark grey represents the zone where NE and AMH coexist. Left panel shows the demographic forward phase and right panel shows the coalescent backward phase. Pink and red dots on the right panel represent at any time the location of lineages ancestral to the AMH sampled genes in AMH and NE demes, respectively. The parameters used for the simulations are those of scenario A in Table 12.2

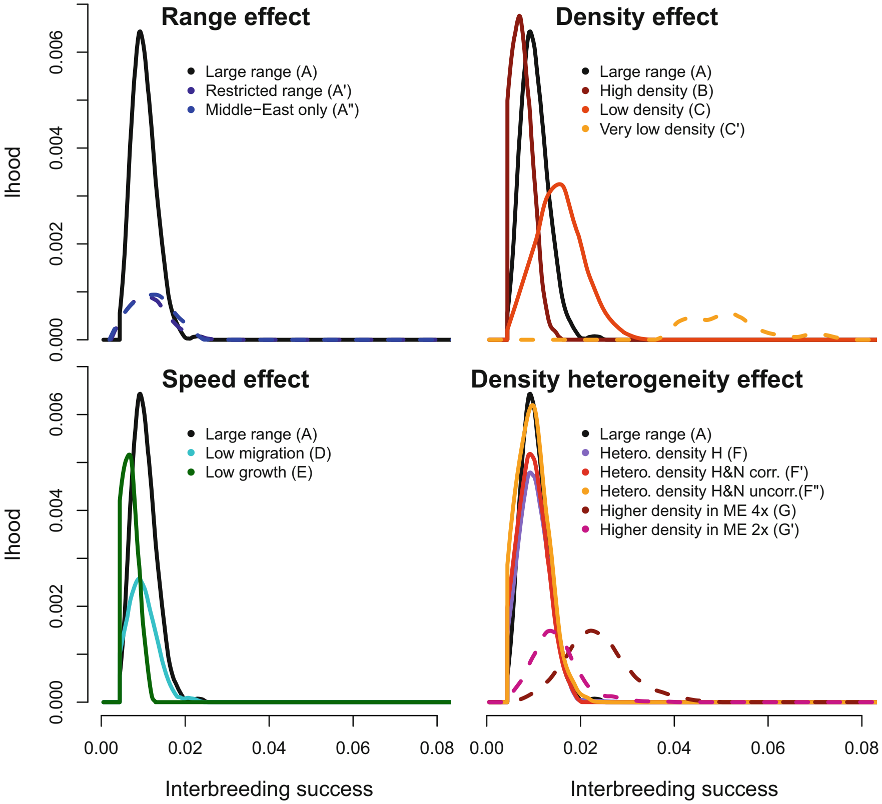


Fig. 12.4 Distribution of the proportion of simulations (among 10,000) resulting in Neanderthal introgression levels higher in the Chinese sample but still compatible with observations (1.9–3.1% in both French and Chinese samples). Each likelihood curve corresponds to a different demographic scenario described in Table 12.2. Results were obtained by assuming a deme area of 100 × 100 km^2. Solid lines correspond to scenarios that are equally likely (within two AIC units from the scenario with the highest likelihood), whereas scenarios shown with a dotted line have an associated AIC more than two units larger and thus cannot be considered as equally well supported by the data (details of the estimation procedure may be found in Currat and Excoffier 2011)

Table 12.2 Demographic parameters, interbreeding rate estimates, and relative probabilities of the simulated scenarios

Models	K_N[a]	K_H[b]	r[c]	m_N[d]	m_H[e]	Colonization time[f]	Estimated interbreeding success[g]		Model A relative probability[h,i]
A. Large neanderthal range	200	800	0.8	0.1	0.2	220	0.0093	[0.0049–0.0159]	–
A′. Restricted neanderthal range	200	800	0.8	0.1	0.2	220	0.0105	[0.0031–0.0213]	7.05
A″. Hybridization in middle East only	200	800	0.8	0.1	0.2	180*i*	0.0126	[0.0027–0.0224]	6.84
B. Large *K*	400	1600	0.8	0.1	0.2	220	0.0070	[0.0044–0.0121]	0.95
C. Small *K*	100	400	0.8	0.1	0.2	220	0.0155	[0.0054–0.0259]	1.98
C′. Very small *K*	25	100	0.8	0.1	0.2	240	0.0519	[0.0380–0.0731]	11.46
D. Small *m*	200	800	0.8	0.05	0.1	290	0.0091	[0.0044–0.0163]	2.50
E. Small *r*	200	800	0.4	0.1	0.2	300	0.0066	[0.0044–0.0109]	1.25
F. Variable K_H	200	200–1600	0.8	0.1	0.2	220	0.0093	[0.0044–0.0169]	1.34
F′. Variable K_H and K_N (correlated)	50–400	200–1600	0.8	0.1	0.2	220	0.0093	[0.0044–0.0163]	1.24
F″. Variable K_H and K_N (uncorrelated)	50–400	200–1600	0.8	0.1	0.2	220	0.0098	[0.0044–0.0165]	1.04
G. K 4× higher in ME	200 (50)	800 (200)	0.8	0.1	0.2	220	0.0222	[0.0105–0.0369]	4.32

