ANALYSIS OF HUMAN GENETIC VARIATION IN CANDIDATE GENES UNDER POSITIVE SELECTION ON THE HUMAN LINEAGE

Andrés Moreno Estrada

DOCTORAL THESIS UPF / 2009

THESIS DIRECTOR

Dra. Elena Bosch Fusté

Department of Experimental and Health Sciences

UNIVERSITAT POMPEU FABRA

Front Cover Artwork: World Map downloaded from TheLivingEarth.com © 2008;
Human-Chimp Image adapted from Pääbo S., *Science* 291:1219–1220, © 2001.

Por seguirnos tan de cerca estando tan lejos

A André Michel Aubry

In Memoriam

DARWIN200

“When we can fell assured that all the individuals of the same species, and all the closely allied species of most genera, have within a not very remote period descended from one parent, and have migrated from some one birthplace… then… we shall surely be enabled to trace in an admirable manner the former migrations of the inhabitants of the whole world.”

Charles Darwin, *On the Origin of Species by Means of Natural Selection* (1859), pp. 486–487

Acknowledgements

This work could not have been possible without the valuable support of many people from both academic and personal areas. Rather than consisting of an individual achievement, I believe that one can devote so much time and effort to research thanks to the whole team that is behind every step one makes, so that this final reward is theirs too.

At the first line of the team is, of course, my family. My parents, Arturo & Cristina, and my sister and niece, Sofía & Alina, have always been there for me, supporting me and respecting my crazy isolation and obsession when following an idea. My father has always said that he wants me to go farther than him, and despite that is impossible, those words have taught me to believe that I can reach the top of the mountain, as he says. My mother has always been one step ahead and has guided me so much, understanding so well every stage of my life. They both gave the opportunity of choosing what I wanted to do and encouraged me to do whatever it takes to achieve my goals. Now that we are far away I know that they have enjoyed and suffered every step as much as I have. Thanks for supporting me and loving me so much. Thanks for making me learn from new experiences. Thanks for your example of how to come across equally well anywhere, from indigenous villages to big cities overseas. Thanks for showing me the world as it is. Thanks to all the family for caring about us all these years. I am especially grateful to my uncle Enrique Estrada Faudon who has largely influenced my interest in science. Thank you so much to all of you for sharing this moment.

I would also like to express my sincerest gratitude to my thesis director, Elena Bosch, who played a major role in this work. Thanks for caring about every single issue both professionally and personally to assure throughout these years a successful ending of my PhD. Thanks for teaching me so many concepts with your expertise in this field. Thanks for showing me how to do really good research. Thanks for understanding my meticulous way of doing things but also for pushing me when necessary, like when writing this thesis! Thanks for all these years of teamwork.

Academic life is made of opportunities, so I would like to thank Jaume Bertranpetit for the opportunity of undertaking this PhD in such a leading group of scientists, from whom I learned much of what is written in this thesis. Special thanks to Francesc Calafell, David Comas, Carles Lalueza, Arcadi Navarro, and Hafid Laayouni. Thank you Carles for being part of the evaluating committee, your opinion is very valuable to me.

I also feel especially grateful to Mark Stoneking for guidance during my search of PhD programs and for giving me the opportunity of carrying out part of this work in his group at the Max Planck Institute for Evolutionary Anthropology in Leipzig, Germany. Thanks for honoring us by presiding the examining committee of my dissertation defense. Special thanks also go to Kun Tang for helping me applying his method to my data and for showing me that crazy schedules are not so crazy. Also from the MPI-EVA I want to thank David López Herráez, David Hughes, Sean Myles, Frederick Delfin, and Guido Valverde for help during my stays in Leipzig.

I could not mention the many colleagues that have also contributed to this work without making the mistake of not including them all. Special thanks to my closest mates at work, Martin Sikora and Johannes Engelken for valuable help and unforgettable scientific chats, as well as clandestine matches in the cubicle! I really appreciate Martin's support in many critical moments of this work, not to mention his friendship. I also thank present and former members of the Evolutionary Biology Unit, especially Chiara Batini, Michelle Gardner, Kristin Kristjánsdóttir, Olga Fernando, Isabel Mendizábal, Lourdes Sampietro, Gemma Berniell, Marta Soldevila, Aida Andrés, and Oscar Lao, as well as friends and colleagues from the PRBB, especially Citlali Vázquez, Eidi Alvarado, and Leonardo Mina for sharing many special moments throughout our PhDs. Life in Barcelona was more bearable thanks to my friends Francisco Franco, Blanca Chávez, Loli Miguel, Marta Navarro, Uri Carbonell, Fernanda Álvarez, María Castellano, and Víctor Sánchez, thank you so much to all of you. Best friends living abroad are also part of this achievement, especially Hugo Guerrero and Pablo Carrillo with whom I have always shared the vocation to be scientists, so now we all can celebrate to have become doctors!

I want to dedicate this work to one of the persons that have influenced my life the most, our godfather Andrés Aubry. He is unfortunately not with us anymore but I am sure that he would have been very proud today because, despite the great geographical distance between us, he was always aware of our main steps and even of the implications of our present and future projects. I am sure that his legacy will bear fruit in the years ahead not only in our personal projects but also in the community to which he devoted most of his life.

To finish, I would like to specially thank Karla Sandoval for her unconditional love and support, without which I would not have finished this thesis. Thanks for having such unique supportive attitude. We work in a field where human beings are the main subject of study, but I have realized that very few people really worry about other humans as much as you do. I admire you for this and I am surely not the only one, many people that have crossed your path during these years have offered their friendship, which gives proof of that. Thanks for teaching me how to see the world from a point of view that I would have never had without you. Thanks for being there when I was looking for your supervisor in your former institute, to which I am sure you will go back stronger and well prepared to lead your own projects. Thanks for being the most encouraging wife. Thanks for your passion for life. Thanks for sharing many ambitious projects with me and for saying yes when we decide to go for them even if we both ignore what it takes to accomplish them, but don’t worry, we will.

Andrés Moreno, February 2009

This work has been carried out while I was supported by a fellowship from the Mexican Council for Science and Technology (CONACYT) during the period between October 2003 and September 2008 (grant 179339). I also was awarded with a travel grant from the Boehringer Ingelheim Foundation to undertake part of this work at the Max Planck Institute for Evolutionary Anthropology in Leipzig, Germany (four different months during 2007 and 2008). I must also acknowledge the financial assistance of the Universitat Pompeu Fabra for presenting this work in several international congresses and for supporting me with an additional short-term grant while finishing this thesis.

Abstract

Natural selection has played an important role in shaping human genetic variation, thus, finding variants that have been targeted by positive selection can provide insights about which genes influence human phenotypic variability. In this work we conduct a genome-wide survey of protein-coding genes comparing humans, chimpanzees, and closely related species in order to detect the fraction of genes undergoing positive selection on the human lineage, and further investigate intraspecific variation in a subset of candidate genes in the search of recent selective events in worldwide human populations. Our results suggest that most of the genes implicated in selective events during early human evolution differ from those involved in recent human adaptations, implying distinct selective pressures during varying stages of human evolutionary history. We also found three genome regions with evidence of recent positive selection, which were dissected to propose targets of selection and discuss on the possible underlying selective pressures in each case.

Resumen

La selección natural ha moldeado de forma importante la variación genética humana, por lo que encontrar variantes que hayan sido seleccionadas positivamente puede dar indicios acerca de los genes que determinan la diversidad fenotípica humana. En este trabajo comparamos los genes del genoma humano, del chimpancé y de varias especies cercanamente emparentadas para detectar aquellos genes bajo selección positiva en el linaje humano, y posteriormente evaluar la variación intraespecífica en un subconjunto de genes con la intención de buscar eventos de selección reciente en poblaciones humanas de todo el mundo. Nuestros resultados sugieren que la mayoría de genes implicados en eventos selectivos durante la evolución temprana del hombre no son los mismos que aquellos involucrados en adaptaciones humanas recientes, lo que implica la existencia de diferentes presiones selectivas a lo largo de las distintas etapas de la historia evolutiva humana. También encontramos tres regiones genómicas con evidencias de selección positiva reciente, las cuales fueron analizadas en profundidad para proponer posibles dianas de selección y discutir las presiones selectivas subyacentes en cada caso.

Preface

It is quite exciting, and challenging, to write a thesis about natural selection when the scientific community and the society is talking about it the most, and of course they are: 200 years ago today (on February 12, 1809) Charles Darwin was born. He postulated that all species of life have evolved over time from common ancestors through the process he called natural selection. His famous book *On the Origin of Species by Means of Natural Selection* was published in 1859 so that besides the bicentenary of Darwin's birth, this year commemorates the 150th anniversary of the publication of a theory that still influences modern research today.

This thesis focuses on natural selection in humans, but at the same time it tries constantly to put in context the position of our species within nature. This obeys to methodological reasons but also, and more importantly, to my intention of stressing the importance of transmitting a broad view of ourselves in which we acknowledge to be just a little branch of the tree of life.

This may seem an obvious idea for most biologists, but within medical circles this is not yet a common concept as the study of the human body is usually intended to solve health problems and not to understand the underlying mechanisms that physiologically (and pathologically) make us the way we are and, thus, set us apart from other forms of life on Earth. The answer to these intriguing issues must be sought in the light of biology and evolution, which before starting this PhD I mostly ignored.

Back in late 2002, while seated in the middle of the campus of the Mexican University of Guadalajara, right after completing my medical studies, I was reading a book about Mendel's biography. The way he ended up establishing the basic laws of inheritance by means of his hybridization experiments in peas, or the way in which Darwin reasoned his theory triggered by his 5-year voyage on the HMS *Beagle*, was telling me how personal circumstances greatly influence scientists' future steps.

Then, I suddenly started wondering: if I had the privilege of choosing the research field to which devote myself, what question would I like to answer? So I wrote two questions in the last page of the book as a personal commitment to find some place where to learn how to answer them, one of which was: "*what make us humans?*". So without even knowing that fields such as "biological anthropology" or "evolutionary genetics" existed, I started a personal journey that led me to end up getting married to a physical anthropologist, and writing a PhD thesis about human evolutionary genetics.

This document is therefore not only the presentation of a number of results obtained during five years of work, but also the compilation of what I have learned during this time, so that I hope it will be consulted by future students and possibly may help to encourage them to get involved in this fascinating field.

For this reason, besides trying to provide as complete as possible information about the current state of the most recent advances, I intentionally also tried to keep the main concepts as basic as possible, so that researchers and students from a broad spectrum of disciplines can go through the text and hopefully find it useful.

In the first part, an overview is given about human origins based on genetic evidence, followed by an introduction to human genetic variation including the recent generation of genome-wide catalogues of variation. Next, the main concepts of the evolutionary processes shaping diversity are explained with particular remarks relevant to the study of human populations. From here on, the thesis focuses on natural selection, particularly on positive selection. The most widely used methods for detecting selection in humans are then reviewed and some examples of positive selection in the human genome are given. This part is intentionally divided into methods (and examples) based on **interspecific divergence** and **intraspecific diversity** because subsequent presented results consist on the separate analysis of divergence and diversity in a set of candidate genes. By doing so, we address natural selection occurring at different stages of human evolutionary history. Finally, we derive conclusions based on our results about specific adaptive changes occurring in human populations and argue how these may contribute to define our species and in some cases to differentiate particular populations around the world.

This work is a small piece of knowledge towards the end of providing some new insight about the genetic basis of human adaptation, whose traces have been imprinted in our genomes for ages, but only since very recently we are able to explore them at the molecular level.

Barcelona, February 2009

Index

Abreviations

AAAS: American Association for the Advancement of Science
BEB: Bayes Empirical Bayes
CEPH: Centre d'Étude du Polymorphisme Humain
CLR: Composite likelihood ratio
cM: centiMorgan
CNS: Coding non-synonymous
CNV: Copy number variant
CRTRs: Contiguous Regions of Tajima's D Reduction
DAF: Derived allele frequency
DNA: Deoxyribonucleic acid
EHH: Extended Haplotype Homozygosity
Hb: Hemoglobin
HGDP: Human Genome Diversity Project
ITIM: Immunoreceptor tyrosine-based inhibitory motif
Kb: Kilo bases
Kya: Thousand years ago
Kyr: Thousand years
LD: Linkage disequilibrium
LRH: Long range haplotype
LRT: Likelihood ratio test
LUCA: Last universal common ancestor
MAF: Minor allele frequency
Mb: Megabase
mtDNA: mitochondrial DNA
Mya: Million years ago
Myr: Million years
OLA: Oligonucleotide ligation assay
OR: Olfactory receptor
PAML: Phylogenetic Analysis by Maximum Likelihood
PCA: Principal component analysis
POPRES: Population reference sample
RFLPs: Restriction fragment length polymorphisms
SFS: Site frequency spectrum
SNP: Single nucleotide polymorphism
STR: Short tandem repeat

INTRODUCTION

1 Human Origins

We humans have inevitably an anthropocentric view of nature. We see ourselves as unique in comparison to other living organisms. Indeed, after thousands of years of evolution, humans have become the most dominant primate species on Earth. An incomparable capacity for culture has allowed us to grow in number, to extend our range to almost all regions of the planet, and to impact—for better and for worse—the lives of many other animals, plants, and ecosystems. This raises the question of what processes during human evolution have led to such a clear distinction between our species and other forms of life. In other words, what makes us humans?

A straightforward and fascinating aspect of human uniqueness is the same fact of contemplating questions such as the one posed above. For answering it, however, we definitely need to step back and take a broader view of ourselves. Although unique, we are just one of many millions of living species on the planet, and to understand ourselves, we need to understand the evolutionary history that we share with all forms of life on Earth, from the Last Universal Common Ancestor (LUCA) to our closest living relatives. First, it is clear that humans are not at the *top* of evolution neither *more evolved* than other living organisms, as some misconceptions in popular thinking point out. Instead, all modern species are equally derived from their common ancestors in terms of time (see **Figure 1**).

All living beings are adapted to their present ecological niches. Natural selection is ongoing, and as a consequence important adaptive changes have taken place in all organisms. All the phenotypes and genes of organisms alive today derive from ancestors that can be traced back in time. Most of these features have been shaped by the environmental challenges faced by these organisms and their ancestors. It is thanks to our shared evolutionary heritage with every other species on the planet that comparative analyses with bacteria, yeasts, flies, worms, fish, mice or chimpanzees, have much to say about ourselves.

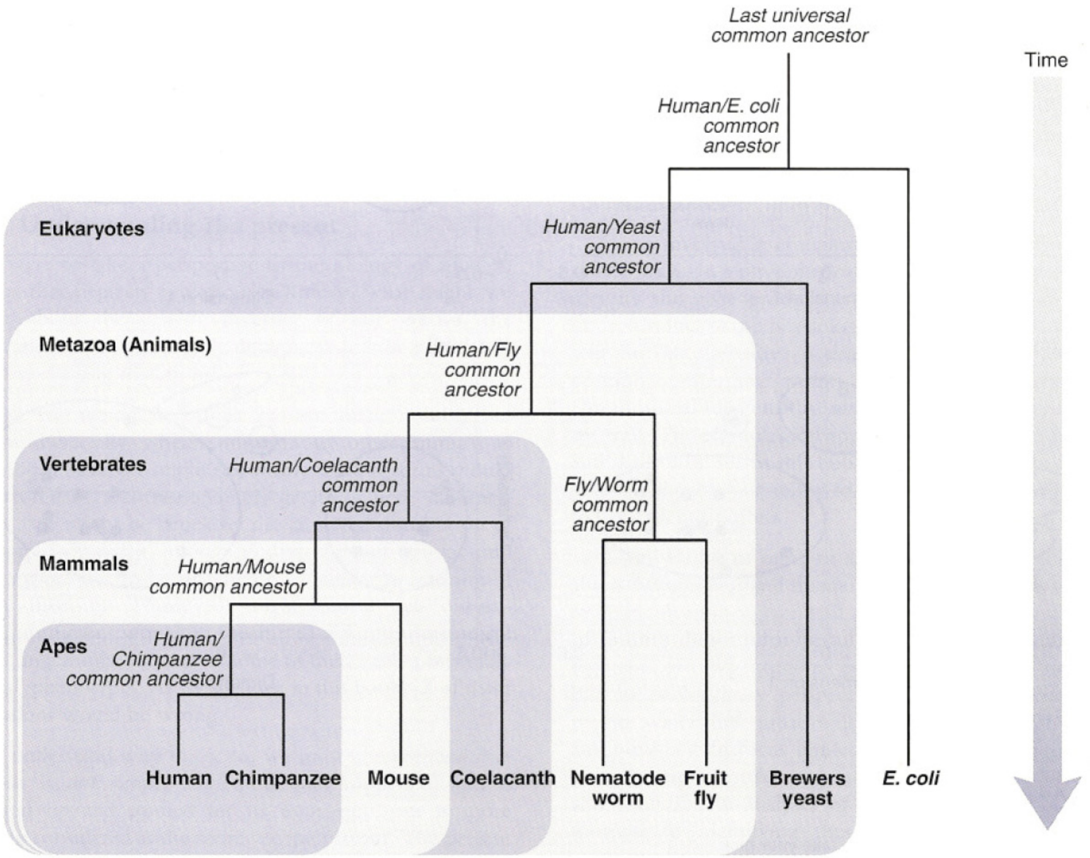

Last universal common ancestor
Time
Human/E. coli common ancestor
Eukaryotes
Human/Yeast common ancestor
Metazoa (Animals)
Human/Fly common ancestor
Vertebrates
Human/Coelacanth common ancestor
Fly/Worm common ancestor
Mammals
Human/Mouse common ancestor
Apes
Human/ Chimpanzee common ancestor
Human
Chimpanzee
Mouse
Coelacanth
Nematode worm
Fruit fly
Brewers yeast
E. coli

Figure 1: A phylogenetic tree showing that all living organisms share a common ancestor some time in the past. Contemporary species shown here are just representative examples of each clade (adapted from (Jobling et al. 2004)).

To understand how genomes evolve, it is important to know how the species carrying the genomes are related to each other, that is, their phylogeny.

1.1 The Human Lineage

We are one of the ~300 primate species that currently live on Earth. Within primates, we share the group of hominoids with our closest living relatives: the apes. The apes in turn include the gibbons and orangutans in Asia, and the gorillas, chimpanzees, and bonobos in Africa. The phylogeny of the hominoids is generally accepted (see **Figure 2**), in which we observe that the African apes are more closely related to humans than the orangutans. Early evidences for that date back to 1863, when Thomas Huxley stated: “it is quite certain that the Ape which most nearly approaches Man, in the totality of its organization, is either the Chimpanzee or the Gorilla” (Huxley 1863).

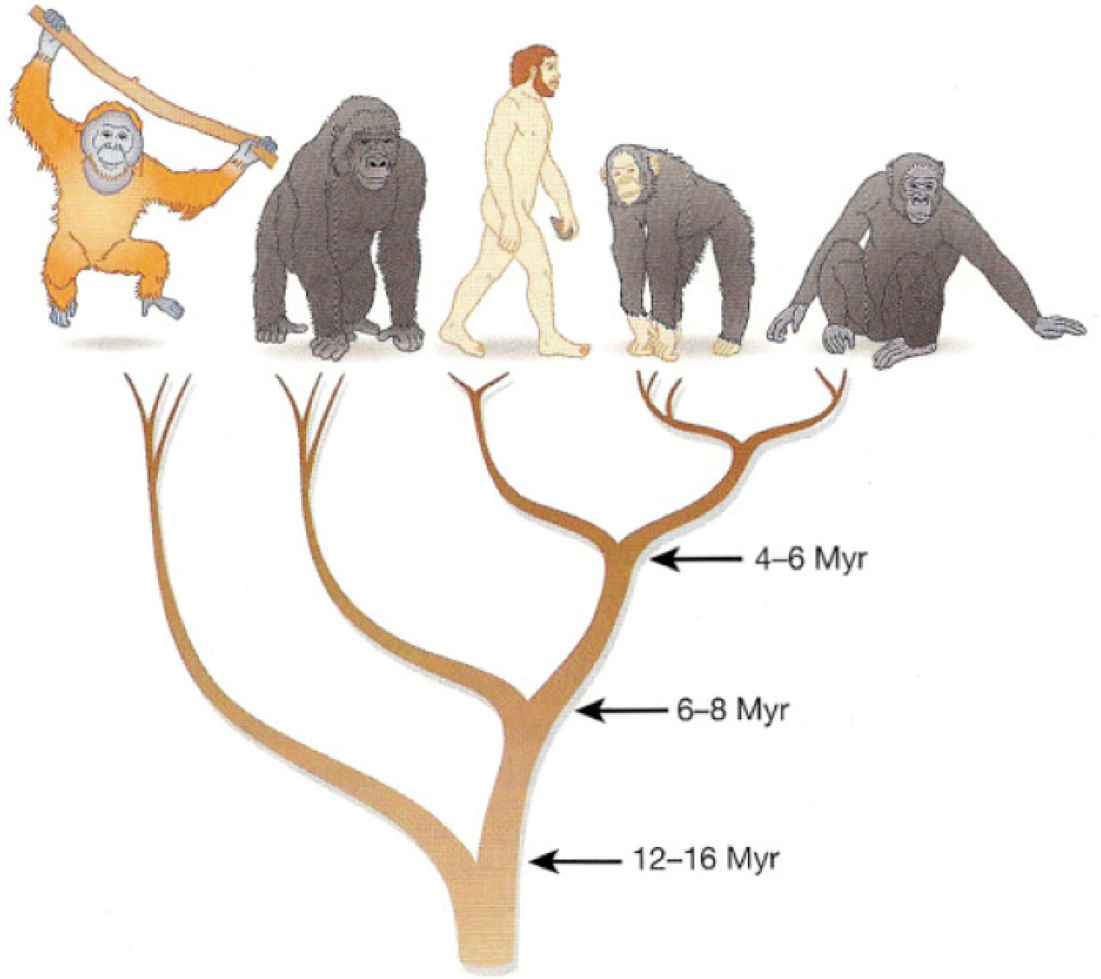


Figure 2: Illustration of the divergence of human and ape species. Approximate dates of divergence are given, from left to right, for orangutan, gorilla, human, bonobo and chimpanzee (adapted from (Paabo 2003)).

However, it remained unclear for a long time whether chimpanzees or gorillas are closer to humans. Nowadays, deoxyribonucleic acid (DNA) complete sequences of both human and chimpanzee genomes are available, and the gorilla genome project is ongoing; hence this question can be addressed from a genomic perspective. Humans are ~98.8% similar to chimpanzees at the nucleotide level (Consortium 2005a) and the majority of regions in our genome are indeed most closely related to chimpanzees and bonobos, although a non-trivial fraction is more closely related to gorillas (Chen et al. 2001). This is because the speciation events that separated these lineages occurred so closely in time that genetic variation in the first ancestral species, from which the gorilla lineage derived, survived until the second speciation event between the human and chimpanzee lineages (Enard and Paabo 2004). The estimation of divergence times between primate groups is more controversial than the phylogeny because dating is based on paleontological data which has uncertainties associated with fossil calibration points. Nevertheless, it seems clear that the human evolutionary lineage diverged from that of chimpanzees about 4-6 million years ago, from that of gorillas about 6-8 million years ago, and from that of the orangutans about 12-16 million years ago (Chen et al. 2001; Glazko and Nei 2003) (see **Figure 2**). By no means, however, this implies that we were alone in our evolutionary lineage. Many other species existed, but eventually became extinct. We are then the only surviving species of a formerly rich and diverse *Homo* lineage. For example, molecular information is available from *Homo neanderthalensis* remains. Neanderthals lived in Europe and western Asia until about 30,000 years ago. Mitochondrial (mt) DNA sequence data indicates that Neanderthals carried mtDNA sequences that fall outside the variation of modern humans and that diverged approximately 500,000 years ago (Lahr and Foley 1998) (see **Figure 3**). Thus, although extinct, they are the group of hominids most closely related to contemporary humans.

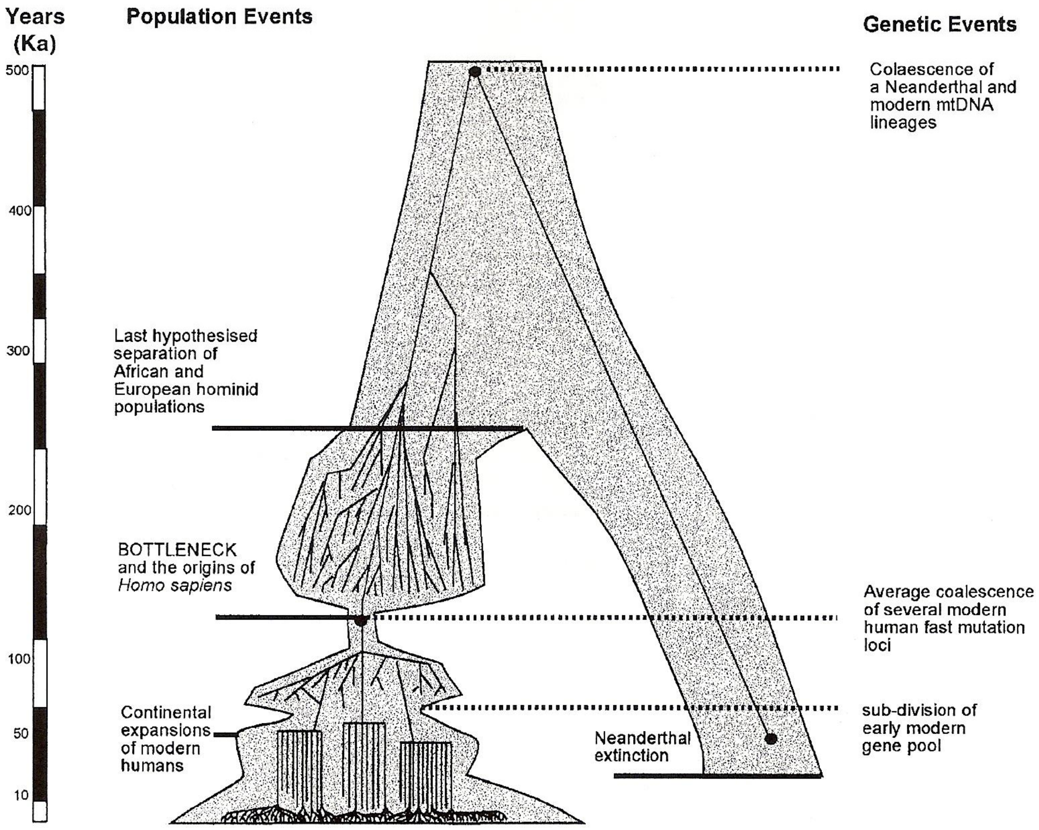

Years (Ka)
500
400
300
200
100
50
10
Population Events
Genetic Events
Colaescence of a Neanderthal and modern mtDNA lineages
Last hypothesised separation of African and European hominid populations
BOTTLENECK and the origins of *Homo sapiens*
Average coalescence of several modern human fast mutation loci
Continental expansions of modern humans
Neanderthal extinction
sub-division of early modern gene pool

Figure 3: Schematic representation of the relationship between the population history of humans and Neanderthals based on a mtDNA genealogy (adapted from (Lahr and Foley 1998)).