G′. K 2× higher in ME	200 (100)	800 (200)	0.8	0.1	0.2	220	0.0135	[0.0056–0.0229]	4.32

[a] Neanderthal carrying capacity

[b] Human carrying capacity

[c] Intrinsic rate of growth

[d] Migration rate between Neanderthal demes

[e] Migration rate between human demes

[f] Approximate time (in generations) for the colonization of Europe from the Middle East estimated from the simulations. This statistic is determined by the growth and migration rates and varies also slightly with the hybridization rate

[g] Maximum likelihood estimates of interbreeding success between humans and Neanderthals are reported with limits of a 95% CI within brackets

[h] Probability of scenario A relative to the other scenarios computed from weighted AIC's (see Material and Methods)

[i] Hybridization occurs for about 80 generations in the Middle East

Acknowledgments This research was supported by a grant of the Center for Advanced Modelling Science (CADMOS) to MC and by Swiss NSF grants 31003A-143393 and 310030_188883 to LE and 31003A_182577 to MC. CSQ acknowledges support from CONICYT-Becas Chile and the Institute of Genetics and Genomics in Geneva (IGE3). All simulations were performed in the HPC cluster baobab.unige.ch.

References

Ackland GJ et al (2007) Cultural hitchhiking on the wave of advance of beneficial technologies. Proc Natl Acad Sci U S A 104(21):8714–8719

Anderson DG, Gillam C (2000) Paleoindian colonization of the Americas: implications from an examination of physiography, demography, and artifact distribution. Am Antiq 65(1):43–66

Arenas M et al (2012) Consequences of range contractions and range shifts on molecular diversity. Mol Biol Evol 29(1):207–218

Arenas M et al (2013) Influence of admixture and paleolithic range contractions on current European diversity gradients. Mol Biol Evol 30(1):57–61

Austerlitz F et al (1997) Evolution of coalescence times, genetic diversity and structure during colonization. Theor Popul Biol 51(2):148–164

Balloux F (2001) EASYPOP (version 1.7): a computer program for population genetics simulations. J Hered 92(3):301–302

Banks WE et al (2008) Human ecological niches and ranges during the LGM in Europe derived from an application of eco-cultural niche modeling. J Archaeol Sci 35(2):481–491

Barbujani G, Sokal RR, Oden NL (1995) Indo-European origins: a computer-simulation test of five hypotheses. Am J Phys Anthropol 96(2):109–132

Barton NH et al (2013a) Genetic hitchhiking in spatially extended populations. Theor Popul Biol 87:75–89

Barton NH, Etheridge AM, Veber A (2013b) Modelling evolution in a spatial continuum. J Stat Mech 2013:P01002

Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian computation in population genetics. Genetics 162(4):2025–2035

Bellwood P (2001) Early agriculturalist population. Annu Rev Anthropol 30:181–207

Biraben JN (1979) Essay on the evolution of numbers of mankind. Population 34(1):13–25

Brown JM, Savidge K, McTavish EJ (2011) DIM SUM: demography and individual migration simulated using a Markov chain. Mol Ecol Resour 11(2):358–363

Caspermeyer J (2014) Sunlight adaptation region of neanderthal genome found in up to 65% of modern East Asian populations. Mol Biol Evol 31(3):763

Cavalli-Sforza LL, Hewlett B (1982) Exploration and mating range in African Pygmies. Ann Hum Genet 46(Pt 3):257–270

Chadeau-Hyam M et al (2008) Fregene: simulation of realistic sequence-level data in populations and ascertained samples. BMC Bioinf 9:364

Currat M, Excoffier L (2004) Modern humans did not admix with Neanderthals during their range expansion into Europe. PLoS Biol 2(12):2264–2274

Currat M, Excoffier L (2005) The effect of the Neolithic expansion on European molecular diversity. Proc R Soc B 272(1564):679–688