An ongoing project aimed at sequencing the Neanderthal genome has recently completed the first draft version (AAAS press release, February 12, 2009), which will allow closer, and previously impossible, genome-wide comparisons to our genome.

In comparison to genetic data from chimpanzees, gorillas, and orangutans, humans display lower levels of intraspecific diversity (Kaessmann et al. 2001). As discussed in more detail later, humans differ on average at only 1 of every 500-1,000 nucleotides between chromosomes. This degree of diversity is less than what typically exists for instance among chimpanzees; which is surprising because humans are by far much more numerous and because humans are distributed over the entire world whereas chimpanzees and other apes are more restricted in their distribution. One explanation for this is that humans can be traced back on time to a small population that expanded in the relatively recent past, which is a concept directly related to the different theories of modern human origins and their subsequent dispersal.

1.2 Dispersal of Modern Humans

Approximately 100,000 years ago the planet was inhabited by a morphologically heterogeneous group of hominids: *Homo neanderthalensis* in Europe, *Homo erectus* in Asia, and *Homo sapiens* in Africa. Around 30,000 years ago, most of this diversity had disappeared: only anatomically modern humans were occupying the entire Old World. How did this transition occur?

Early models proposed that human "races" were distinct biological species that originated independently with no gene flow between them (Gould 1981). Later models based on the fossil record, such as the Multiregional Origin model, propose that after the migration of *H. erectus* out of Africa ~0.8 to 1.8 million years ago, there has been parallel evolution from *H. erectus* to *H. sapiens* among geographically dispersed populations, with limited gene flow between populations (Wolpoff 1996).

In contrast to these models, which predict that populations from distinct geographic regions have been differentiating over long periods of time, the genetic data accumulated over the past two decades overwhelmingly support the Recent African Origin model, also called the Out of Africa model (Cann et al. 1987; Vigilant et al. 1991). According to this model, all non-African populations descend from an anatomically modern *H. sapiens* ancestor that evolved in Africa ~200 thousand years ago (Kya) and then spread and diversified throughout the rest of the world starting ~50-100 Kya, replacing any archaic *Homo* populations still present outside of Africa, such as *H. neanderthalensis* in Europe and *H. erectus* in Asia (Stringer and Andrews 1988). Thus, the Out of Africa model predicts a recent common African ancestor with subsequent recent expansions after the initial migration(s) out of the African continent ~100 Kya (see **Figure 4**).

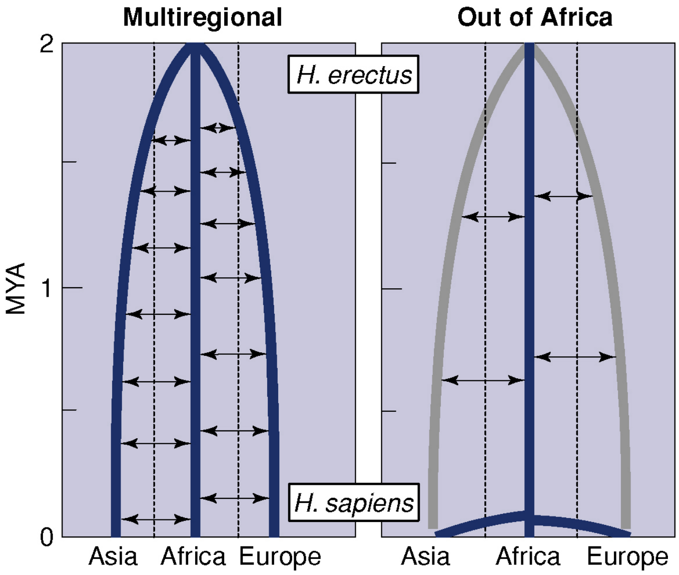


Figure 4: Models of recent human evolution. These diagrams illustrate two contrasting models of modern human origins, both beginning with *H. erectus* shortly after 2 MYA and leading to contemporary humans. Horizontal arrows indicate gene flow between populations. Blue lines denote ancestors of modern humans and gray lines represent lineages that became extinct. The Multiregional Model is now very much a minority view and, thus, no longer one of the serious models (adapted from (Jobling et al. 2004)).

As humans colonized the world (see **Figure 5**), populations diverged and thus differentiated genetically. Hence the dispersal of modern humans can be inferred from extant diversity of different parts of our genome (i.e., autosomes, mtDNA or the Y chromosome). Indeed, the history of human expansions into Australo-Melanesia (~60 Kya), Europe (~45 Kya), Asia (~35 Kya), the New World (~15-20 Kya), and the Pacific (~3 Kya) is genetically supported by patterns of allele frequency variation (Cavalli-Sforza et al. 1994). This begs the question of how much genetic variation occurs within humans; which in agreement with our curiosity at any level, means that we want to understand what make us unique, not only as species but also as individuals. For discussing about that, some basic concepts about the structure and variation of the human genome need to be reviewed.

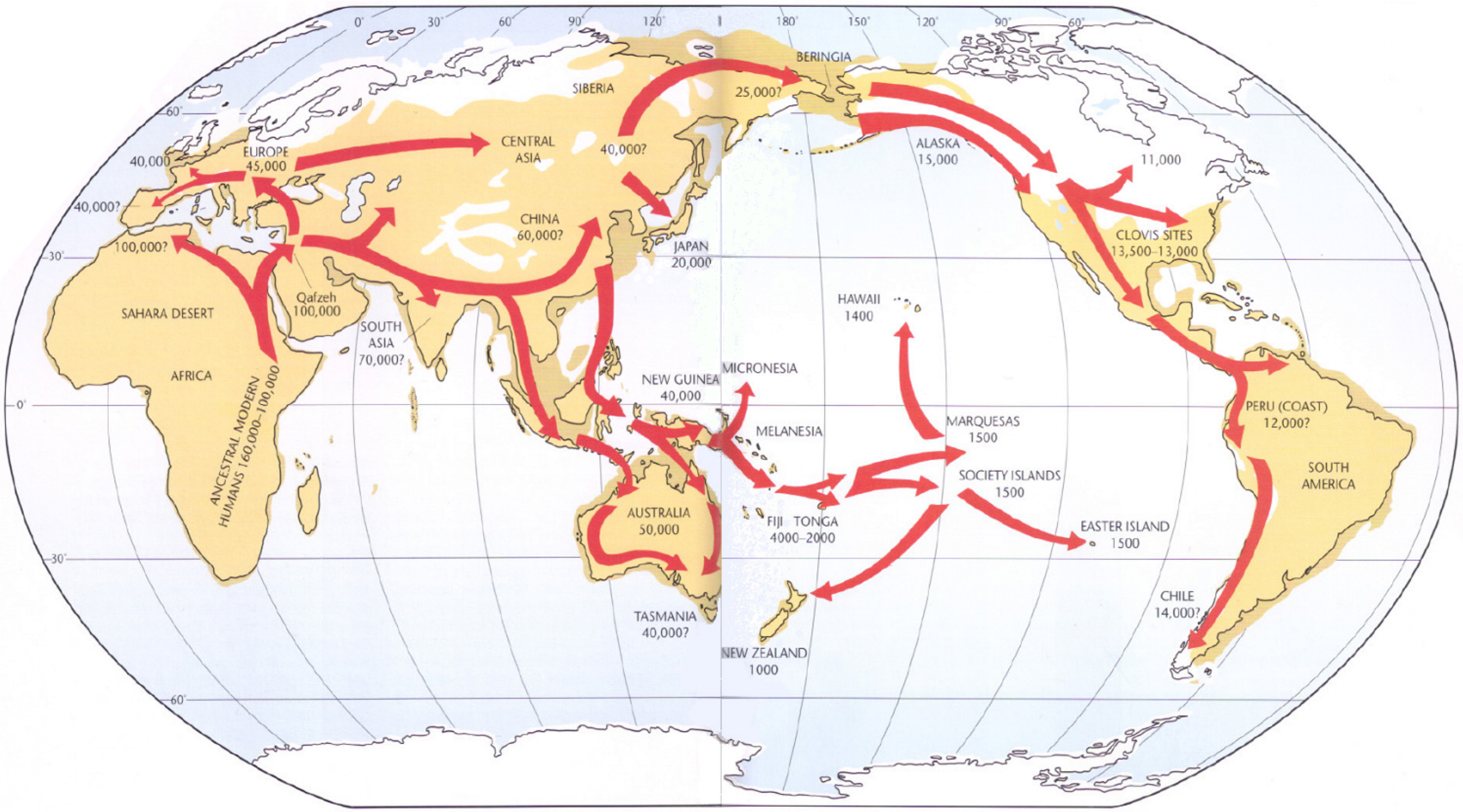


Figure 5: World map of the dispersal of modern humans. Starting in Africa, arrows indicate migrations with approximate dates (years ago). Ice sheets and low sea levels of *circa* 18,000 years ago are also shown. This illustration is orientative and not necessarily reflects the accuracy of all studies carried out to date (adapted from (Stringer and Andrews 2005)).

2 Human Genetic Variation

Our genome consists of about 3 billion nucleotides, which are the monomeric subunits making up the DNA molecule. Nucleotides vary in their *bases*, which can be adenine (A), guanine (G), cytosine (C) or thymine (T). It is the sequence of these four different nucleotides that carries the genetic information, which have been passed down to us from our ancestors. In every generation, several of these nucleotides are affected by mutations in the male and female germ-line so that subsequent generations receive slightly different versions of the ancestral genomes. Single nucleotide polymorphisms (SNPs) represent by far the majority of these variants (Build 129 of dbSNP database contains ~15 million SNPs, of which 6.5 million are validated), so that common estimates initially pointed that all humans are ~99.9% identical at the nucleotide sequence level (Przeworski et al. 2000; Reich et al. 2002). Nonetheless, this may be an overestimate since genomes can differ in many other ways. Bits of DNA ranging from a few to many thousands of bases (Kb) can get lost (deletions), added (insertions), or turned around (inversions). Even more, there are structural variations, known as copy number variants (CNVs), that change the number of copies of a gene or any piece of DNA (see **Figure 6**).

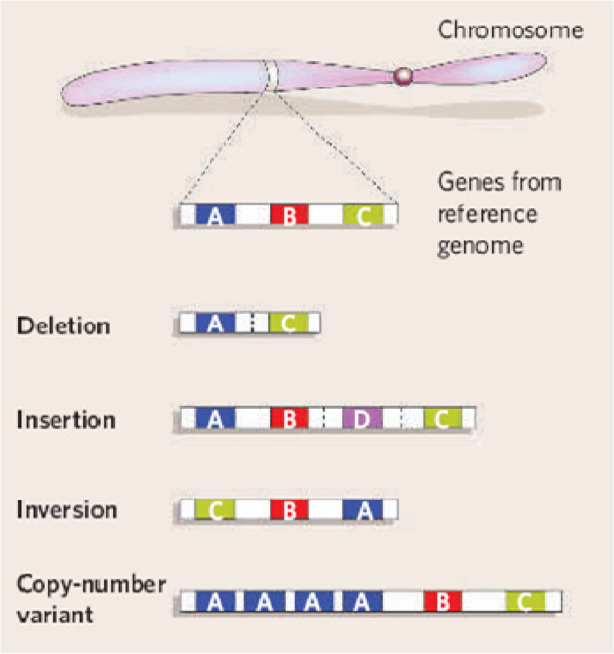


Figure 6: Examples of structural variations in the genome. Rectangles represent segments of DNA and letters A to D schematically illustrate different genes (adapted from (Pennisi 2007)).

Therefore, current estimates range from ~99.6 to 99.8% of genome-sequence identity between humans (Feuk et al. 2006). The genetic variation contained in the remaining 0.2–0.4%, in spite of representing a very small fraction of the total genome, is enough to ensure most of our individual uniqueness at the DNA level. Both structural variants and SNPs may be equally important in modern population genetics, however, given the nature of the data presented in this work, a more detailed background will focus on SNP variation as we move further on.

2.1 SNPs

As briefly introduced above, SNPs are the smallest unit of polymorphism and the most common one in the human genome. A SNP is said to be present at a particular nucleotide site if the DNA molecules in the population frequently differ in the *alleles* observed at that position. For example, some DNA molecules in a population may have an **A** at a particular site, whereas others in the same population may have a **C** at the same site. Most SNPs are biallelic, that is, two alleles or forms for the polymorphism exist. The most frequent is usually referred to as "major allele" and the other as "minor allele". Most SNPs occur in DNA of no known function, but common variants also occur in coding regions of genes, altering amino acid sequences of proteins (see **Figure 7**), and in regulatory regions that affect gene expression.