Currat M, Excoffier L (2011) Strong reproductive isolation between humans and Neanderthals inferred from observed patterns of introgression. Proc Natl Acad Sci U S A 108 (37):15129–15134

Currat M, Silva NM (2013) Investigating European genetic history through computer simulations. Hum Hered 76(3-4):142–153

Currat M, Ray N, Excoffier L (2004) SPLATCHE: a program to simulate genetic diversity taking into account environmental heterogeneity. Mol Ecol Notes 4(1):139–142

Currat M et al (2006) Comment on “Ongoing adaptive evolution of ASPM, a brain size determinant in homo sapiens” and “microcephalin, a gene regulating brain size, continues to evolve adaptively in humans”. Science 313:5784

Currat M et al (2008) The hidden side of invasions: massive introgression by local genes. Evolution 62(8):1908–1920

Currat M, Poloni ES, Sanchez-Mazas A (2010) Human genetic differentiation across the Strait of Gibraltar. BMC Evol Biol 10:237

DeGiorgio M, Degnan JH, Rosenberg NA (2011) Coalescence-time distributions in a serial founder model of human evolutionary history. Genetics 189(2):579–593

Dellicour S, Hardy OJ, Mardulyn P (2014) PhyloGeoSim 1.0. A program to simulate the evolution of DNA sequences under a spatially explicit model of coalescence. Mol Biol Evol 31 (12):3359–3372

Deshpande O et al (2009) A serial founder effect model for human settlement out of Africa. Proc Biol Sci 276(1655):291–300

Di D, Sanchez-Mazas A (2011) Challenging views on the peopling history of East Asia: the story according to HLA markers. Am J Phys Anthropol 145(1):81–96

Ding Q et al (2014) Neanderthal introgression at chromosome 3p21.31 was under positive natural selection in East Asians. Mol Biol Evol 31(3):683–695

Edmonds CA, Lillie AS, Cavalli-Sforza LL (2004) Mutations arising in the wave front of an expanding population. Proc Natl Acad Sci U S A 101(4):975–979

Eriksson A, Manica A (2012) Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. Proc Natl Acad Sci U S A 109(35):13956–13960

Eriksson A et al (2012) Late Pleistocene climate change and the global expansion of anatomically modern humans. Proc Natl Acad Sci U S A 109(40):16089–16094

Eswaran V (2002) A diffusion wave out of Africa. Curr Anthropol 43:749–764

Eswaran V, Harpending H, Rogers AR (2005) Genomics refutes an exclusively African origin of humans. J Hum Evol 49(1):1–18

Ewing G, Hermisson J (2010) MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics 26(16):2064–2065

Excoffier L (2004) Patterns of DNA sequence diversity and genetic structure after a range expansion: lessons from the infinite-island model. Mol Ecol 13(4):853–864

Excoffier L, Foll M (2011) Fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. Bioinformatics 27(9):1332–1334

Excoffier L, Ray N (2008) Surfing during population expansions promotes genetic revolutions and structuration. Trends Ecol Evol 23(7):347–351

Excoffier L, Novembre J, Schneider S (2000) SIMCOAL: a general coalescent program for the simulation of molecular data in interconnected populations with arbitrary demography. J Hered 91:506–510

Excoffier L, Quilodran CS, Currat M (2014) Models of hybridization during range expansions and their application to recent human evolution. In: Derevianko AP, Shunkov MV (eds) Cultural developments in the eurasian paleolithic and the origin of anatomically modern humans. Publishing Department of the Institute of Archaeology and Ethnography SB RAS, Novosibirsk, pp 122–137

Fisher RA (1937) The wave of advance of advantageous genes. Ann Eugenics 7:355–369

Fix AG (1996) Gene frequency clines in Europe: demic diffusion or natural selection? J Roy Anthropol Inst 2:625–643

Fix AG (1997) Gene frequency clines produced by kin-structured founder effects. Hum Biol 69 (5):663–673

Fort J, Mendez V (1999) Reaction-diffusion waves of advance in the transition to agricultural economics. Phys Rev E 60(5 Pt B):5894–5901

Fort J, Pujol T (2008) Progress in front propagation research. Rep Prog Phys 71(8):086001

Fort J, Pujol T, Cavalli-Sforza LL (2004) Palaeolithic populations and waves of advance (human range expansions). Camb Archaeol J 14(1):53–61

Gibbons A (2012) Human evolution. Turning back the clock: slowing the pace of prehistory. Science 338(6104):189–191

Green RE et al (2010) A draft sequence of the neandertal genome. Science 328(5979):710–722