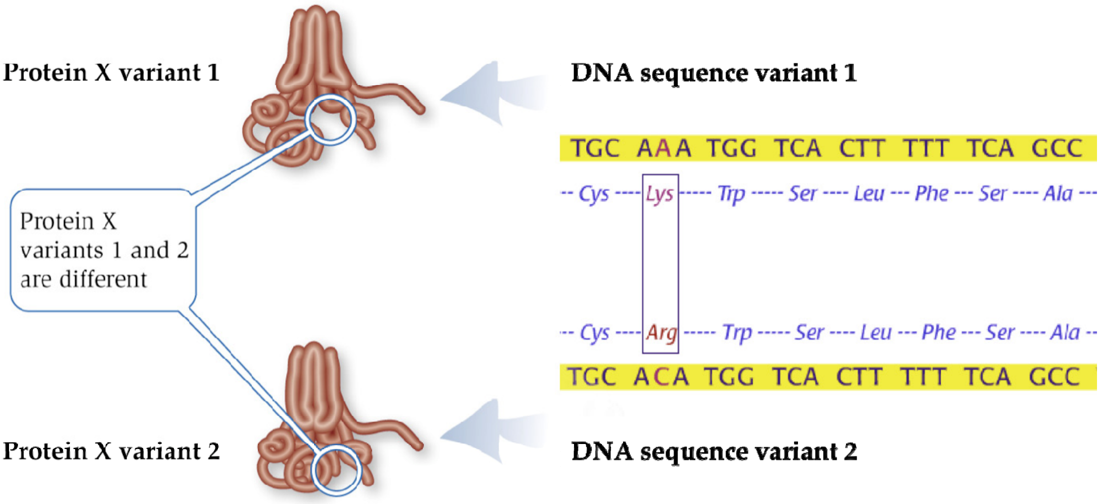

Protein X variant 1
DNA sequence variant 1
TGC AAA TGG TCA CTT TTT TCA GCC
-- Cys ---- Lys ----- Trp ----- Ser ---- Leu --- Phe --- Ser ---- Ala ---
Protein X variants 1 and 2 are different
-- Cys ---- Arg ----- Trp ----- Ser ---- Leu --- Phe --- Ser ---- Ala --
TGC ACA TGG TCA CTT TTT TCA GCC
Protein X variant 2
DNA sequence variant 2

Figure 7: Example of a nonsynonymous SNP affecting the structural conformation of the encoded protein (hypothetically denoted as Protein X) (modified from educational resources of the Genetic Science Learning Center, University of Utah).

Coding SNPs that result in amino acid replacements are known as nonsynonymous SNPs, which will be referred to as CNS (Coding Non-Synonymous) in the chapters of results of this work. On the contrary, synonymous SNPs (CS or Coding Synonymous) are those present in coding regions but that do not result in an amino acid replacement when the sequence is translated into protein. Unlike synonymous and most noncoding SNPs, nonsynonymous SNPs together with functional SNPs in regulatory regions, are believed to possibly have a straightforward effect on the phenotype, and may therefore be important for the adaptation of the organism.

2.2 Haplotypes

A haplotype is a unique combination of alleles along a given stretch of a single chromosome. The variation found along chromosomes is often structured in blocks of haplotypes, which vary in length both among populations and genome regions. These "haplotype blocks" are likely to result from the fact that recombination (the reshuffling of chromosome segments that occurs during meiosis, that is, the formation of gametes), tends to occur in certain areas of the chromosomes more often than in others (Daly et al. 2001). Such areas have been referred to as "recombination hotspots" (see **section 3.1**). Two areas of the genome are exceptional (i.e., mtDNA and the non-recombining region of the Y chromosome), since they are not reshuffled during the formation of gametes. In the rest of the genome, although the extension of haplotype blocks depends on the methods used to define them, they are typically 5–200 Kb in length, and as few as four to five common haplotypes account for most of the variation in each block (Paabo 2003) (see **Figure 8**).

The catalogue of haplotypes for every block makes up the "haplotype map" of the human genome. Since 2005 this haplotype map is available for three different human populations as a result of the International HapMap Project (Consortium 2005b), which is currently in the phase of increasing the number of populations up to eleven (HapMap phase 3). Before discussing in more detail the influence that this and other ongoing projects are having in modern population genetics, a basic concept –directly related to the existence of haplotype blocks– needs to be reviewed.

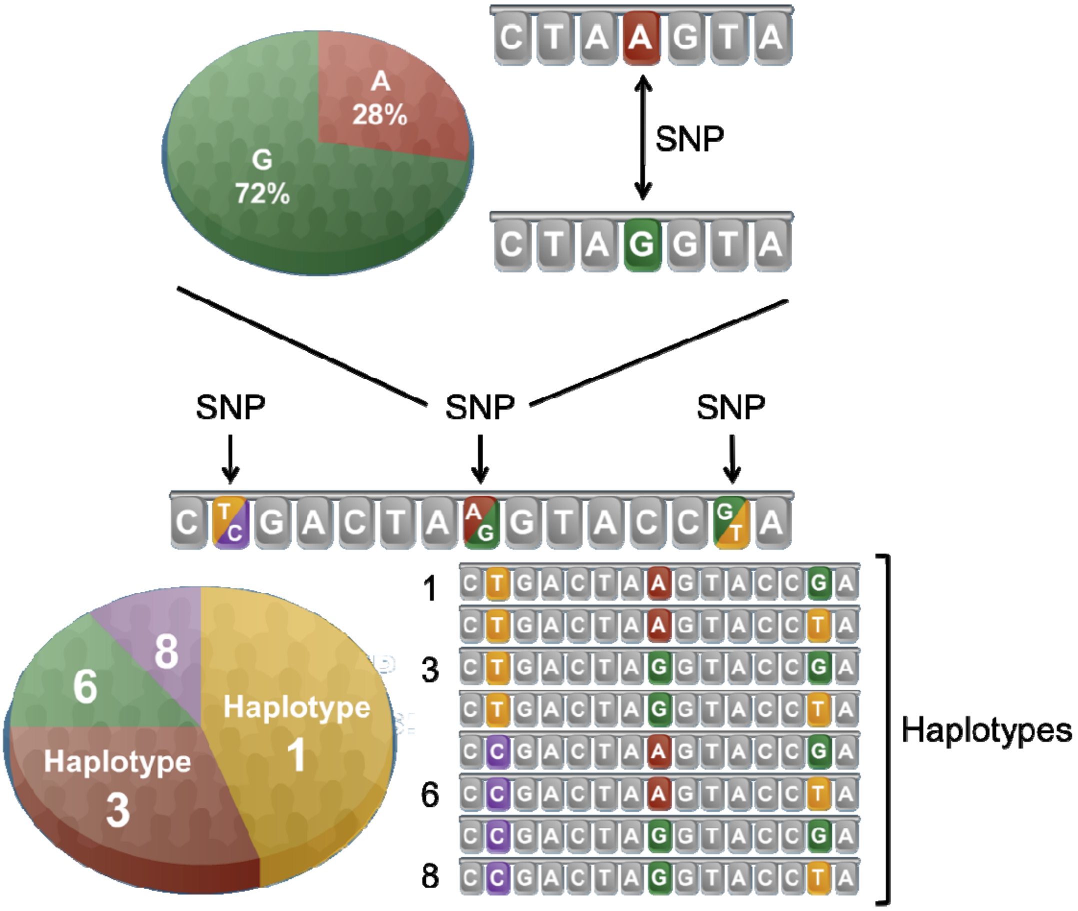
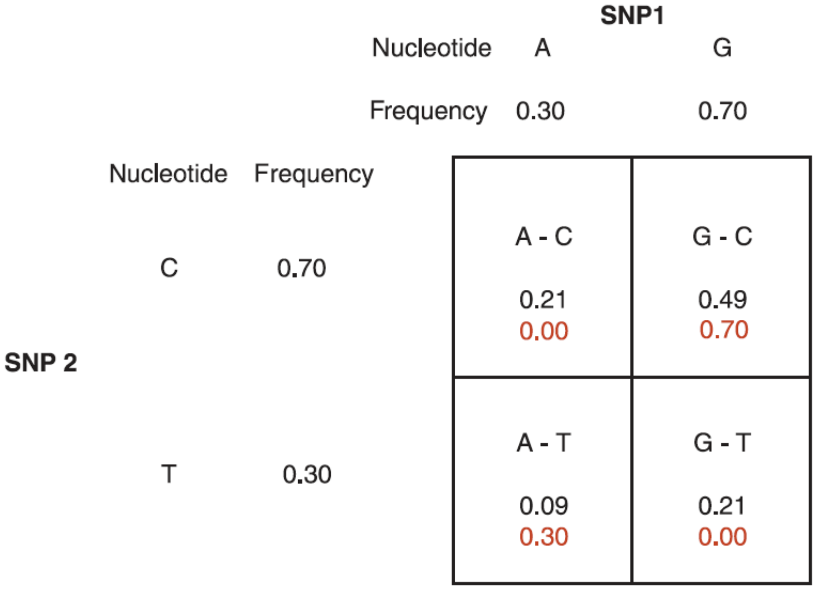


Figure 8: Schematic representation of a haplotype block. For simplicity, no more than three SNPs are illustrated along the DNA sequence. Note that, out of eight possible haplotypes, only four of them account for all the variation found in the population (modified from resources of the Genetic Science Learning Center, University of Utah).

2.3 Linkage Disequilibrium

Before defining linkage disequilibrium (LD), it is convenient to explain what linkage equilibrium is. For that, a simple example will be given. Assume the presence of two SNPs in a DNA sequence, where the alleles of SNP 1 (A/G) are present in 30% and 70% of the population respectively; and SNP 2 (C/T) has frequencies of 70% and 30% respectively. If recombination has acted long enough on the DNA sequence to exchange information between chromosomes, the frequencies of the haplotype A–C will be 0.30 x 0.70 = 0.21, those of the haplotype G–C will be 0.70 x 0.70 = 0.49 and so on (see **Figure 9**). In other words, haplotype frequencies are equal to the product of allele frequencies (i.e., there is **linkage equilibrium**). Conversely, if there are nonrandom associations between alleles due to their tendency to be coinherited more often than other combinations, the two SNPs are said to be in **linkage disequilibrium**. Measures of LD include *D* (Lewontin 1964), |D'| and r^2. Both |D'| and r^2 are calculated in a way that complete or perfect LD equals 1, indicating that the two alleles have not been separated by recombination. However, r^2 –defined as the squared correlation coefficient between two loci– is less affected by small sample sizes than |D'| and thus is one of the most widely used measures.



			SNP1	
		Nucleotide	A	G
		Frequency	0.30	0.70
SNP 2	Nucleotide	Frequency		
	C	0.70	A - C 0.21 0.00	G - C 0.49 0.70
	T	0.30	A - T 0.09 0.30	G - T 0.21 0.00

Figure 9: Linkage disequilibrium illustration. An example of two SNPs that at either in linkage equilibrium (random association of alleles according to allele frequencies -shown in black-) or linkage disequilibrium (nonrandom association of alleles -shown in red-) (adapted from (Kaessmann and Paabo 2002)).

LD is basically the result of reduced recombination between loci, reflecting that not enough time has elapsed since the mutations arose for recombination to shuffle them around and reach linkage equilibrium. Therefore, the extent of LD is extremely informative about the evolutionary history of the sample being analyzed. In fact, many studies have demonstrated that there is extensively more LD outside Africa than what has been found in African populations (Calafell et al. 1998; Kidd et al. 2000; Reich et al. 2001; Tishkoff et al. 1996), meaning that the gene pool in Africa is older than that outside Africa, reinforcing the theory of an African origin of modern humans.

Besides varying among populations, LD also differs between different parts of the genome. As a consequence of the probability of recombination events, LD typically declines with increasing physical distance between SNPs, meaning that far-apart loci are not expected to be in strong LD. On the other hand, closely related SNPs typically show high values of LD between them.

This fact has important implications, as a limited set of SNPs, by being in LD with other SNPs in their vicinity, would be enough for capturing indirectly most of the variation in a given region or even in the whole genome. Moreover, the same approach would enable anonymous markers (i.e., SNPs with unknown effect on gene function) to serve as proxies to tag functional variants associated with a particular phenotype, for instance a complex disease. The construction of a genome-wide catalogue of such markers, known as *tagSNPs*, was a strong motivation for the launching of the International HapMap project, as tagSNPs would facilitate the conduction of large-scale genetic association studies of human diseases.

The HapMap project took advantage of the existence of LD to genotype with a few millions of markers the whole genome in three human populations (see below). Such approach allowed avoiding the much more expensive and time-consuming task of resequencing the complete genomes of the studied individuals. Nonetheless, current genotyping and sequencing technologies are surprisingly becoming almost equally accessible, so this situation may change in the near future. Actually, as of November 2008, current sequencing efforts aimed at producing finer catalogs of human variation (Kaiser 2008) are already generating resequencing data equivalent to >1,000 x coverage of the human genome. Such dramatic change has occurred in less than 10 years since the initial draft of the human genome was announced, proving that the scales of generating and analyzing data are rapidly increasing (and do not seem to stop). Next, some of the main projects influencing the genomic era and the way we currently conduct our research will be reviewed.

2.4 Catalogs of Human Variation: from The Human Genome to 1000 Genomes

Since the completion of the Human Genome Project (HGP), research focused on human genetic variation has intensified. The first draft of the human genome sequence was provided by two independent sequencing efforts (Lander et al. 2001; Venter et al. 2001). Far from consisting in a single sequence of *the* human genome, it was a mosaic of many different genomes. Therefore, the project yielded about 4 million SNPs that were discovered as the differences between individual sequences arose during the assembly.

As a complement of the Human Genome Project, the Human Genome Diversity Project (HGDP) was proposed in 1991 by a group of geneticists led by Luca Cavalli-Sforza and Allan Wilson. The HGDP project, aimed at collecting, analyzing, and making available for research a broad set of human samples from al around the globe, unfortunately did not achieve large-scale success due to ethical and political problems. Nevertheless its greatest achievement has been the establishment of a diversity panel of 1,064 cell lines (Cann et al. 2002).