Gronenborg D (1999) A variation on a basic theme: the transition to farming in southern central Europe. J World Prehist 13(2):123–210

Guillaume F, Rougemont J (2006) Nemo: an evolutionary and population genetics programming framework. Bioinformatics 22(20):2556–2557

Hallatschek O et al (2007) Genetic drift at expanding frontiers promotes gene segregation. Proc Natl Acad Sci U S A 104(50):19926–19930

Henn BM, Cavalli-Sforza LL, Feldman MW (2012) The great human expansion. Proc Natl Acad Sci U S A 109(44):17758–17764

Hewitt GM (2000) The genetic legacy of the quartenary ice ages. Nature 405:907–913

Hewlett B, Van de Koppel JMH, Cavalli-Sforza L (1982) Exploration ranges of Aka pygmies of the Central African Republic. Man 17:418–430

Hudson RR (1990) Gene genealogies and the coalescent process. In: Futuyma D, Antonovics J (eds) Oxford surveys in evolutionary biology, vol 7. Oxford University Press, Oxford, p 300

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

Itan Y et al (2009) The origins of lactase persistence in Europe. PLoS Comput Biol 5(8):e1000491

Kimura M (1953) Stepping-stone model of population. Genetics 3:62–63

Kingman JFC (1982) The coalescent. Stoch Process Appl 13:235–248

Klopfstein S, Currat M, Excoffier L (2006) The fate of mutations surfing on the wave of a range expansion. Mol Biol Evol 23(3):482–490

Lahr MM, Foley RA (1998) Toward a theory of modern human origins: geography, demography, and diversity in recent human evolution. Yearb Phys Anthropol 41:137–176

Landguth EL, Cushman SA (2010) CDPOP: a spatially explicit cost distance population genetics program. Mol Ecol Resour 10(1):156–161

Laval G, Excoffier L (2004) SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. Bioinformatics 20 (15):1485–2487

Lavrentovich MO, Korolev KS, Nelson DR (2013) Radial Domany-Kinzel models with mutation and selection. Phys Rev E 87(1):012103

Liu H et al (2006) A geographically explicit genetic model of worldwide human-settlement history. Am J Hum Genet 79(2):230–237

Lotka AJ (1932) The growth of mixed populations: two species competing for a common food supply. J Wash Acad Sci 22:461–469

Macaulay V et al (2005) Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. Science 308(5724):1034–1036

Martino LA et al (2007) Fisher equation for anisotropic diffusion: simulating South American human dispersals. Phys Rev E 76(3):031923

Meirmans PG (2011) MARLIN, software to create, run, and analyse spatially realistic simulations. Mol Ecol Resour 11(1):146–150

Moreau C et al (2011) Deep human genealogies reveal a selective advantage to be on an expanding wave front. Science 334(6059):1148–1150

Morton NE (1977) Isolation by distance in human populations. Ann Hum Genet 40(3):361–365

Morton NE (1982) Estimation of demographic parameters from isolation by distance. Hum Hered 32:37–41

Neuenschwander S et al (2008) quantinemo: an individual-based program to simulate quantitative traits with explicit genetic architecture in a dynamic metapopulation. Bioinformatics 24 (13):1552–1553

Nordborg M (1998) On the probability of Neanderthal ancestry. Am J Hum Genet 63(4):1237–1240

Novembre J et al (2008) Genes mirror geography within Europe. Nature 456(7218):98–101

Peischl S et al (2013) On the accumulation of deleterious mutations during range expansions. Mol Ecol 22(24):5972–5982

Pennington R (2001) Hunter-gatherer demography. In: Panter-Brick C, Layton RH, Rowley-Conwy P (eds) Hunter-gatherers: an interdisciplinary perspective. Cambridge University Press, Cambridge, pp 170–204

Petit RJ, Excoffier L (2009) Gene flow and species delimitation. Trends Ecol Evol 24(7):386–393

Powell A, Shennan S, Thomas MG (2009) Late pleistocene demography and the appearance of modern human behavior. Science 324(5932):1298–1301

Prufer K et al (2014) The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505(7481):43–49

Prugnolle F, Manica A, Balloux F (2005) Geography predicts neutral genetic diversity of human populations. Curr Biol 15(5):R159–R160

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

Rasmussen M et al (2011) An aboriginal Australian genome reveals separate human dispersals into Asia. Science 333(6052):94–98

Rasteiro R, Chikhi L (2013) Female and male perspectives on the neolithic transition in Europe: clues from ancient and modern genetic data. PLoS One 8(4):e0060944

Rasteiro R et al (2012) Investigating sex-biased migration during the Neolithic transition in Europe, using an explicit spatial simulation framework. Proc R Soc B 279(1737):2409–2416

Ray N (2003) Modélisation de la démographie des populations humaines préhistoriques à l'aide de données environnementales et génétiques. Université de Genève, Genève

Ray N, Adams JM (2001) A GIS-based vegetation map of the world at the last glacial maximum (25,000-15,000 BP). Internet Archaeol 11:319

Ray N, Excoffier L (2010) A first step towards inferring levels of long-distance dispersal during past expansions. Mol Ecol Resour 10(5):902–914

Ray N, Currat M, Excoffier L (2003) Intra-deme molecular diversity in spatially expanding populations. Mol Biol Evol 20(1):76–86

Ray N et al (2005) Recovering the geographic origin of early modern humans by realistic and spatially explicit simulations. Genome Res 15(8):1161–1167

Ray N, Currat M, Excoffier L (2008) Incorporating environmental heterogeneity in spatially-explicit simulations of human genetic diversity. In: Matsumura S, Forster P, Renfrew C (eds) Simulations, genetics and human prehistory. McDonald Institute for Archaeological Research, Cambridge, pp 103–117

Ray N et al (2010) SPLATCHE2: a spatially-explicit simulation framework for complex demography, genetic admixture and recombination. Bioinformatics 26(23):2993–2994

Rebaudo F et al (2013) SimAdapt: an individual-based genetic model for simulating landscape management impacts on populations. Methods Ecol Evol 4(6):595–600

Reich D et al (2010) Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468(7327):1053–1060

Reich D et al (2011) Denisova admixture and the first modern human dispersals into southeast Asia and Oceania. Am J Hum Genet 89(4):516–528

Rendine S, Piazza A, Cavalli-Sforza L (1986) Simulation and separation by principal components of multiple demic expansions in Europe. Am Nat 128(5):681–706

Sanchez-Quinto F et al (2012) Genomic affinities of two 7,000-year-old Iberian hunter-gatherers. Curr Biol 22(16):1494–1499

Serre D et al (2004) No evidence of neandertal mtDNA contribution to early modern humans. PLoS Biol 2(3):E57

Sgaramella-Zonta L, Cavalli-Sforza L (1973) A methode for the detection of a demic cline. In: Morton NE (ed) Genetic structure of population. University of Hawaii Press, Honolulu

Skoglund P et al (2012) Origins and genetic legacy of Neolithic farmers and Hunter-gatherers in Europe. Science 336:466–469

Skoglund P et al (2014) Genomic diversity and admixture differs for stone-age scandinavian foragers and farmers. Science 344(6185):747–750

Sokal RR (1991) Ancient movement patterns determine modern genetic variances in Europe. Hum Biol 63(5):589–606

Steele J (2009) Human dispersals: mathematical models and the archaeological record. Hum Biol 81(2-3):121–140

Steele J, Adams JM, Sluckin T (1998) Modeling Paleoindian dispersals. World Archaeol 30:286–305

Stewart JR, Stringer CB (2012) Human evolution out of Africa: the role of refugia and climate change. Science 335(6074):1317–1321

Stringer C (2014) Why we are not all multiregionalists now. Trends Ecol Evol 29(5):248–251

Sukumaran J, Holder MT (2011) Ginkgo: spatially-explicit simulator of complex phylogeographic histories. Mol Ecol Resour 11(2):364–369

Tattersall I (2009) Human origins: out of Africa. Proc Natl Acad Sci U S A 106(38):16018–16021

Veeramah KR, Hammer MF (2014) The impact of whole-genome sequencing on the reconstruction of human population history. Nat Rev Genet 15(3):149–162

Wall JD et al (2013) Higher levels of neanderthal ancestry in East Asians than in Europeans. Genetics 194(1):199

Wegmann D, Currat M, Excoffier L (2006) Molecular diversity after a range expansion in heterogeneous environments. Genetics 174(4):2009–2020

Wegmann D et al (2010) ABCtoolbox: a versatile toolkit for approximate Bayesian computations. BMC Bioinf 11:116

Wright S (1943) Isolation by distance. Genetics 28:114–138

Zvelebil M (2001) The agricultural transition and the origins of Neolithic society in Europe. Documenta Praehistorica XXVlll. Neolithic Stud 8:1–26

Index

N. Saitou (ed.), *Evolution of the Human Genome II*, Evolutionary Studies,
https://doi.org/10.1007/978-4-431-56904-6