In 2003 the International HapMap project was launched, with the primary goal of developing a haplotype map of the human genome that describes the common patterns of genetic variation. As part of phase I and II of the project, approximately 3.9 million SNPs have been genotyped in 270 individuals from three different human populations (Consortium 2005b). The samples came from an African population in Nigeria (Yoruba), a mostly Utah (U.S.) population of European ancestry, and a sample drawn from the Japanese and Han Chinese populations. The third phase of the project (HapMap 3) comprises 1,301 samples (including the original 270 individuals) from 11 populations. Newly introduced ancestries include Gujarati Indians and individuals of Mexican ancestry. HapMap 3 data, however, is not as dense as that available for the three original populations. Besides providing a wealth of information about the patterns of LD in human populations, the HapMap data set constitutes a valuable resource for many SNP-based studies in human populations.

Other catalogs of human variation have been generated, such as the Perlegen data set, which consists in a survey of human genetic variation by genotyping 1.6 million SNPs in 71 Americans of European, African, and Asian ancestry (Hinds et al. 2005). A more recent survey (Li et al. 2008) analyzed a smaller number of SNPs (~650, 000) but in a considerably larger set of samples (1,043 individuals from 51 populations of the HGDP-CEPH Diversity Panel) consisting in the highest resolution map of worldwide human genetic diversity to date.

At the same time, sequencing efforts of complete human genomes have been ongoing. In May 2007, DNA co-discoverer James Watson’s genome sequence was deciphered by massively parallel DNA sequencing (Wheeler et al. 2008). Whereas in November 2008, two groups revealed individual genome sequences of a Yoruba man from Ibadan, Nigeria (Bentley et al. 2008), and of a Han Chinese individual (Wang et al. 2008); and this time for a small fraction of the cost of the human genome's first drafts or subsequently published editions (Consortium 2004).

The most recent and ambitious collaborative effort is the so-called *1000 Genomes Project*, which is the natural consequence of the HapMap Project in the search of a deeper characterization of human genetic variation. The latter includes only the most common variants (i.e., SNPs present in at least 5% of the population), whereas the 1000 Genomes project intends to find essentially all SNPs and detectable CNVs that occur at >1% of frequency in each of multiple human population samples. In order to do so, the project will involve the sequencing of complete genomes of at least one thousand individuals from around the world. By accomplishing this, the 1000 Genomes project will create the most detailed catalogue of human polymorphism. The first official data from a series of pilot projects is expected to be released in January 2009 (Kuehn 2008).

It is essential then to think about "the human genome" not as a single and steady entity across humans, but in terms of diversity as a unique mosaic of sequences, which ultimately make us individuals.

2.5 Global Patterns of Genetic Variation

Much has been learned about the worldwide distribution of human variation since genetic data is available. Evidences come from different types of markers and data range from small fractions of the genome, such as the X, the Y chromosome or the mtDNA, to genome-wide surveys.

In agreement with a genetic bottleneck occurring at the time of migration of modern humans out of Africa, most results point to the existence of higher levels of genetic variation in African populations than in non-Africans (Tishkoff and Verrelli 2003). Additionally, Africans have the largest number of population-specific alleles and non-Africans carry only a fraction of the genetic diversity that is present in Africa (Calafell et al. 1998; Kaessmann et al. 2001; Quintana-Murci et al. 1999). Although some exceptions to this pattern do exist, in which higher variability has been found in Europeans, they may be biased because the polymorphisms (mainly RFLPs – Restriction Fragment Length Polymorphisms– and few SNPs) were first identified in non-African populations (Kidd et al. 2004).

Populations also differ with respect to the organization of haplotypes and linkage disequilibrium. Levels and patterns of LD depend on gene-specific factors, such as mutation and recombination, as well as demographic factors that have a genome-wide effect, such as population size, population structure, founder effect and admixture. Numerous studies of LD between SNPs show greater extent of LD in Eurasians than in Africans and still greater LD in Native Americans (Gabriel et al. 2002; Kidd et al. 2000; Reich et al. 2001). This pattern of LD is consistent with human demographic history; ancestral African populations have had more time for recombination and mutation to reduce LD. The bottleneck associated with the expansion of modern humans out of Africa resulted in many African haplotypes being lost, leading to greater LD in non-African populations. Another bottleneck, associated with the expansion into the Americas, is reflected in the even higher amounts of LD in this region. In agreement with this, Jakobsson et al. (2008), by genotyping 525,910 SNPs in a worldwide sample of 29 populations, have demonstrated that increasing LD is observed with increasing geographic distance from Africa, supporting the hypothesis of a serial founder effect for the out-of-Africa spread of human populations.

Therefore, we can conclude that, within the human gene pool, most variation is found in Africa and what is seen outside Africa is as subset of the variation found within Africa (see **Figure 10**).

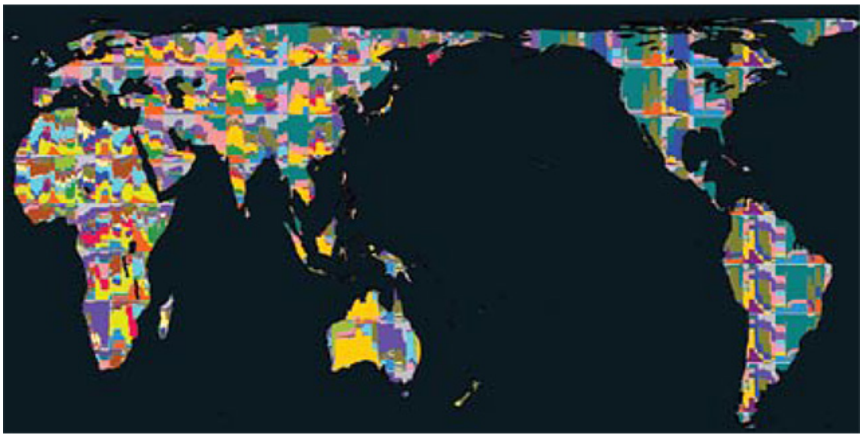


Figure 10: A schematic genomic map of the world showing that human genetic diversity decreases outside of Africa. Each colored tile represents a common haplotype. Africa has more tiles than found on other continents and ones that correspond to haplotypes found nowhere else (Illustration by Martin Soave/University of Michigan based on data by (Jakobsson et al. 2008)).

2.6 Genetic Substructure

Humans, as a species, are extremely similar at the genetic level and the variability that does exist is distributed following a pattern of gradients, rather than discrete clusters, of allele frequencies that extend over the entire world (Lao et al. 2008; Novembre et al. 2008; Serre and Paabo 2004). Therefore, there is no reason to assume that major genetic discontinuities or "races" exist between different continents. Isolation by distance is the norm structuring human populations because humans do not mate at random; individuals living in the same geographic region and sharing a language are more likely to mate with each other than with individuals from more distant regions. Although mobility (and therefore admixture) is becoming increasingly common today, historically it has not been the case. This means that, populations have differentiated over time due to genetic drift and other processes.

The classic measure for partitioning genetic variance within and between populations is Sewall Wright's molecular fixation index (F_{ST}), a statistic ranging from a value of zero (no differentiation) to a value of one (no shared genetic variation) (Wright 1969). Estimates of F_{ST} (or equivalent measures) within and between main geographic regions (e.g. Africa, Europe and Asia) typically range from 0.11 to 0.23 for different types of autosomal polymorphisms, indicating that only 11–23% of observed variation is due to differences among populations (Tishkoff and Kidd 2004).

Overall, despite their worldwide distribution, humans show low levels of differentiation (see also **section 4.2.3**.)

As briefly anticipated along this section, several evolutionary processes interact for shaping the observed patterns of genetic diversity, which will be reviewed in more detail next.

3 Evolutionary Forces Shaping Diversity

For evolution to occur there must be a change in allele frequencies from one generation to another. If this is not the case, a population is said to be in Hardy-Weinberg equilibrium, which means that the allele frequencies in one generation can predict the genotype proportions in the next generation, as a result of the absence of processes modifying diversity. In nature, however, real populations are under the operation of one or more evolutionary forces capable of shaping genetic variation.

3.1 Mutation and Recombination

Without mutation, there would be no evolution. Defined as any heritable change in the genetic material of an organism, mutation is the ultimate source of genetic variation. Mutations range from single base changes to small insertions, duplications, and deletions to chromosomal changes such as translocations and the formation of polyploids. Base substitutions, or *point mutations*, occur when there is a change from one base pair to another at a single position in the DNA sequence. Using one DNA strand as a reference, there are 12 possible point mutations as each base can change to any of the three other bases. There are two forms of DNA bases structurally different, purines (G and A) and pyrimidines (C and T). Changes from one purine to another or one pyrimidine to another are referred to as *transitions*, whereas changes from a purine to a pyrimidine or *vice versa* are called *transversions* (see **Figure 11**).

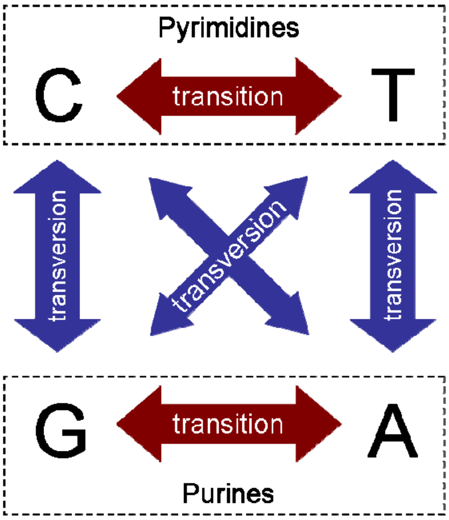


Figure 11: Schematic drawing of the possible transition and transversion mutations.

Transition mutations occur at a much higher rate than transversions, in part because they distort the DNA double helix less than substituting a purine for a pyrimidine or *vice versa*. In contrast, transversions create either purine-purine or pyrimidine-pyrimidine pairs, which go against the base-pairing rules of double-stranded DNA, and thus are more likely to be recognized by the DNA proofreading and repair machinery. In turn, this reduces the rate of transversions compared to transitions. Because of the high fidelity of such DNA repair mechanisms, heritable (i.e., non somatic) mutations are relatively rare for individual nucleotides; however, given the size of the human genome, they are inevitable in every replication cycle of the germ-line. If the probability of mutation per nucleotide pair in humans is 10^{-9} per generation, and the human genome contains approximately 3×10^9 nucleotide pairs, it can be estimated that there would be an average of three new mutations in each human gamete, and therefore, each human zygote would carry six new mutations (Hartl and Clark 2007). In a sense it can be said then that we are all mutants.

New combinations of existing alleles can be generated through the process of genetic recombination, which takes place during meiosis when mixing occurs between different parental chromosomes (see **Figure 12**). Meiotic recombination is a consequence of sexual reproduction, and enhances the ability of populations to adapt to their environment trough the combining of advantageous alleles at different loci. Recombination increases haplotype diversity, and the haplotype structure is therefore affected. Empirical studies in humans and model organisms have revealed that recombination rates are not uniform along a segment of DNA. Chromosomal crossovers appear to be concentrated in *hotspots* between which lie haplotype blocks of low or no recombination in which LD is usually maintained (see **Figure 12**). At larger scales, recombination rates are often low near centromeres and high near telomeres. Fine-scale maps of recombination rates have identified more than 25,000 recombination hotspots across the human genome, together with motifs and sequence contexts that play a role in hotspot activity (Myers et al. 2005).

Both, mutation by creating new alleles and recombination by reshuffling them, are the main processes by which genetic variation is created, providing the raw material on which selection and the other evolutionary forces can act.

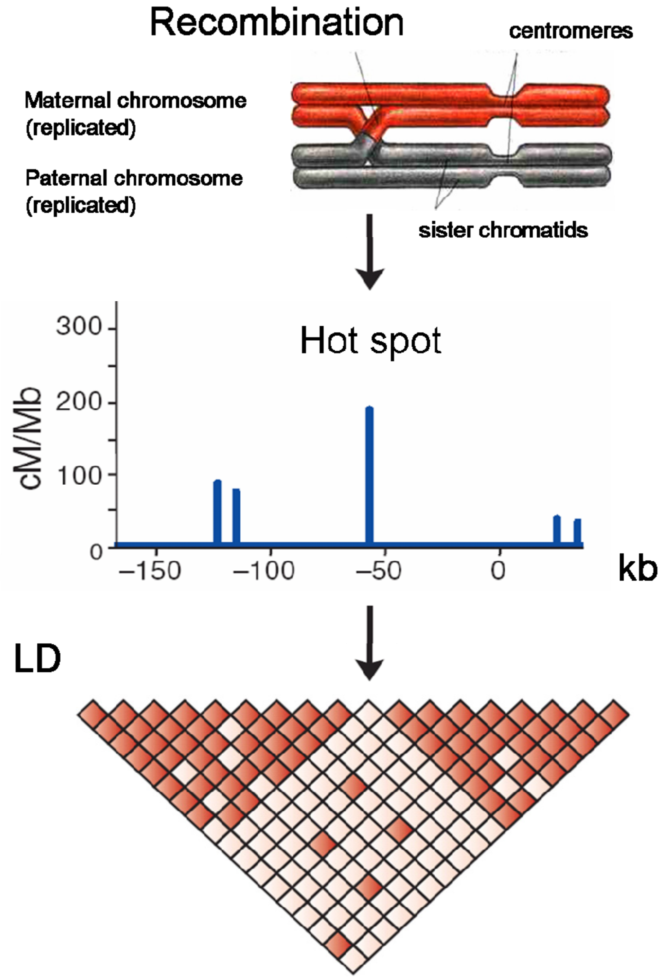


Figure 12: Recombination hotspots and the shaping of haplotype structure in the genome. *Top*: Illustration of a chromosomal crossover during meiosis (adapted from (Alberts et al. 2007). *Middle*: Populational recombination rate, measured in centiMorgan (cM) per Megabase (Mb), over physical distance. *Bottom*: schematic plot of the corresponding pattern of LD. Red squares denote high LD between pairs of markers (modified from (Clark 2005)).

3.2 Genetic Drift

Once mutations arise different outcomes are possible. They can be lost, maintained, or become fixed (when all chromosomes carry the same allele). The expected fate of mutations in natural populations is partly determined by randomness. Because each generation represents a finite sample from the previous one, chance alone can change allele frequency between generations solely through the stochastic process of sampling. This evolutionary process is known as *random genetic drift*.

The magnitude of genetic drift is directly related to the size of the population being sampled. For example, alleles with the same initial frequency (e.g. 0.5) will become either fixed or lost much more rapidly in small populations (e.g. N = 25) than in larger ones (e.g. N = 250 or 2,500), where they will persist over generations with more or less subtle fluctuations in frequency depending on population size. In other words, the smaller the population size, the greater the genetic drift (see **Figure 13**).

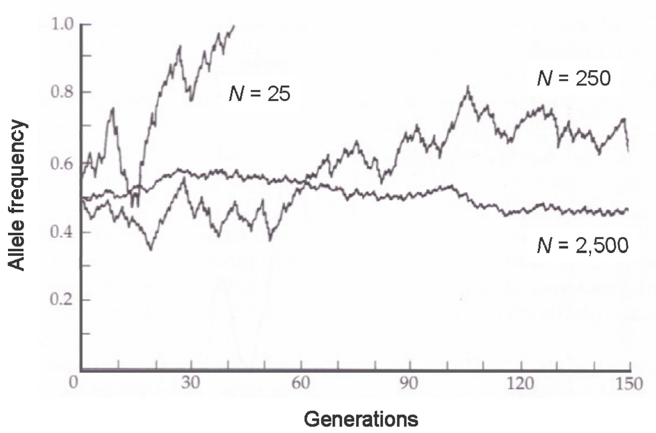


Figure 13: Genetic drift in populations of different sizes (modified from (Graur and Li 2000)).

It is important to make a distinction between the census size of a population (N) and the effective population size (N_e), which is almost always substantially less than the former and it is the actual quotient used to measure the magnitude of genetic drift. For example, the actual size of the human population is approximately 6.5 billion people, whereas the effective population size is only about 10,000 breeding individuals (Harpending et al. 1998; Kaessmann et al. 1999). Although N_e estimates vary among different parts of the genome, this sharp contrast is another reflection of the reduced diversity of our species (Chimpanzees' N_e, for instance, is about 35,000 (Kaessmann et al. 2001)). This is partly due to a recent expansion from a founder population in humans.

Demographic processes, such as *population bottlenecks* and *founder effects* result in a reduced ancestral population size which affects present-day variation. Founder effects imply the genetic separation of a subset of the diversity present within the source population and bottlenecks refer to the reduction in size of a single, previously larger, population with the consequent loss of prior diversity (see **Figure 14**). Since founder effects relate to the process of colonization, human genetic diversity patterns, as discussed in section 2.5, have been strongly shaped by drift as modern humans colonized the world, not to mention the effects of the bottleneck associated with the out-of-Africa expansion. This means that, in the recently expanded human population, the effective population size is still largely determined by the smaller ancestral population sizes in our past.

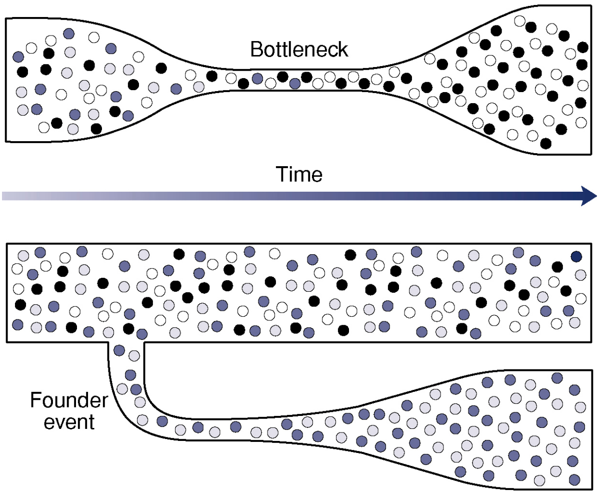


Figure 14: Bottlenecks and founder events. Circles of different colors represent different alleles. Both events result in a loss of genetic diversity due to reduced population sizes in the past (adapted from (Jobling et al. 2004)).

By far the most likely outcome for a new allele is that it will be lost by drift, as opposite to becoming fixed, which is a rare event. But population size directly affects the probability and rate of fixation. The fixation probability of an allele in the absence of selection is equal to its frequency in the population (a new allele would have a frequency of 1/2N). Thus the smaller the population, the higher the probability of fixation and the faster a new mutant becomes fixed. These concepts will be discussed in relation to selection in **section 3.4**.

As opposite to mutation, genetic drift eliminates diversity. Consequently, in a population where no other forces are acting, both mutation and drift will reach an equilibrium in which the number of novel variants (generated by mutation) is balanced by the number of lost variants (eliminated by drift). Such *mutation-drift equilibrium* leads to a stable diversity level in the population (referred to as the *neutral parameter*, θ or "theta"). Nonetheless, as discussed next, other forces such as migration and selection are capable of changing allele frequencies and thus modifying diversity.

3.3 Migration

Most populations are spread over large areas and can be divided into smaller subpopulations within which individuals tend to mate as they share the same local habitat. When such *population structure* exists, there is some genetic differentiation among the subpopulations, meaning that the allele frequencies between them are different (measurable by several fixation indices or *F*-statistics, such as F_{ST} discussed in **section 2.6**). But subpopulations are rarely completely isolated. If there is movement of individuals from one place to another (i.e., migration) and they leave descendents, *gene flow* is said to occur.

Such movement of genes from place to place has a simple effect: it makes different parts of a population more similar to each other. Acting alone, gene flow will eventually homogenize the population. In the presence of random genetic drift, it limits how much genetic divergence can take place. As opposing forces, migration and drift can reach an equilibrium state where differentiation among subpopulations remains constant over time.

To illustrate this, imagine a set of small populations or *demes*, which all start with the same allele frequencies. If these are isolated from each other, then they will drift apart until eventually different demes will be fixed for different alleles. Now, if each deme receives a fraction of immigrant genes (*m*) from some other deme in every generation, the exchange of genes will tend to make the demes more similar and will balance the diversifying effects of random drift (see **Figure 15**).

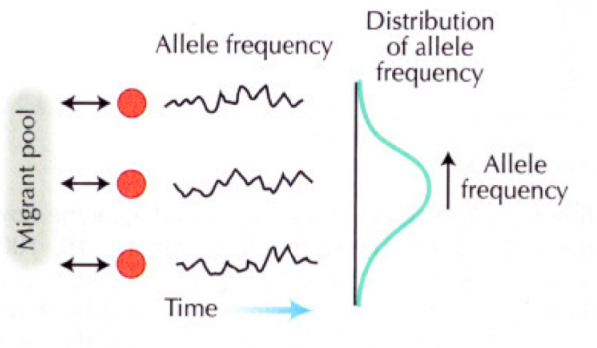


Figure 15: Migration-drift equilibrium. With the exchange of genes with a pool of migrants, allele frequencies in individual demes (red circles) fluctuate at random over time, but the population as a whole reaches a statistical equilibrium (adapted from (Barton et al. 2007)).

Another important effect of gene flow is that it can create linkage disequilibrium by the mixing of populations. In a mixed population, there will be an excess of haplotypes considered as characteristic of the various parental populations, which will be reflected in the increase of LD. Depending on the rate of migration, gene flow can be a potent source of LD.

Clearly, the spatial distribution of humans all over the world is structured in subpopulations, with different barriers to movement and with varying densities and environmental conditions. Such global pattern of a structured human population stirs up interesting questions like whether populations adapt to local environments or how do favorable alleles spread over wide populated areas. These questions involve selection, the core concept of this work, which will be discussed next.

3.4 Natural Selection

Natural selection is the only process that leads to adaptation. The way it shapes diversity is actually manipulating it. When inherited variants cause the organisms to differ in their ability to survive and reproduce (i.e., fitness), there is a generational change in which those individuals with the fittest variants will tend to leave more offspring and, consequently, those variants will tend to increase in frequency. In this way the population becomes progressively better adapted to the environment.

First postulated by Darwin in *The Origin of Species* in 1859, natural selection was not widely recognized as the key evolutionary process until almost a century later. Today, 150 years later, it forms the basis of modern evolutionary theory. Nonetheless, Darwin's concept of natural selection has been made more formal and quantitative; as well as it has been incorporated into models describing the change in allele frequency under selection.

Models of selection compare the relative fitness of a genotype with that of other genotypes competing for the same resources. Mutations that reduce the fitness of the carrier are subject to *negative selection*, also known as purifying selection since they tend to be removed from the population. Alternatively, mutations that increase fitness undergo *positive selection*, and therefore tend to increase rapidly in frequency. Because both processes shift the overall makeup of the population, either by favoring one allele over another or acting against unfavorable or deleterious mutations, they belong to the model of *directional selection* (see **Figure 16**).

Alternative models consider the interaction between alleles to determine the impact of mutations on the fitness of the genotypes (consider for example QQ, PQ, PP). Thus, situations like *overdominance*, *underdominance*, and *codominance*, refer to when the advantage is conferred to the heterozygote, the two homozygotes or one homozygote, respectively. When selection acts on quantitative traits, overdominance can be understood as selection that favors intermediate trait values; this acts to reduce variation and thus is called *stabilizing selection*. As opposite to this, underdominance would be related to *diversifying selection*, which favors extreme phenotypes (i.e. traits in the tails of the distribution) and consequently tends to increase variation (see **Figure 16**).

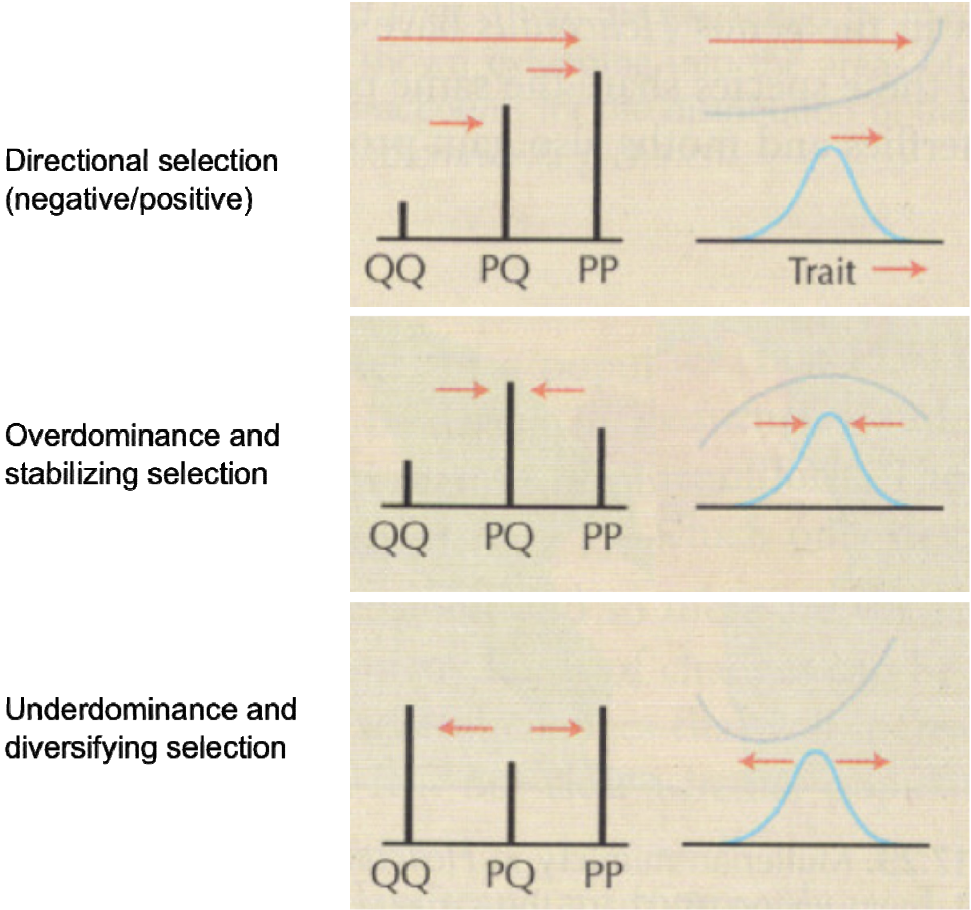

Directional selection (negative/positive)
QQ
PQ
PP
Trait
Overdominance and stabilizing selection
QQ
PQ
PP
Underdominance and diversifying selection
QQ
PQ
PP

Figure 16: Different models of selection. In each model, arrows point towards the favored genotype or trait respectively and gray curved lines follow the level of fitness across each distribution. With discrete alleles, each extra copy of the favored allele increases fitness, and with quantitative traits, fitness increases steadily as the favored trait increases (modified from (Barton et al. 2007)).

In particular, overdominance creates a balanced polymorphism, although this is not the only mechanism by which balanced polymorphisms can be generated. Alternatively, *Frequency-dependent selection* whereby the frequency of a genotype determines its fitness (i.e. rare alleles are favored), also does. Both models then belong to what is referred to as *balancing selection*, which maintains polymorphism in the population.

A well documented case of balancing selection in humans is sickle-cell anemia, which is prevalent in many populations at risk for malaria infection (Allison 1954). The disease is caused by an allele of the *Hemoglobin-B* gene (*HBB*) that codes for an altered form of hemoglobin (Hb^S), which confers an abnormal curved shape to red blood cells. Homozygous individuals for Hb^S suffer from severe anemia and usually do not survive. Despite this, the Hb^S allele is maintained at a relatively high frequency in areas where malaria is endemic, such as Africa, the Middle East, and India, because heterozygous individuals (Hb^S/Hb^A) have only a mild form of the anemia but are quite resistant to malaria. Homozygous for Hb^A are not anemic but, on the other hand, are the most vulnerable to severe malaria. This observation strongly supports that the *HBB* gene has been the target of selection for malaria resistance (Currat et al. 2002; Ohashi et al. 2004).

Malaria has been acting as a selective pressure on the human genome over the last ~10,000 years, so that other genes, like *G6PD*, show evidence for selection mediated by malaria. This and other examples of selection in the human genome will be discussed in section 5.

The fate of selected genes depends to a great extent on genomic context. When mutations are undergoing selective fixation, they tend to drag flanking variation with them through a process of *genetic hitchhiking* (Smith and Haigh 1974). The hitchhiking effect occurs for a simple reason: if two alleles at different loci are in linkage disequilibrium, then positive selection for one of them will cause both to increase. This results in a reduction in the genetic variation for a region surrounding the selected target. This phenomenon, known as *selective sweep*, leaves several characteristic molecular signatures whose recognition is essential for detecting selective events in the genome (see **Figure 17** and **section 4**).

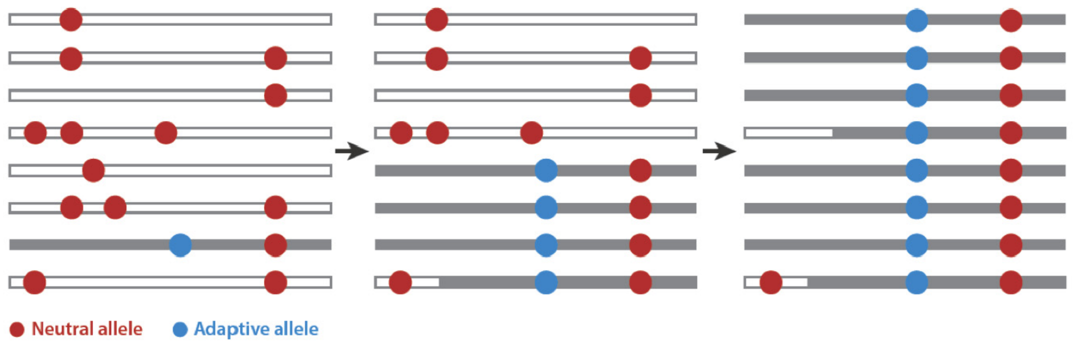


Figure 17: Schematic representation of a selective sweep with recombination. Horizontal bars represent chromosomal regions with neutrally segregating alleles (red dots); the dark chromosome is the one carrying the adaptive allele (blue dot). During the sweep, neutral alleles closely linked to the beneficial allele also rise in frequency (*middle*). After a complete sweep, all variation in the region has been eroded except for the new mutations subsequently arising (*right*) (adapted from (Kelley and Swanson 2008)).

In order to contextualize selection even more, two additional considerations are worth to mention. First, that selection acts on the phenotype, not on the genotype, and the total phenotype is determined by many genes that interact with each other as well as with numerous environmental factors. Second, that selection does not act alone; there is a complex interplay among the different evolutionary forces. It has been discussed that mutation, recombination and migration increase diversity, random genetic drift decreases it, and selection can do either of them.

Now, some of these mechanisms are opposing forces and tend to balance each other, so diversity can eventually reach equilibrium. Interesting patterns then will be seen when the interplay of these forces is able to depart from neutrality.

In 1968 Motoo Kimura proposed that most polymorphisms observed at the molecular level are selectively neutral, so that they are maintained in the population by a balance between mutation and random genetic drift. The frequencies of neutral alleles are not, therefore, determined by natural selection. Consequently, many polymorphisms may have no particular significance in the adaptation of the organism. These principles form the basis of the *neutral theory* of molecular evolution (Kimura 1968; King and Jukes 1969). This theory was developed on the basis of protein polymorphism data, but the model applies to nucleotide sequence data as well.

With the recent flood of DNA sequence availability it has been possible to statistically test the neutral theory against real data, and consequently, numerous evidences that selection is also importantly acting started to deluge the scientific literature. The neutral theory, therefore, serves as a well-understood null hypothesis, and deviations from it may reflect the action of selection.

4 Methods for Detecting Selection

How to detect the action of natural selection? How to discover species-specific changes or even local adaptations of populations? It is all about time. Natural selection has been acting in nature for millions of years and its direct observation is mostly not possible, so that the action of selection has to be mainly inferred.

In humans, this has triggered a special interest as detecting selective events may allow us to identify the functional traits –and the genes underlying them– that set us apart from other species, or those that differentiate populations around the world.

Selection tests then seek to identify genes or genomic regions that have responded to the selection pressures acting in past populations that ultimately made us who we are today. This process, depending on when and how strongly occurred, leaves different signatures in the genome, which we can recognize at the molecular level by analyzing extant patterns of variation.

There are different ways to classify the growing list of methods for detecting selection. Many of them compare the observed diversity to that expected under neutral evolution, so that they are referred to as *neutrality tests*. This can be done by using differences *between* species (**interspecific divergence**) or by using polymorphism *within* species (**intraspecific diversity**). Some other methods compare variation both within and between species. This classification (based on whether the tests look at divergence, diversity or both) will be followed as different signatures of selection are being described. Other characteristics, like the different time scales over which they persist or whether they are detectable using DNA sequence data or SNP data, will be discussed along their description.

4.1 Signatures from Interspecific Divergence

Although only about 1.5% of the mammalian genome codes for proteins (Lander et al. 2001; Waterston et al. 2002), protein-coding regions are of obvious importance for the organism (see **section 2.1**). They tend to be highly conserved (i.e., not variable) across species, and thus, homologous sequences from multiple species (i.e., orthologous) can be aligned together enabling the identification of species-specific genetic changes; some of which, interestingly, may be of functional relevance (Carroll 2003; Olson and Varki 2003).

In the case of our species, a human-specific change is that one that took place at some point on the human evolutionary lineage and that is present in all currently living humans (i.e., that it is fixed among humans). By analyzing one human and one chimpanzee DNA sequence, as well as one or preferably more *outgroups* (i.e., species that are known to diverge earlier than those being studied), one can establish that a change occurred on the human lineage (see **Figure 18**). In order to establish that it is fixed among humans, several individuals from various regions of the world may be needed. However, given the divergence time of ~6 million years between humans and chimpanzees, chances are that a genetic change is specific to humans in more than 80% of the cases if only one human is analyzed (Enard and Paabo 2004). This probability increases drastically if the change is observed in several humans (e.g. as few as 10 sampled chromosomes can increase to 96.4% the probability of being fixed) (see **Figure 18**).

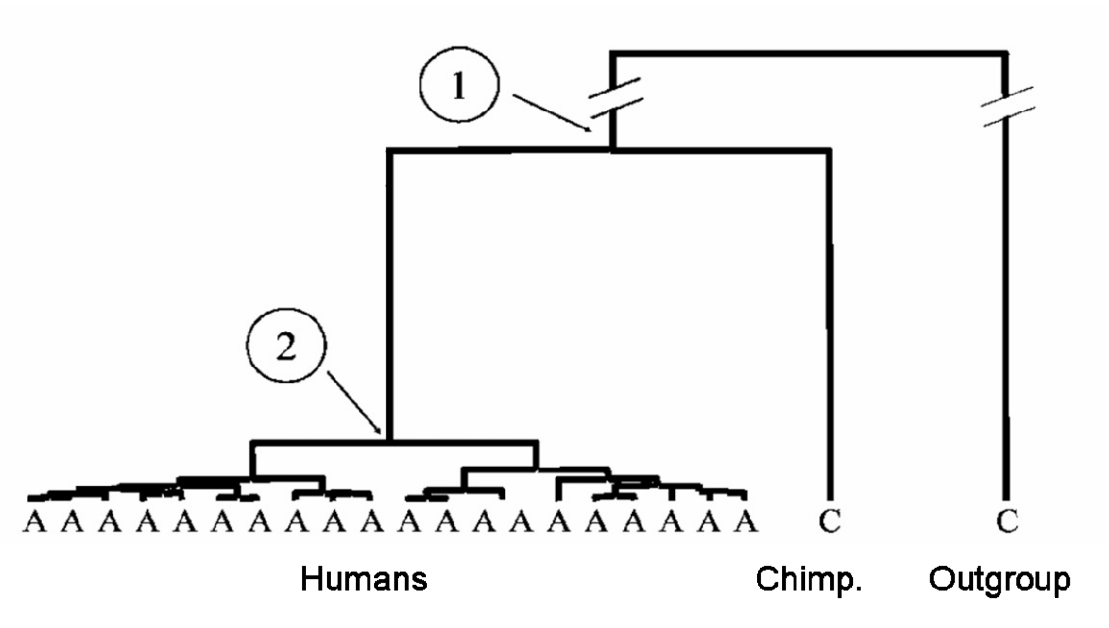


Figure 18: Genealogy of a human-specific change as expected under the neutral model. A genetic change occurring in the phylogenetic branch leading to humans (between points 1 and 2) can be defined as human-specific. Here, a nucleotide substitution from C (ancestral) to A (derived) is depicted (adapted from (Enard and Paabo 2004)).

Now, to test whether a genetic change has been driven to fixation by positive selection in a given lineage, several hypotheses have to be examined. When comparing DNA sequences between species, positive selection can be inferred by testing whether more functional changes occurred in a putatively positively selected region than expected from the neutral mutation rate. The most common method following this approach is the K_a/K_s test (Kimura 1983) which evaluates the proportion of different nucleotide substitutions (i.e., synonymous versus nonsynonymous) in comparison with other lineages. This method allows the pinpointing of specific genes undergoing accelerated protein evolution in specific lineages, which is one of the strongest signatures of positive selection.

4.1.1 High proportion of function-altering mutations

As opposite to silent (or synonymous) changes, nucleotide substitutions that alter protein function (e.g. nonsynonymous) are usually deleterious, or at least disadvantageous, and thus less likely to become fixed. Over a long period of time, however, the action of positive selection can increase the fixation rate of beneficial function-altering mutations, leading to an accelerated accumulation of amino acid substitutions on the targeted protein.

The K_a/K_s ratio (sometimes denoted d_N/d_S or ω in different contexts) is the ratio between the rate of nonsynonymous substitutions (K_a or d_N) and the rate of synonymous substitutions (K_s or d_S). Each rate is calculated per nucleotide site and so, in the absence of any selective pressure, the ratio is expected to equal 1. When genes are constrained by purifying selection they accumulate a higher proportion of silent changes, so K_a/K_s is smaller than 1. However, if the action of positive selection has favored several amino acid replacements between closely related species, then K_a/K_s can be larger than 1. However, the extent of positive selection can be greatly underestimated by simply counting cases where $K_a > K_s$ over the entire gene. The gene as a whole may be under strong selective constraint but a particular region or domain may have evolved under positive selection. Therefore, it is mostly more convenient to use a model in which rates of amino acid evolution (as measured by ω) vary from codon to codon. Such models are referred to as *site models* and those allowing variation in ω among lineages and sites at the same time are known as *branch-site models* (Yang and Nielsen 2002; Zhang et al. 2005). The latter is thought to be one of the most powerful approaches to detect rapidly evolving sites within protein-coding sequences of the genome.

Although similar tests can be applied to noncoding sequences, such as the ζ parameter (Wong and Nielsen 2004) or the K_s/K_i ratio (Chamary et al. 2006) where K_i refers to the intronic substitution rate, they fall outside the scope of this work and therefore will not be discussed in detail.

The signature detected by the K_a/K_s test persists over a large range of evolutionary time scales (i.e., many millions of years), as it results from the total divergence accumulated in each sequence since splitting from their common ancestors. When applied to human sequences and closely related species, this method can reveal the possible genes that played an important role during the early stages of our evolution (see **Figure 19**).

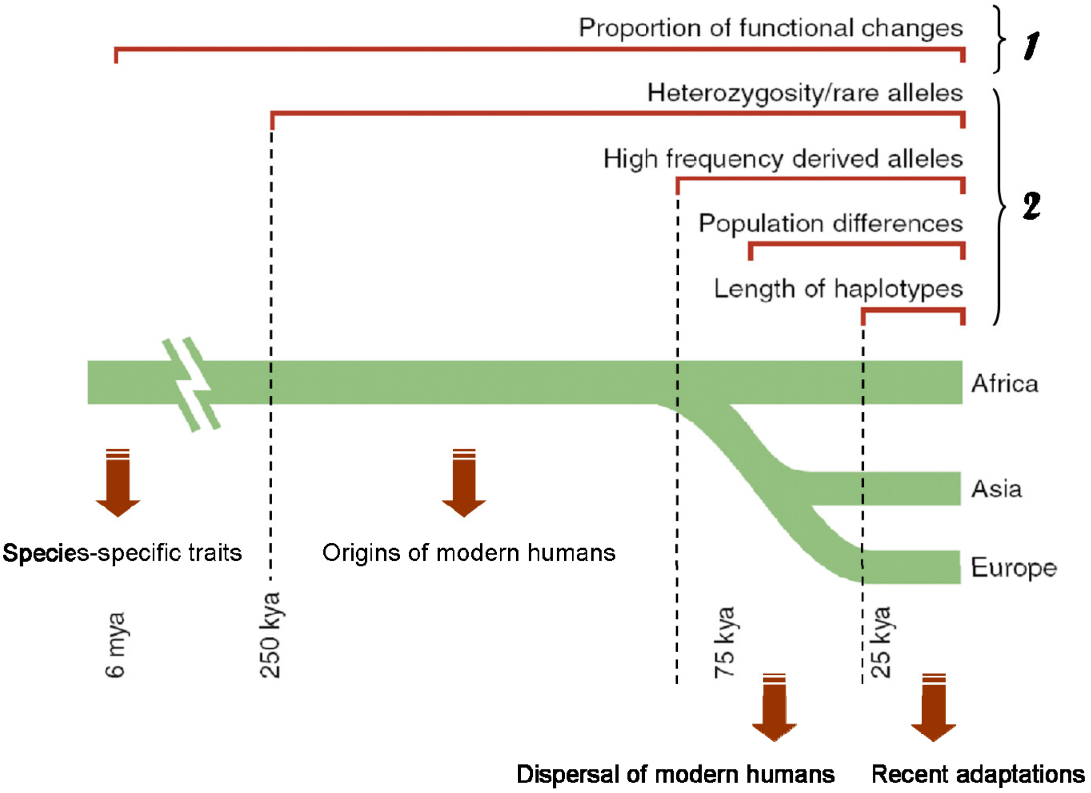


Figure 19: Schematic representation of the time scales over which different signatures of selection persist in the human genome. Different adaptive events involving crucial episodes of our evolutionary history (arrows) can be detected depending on the observed signature. Number (1) indicates signatures based on interspecific divergence, and number (2) those based on intraspecific diversity (adapted from (Sabeti et al. 2006)).

4.2 Signatures from Intraspecific Diversity

Different patterns of variation within species can reveal deviations from the neutral theory, and so indicate signatures of selection. In this approach, instead of sampling just one individual from each of several species, a large sample of individuals is collected from a population of a single species. Moreover, unlike the K_a/K_s test, methods based on intraspecific diversity are not robust to demographic factors, as within-species variation is strongly affected by population history (e.g. population subdivision or population bottlenecks), giving rise to patterns compatible with both the action of selection and that of demography.

For example population bottlenecks, in a similar way to a selective sweep (see **section 3.4**), lead to a drastic reduction of genetic variation. This is because individuals will tend to share recent ancestry as many lineages trace back to coalesce in the short time when the population was small. Given their similar effects on variation, a selective sweep and a population bottleneck could be undistinguishable using data from a single locus. However, in the case of a selective sweep, such reduction in genetic variation should be localized to a particular genomic region (around the favorable mutation that is undergoing fixation); whereas, in contrast, a population bottleneck is expected to affect all parts of the genome in the same way. Therefore, we can distinguish the action of selection from demographic events by comparing patterns of variation across many genes.

Both selection and demography shape the frequency distribution of polymorphic nucleotide sites within the population, also known as the *site frequency spectrum*, in which we can look for several interesting skewed distributions. Selection against deleterious mutations (i.e., negative selection) will increase the fraction of mutations segregating at low frequencies in the sample. Conversely, positive selection will tend to increase the fraction of mutations segregating at high frequencies but at the same time, especially when a selective sweep has been created, it will also increase the fraction of rare alleles. Both selective sweeps and negative selection have roughly the same effect in the leftmost part of the spectrum (i.e., an excess of rare alleles) but in the latter there is an absence of high-frequency alleles (see **Figure 20**).

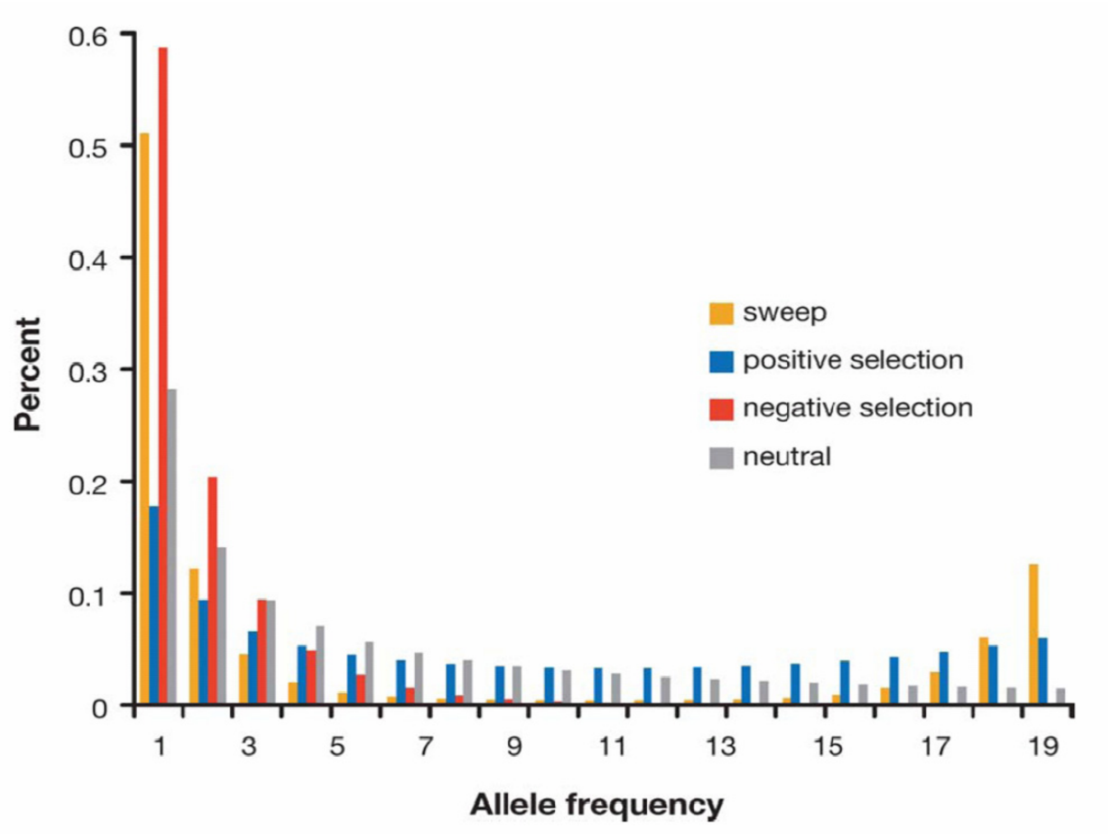


Figure 20: Summary of the effect of selection on the frequency spectrum. Note that both negative selection and a selective sweep can have a similar effect on the leftmost frequency classes (i.e. an excess of rare alleles). The former, however, does not allow mutations to segregate at high frequencies (reflected by the absence of mutations in the rightmost frequency classes) (adapted from (Nielsen 2005); spectra are based on data presented therein).

4.2.1 Excess of rare alleles

In section 3.4 it was described how the neighboring region of a beneficial mutation favored by selection is affected by the so-called hitchhiking effect. Therefore, a selective sweep not only causes the selected variant to rise in frequency but also brings closely linked variants with it (see **Figure 17**). This eliminates diversity in the vicinity with a variable extent that obviously depends on the strength of selection and the level of LD in the region. After that, new mutations tend to restore diversity, but are initially at low frequency. Overall, the signature consists of a region of low genetic diversity, with an excess of rare alleles.

Usually, methods detecting this kind of signature are based on DNA sequences, as this type of data allows capturing the whole spectrum of variation; although SNP data can also be exploited in the search of these signals (e.g. MAF analysis; see below). Commonly used statistics include Tajima's *D* (Tajima 1989) and related tests such as Fu and Li's *D* and *D** (Fu and Li 1993), as well as Fu and Li's *F* and *F** (Fu 1996, 1997).

Tajima's *D* statistic compares two estimates of θ (see **section 3.2**), based on the number of segregating sites (S) and nucleotide diversity (π), respectively. Since different estimates of θ should be equal under neutrality, Tajima's *D* is expected to be zero. Significant positive values indicate balancing selection or population subdivision, whereas negative values indicate positive selection or population expansion.

As mentioned above, the extent of reduced genetic diversity caused by a selective sweep –as reflected in Tajima's *D* or in variable sites (S) for instance–, correlates with the extent of LD in the region. As LD is increased around the beneficial allele, both Tajima's *D* and S remain low; whereas variability is restored as LD decreases further away from the location of the sweep (see **Figure 21**).

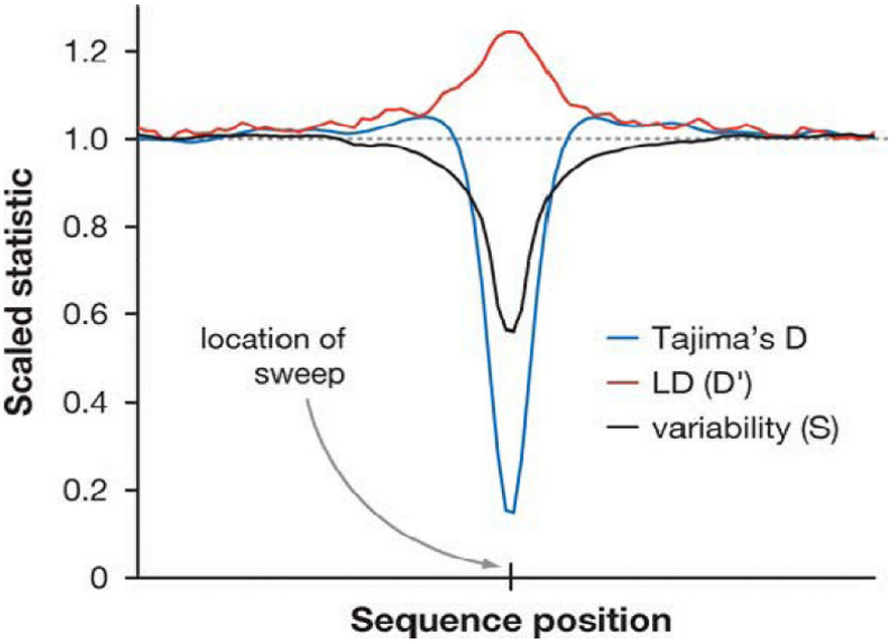


Figure 21: Illustration of the effect of a selective sweep on genetic variation. All statistics are based on simulated data and scaled so that the expected value under neutrality equals one (see text for details) (adapted from (Nielsen 2005)).

Because mutation is rare and new mutations take time to drift to high frequencies under neutral evolution (~1 million years in the human population), it is estimated that reduction in genetic diversity can persist in the human genome for up to ~250,000 years (Sabeti et al. 2006). It is actually the signature with the oldest time scale among those detectable from intraspecific data (see **Figure 19**). This is particularly important because such a time scale comprises, with no doubt, the origins of modern humans.

4.2.2 High-frequency derived alleles

Different related methods take into account the ancestral state of genetic variants. When a particular polymorphism exists, for instance in humans, is rarely expected to have occurred in other primates at the same base. Therefore, the ancestral and derived states of a SNP can be determined by looking at the genome of closely related species, assuming that the mutation occurred after the two species diverged. In humans, this task has been facilitated by the availability of the chimpanzee genome sequence (Consortium 2005a) and the growing data from additional primate genomes (see **Figure 22**).

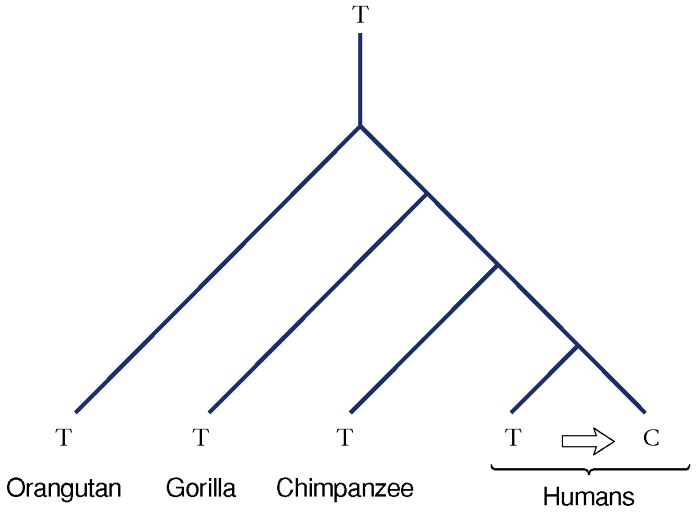


Figure 22: Ancestral state determination of a SNP. The ancestral allele is inferred from the allele present in closely related species. Here, a T (ancestral) to C (derived) transition is assumed to underlie the T/C polymorphism observed in humans (adapted from (Jobling et al. 2004)).

Since they are younger, derived alleles under neutrality are usually expected to be present at lower frequencies than ancestral alleles. However, in a selective sweep scenario, any derived allele linked to a beneficial allele can hitchhike to high frequency; creating a signature of a region with an excess of high-frequency derived alleles. This can be formally tested with Fay and Wu's *H* statistic (Fay and Wu 2000), which applies to DNA sequence data.

For SNP data there is an alternative approach that, although not considered a formal neutrality test, may capture this signature of selection as it explores the proportion of high-frequency derived alleles above a certain threshold (usually 0.8) within a given genomic region. Similarly, this approach can also be applied to detect an excess of rare alleles (see **section 4.2.1**) by exploring the proportion of low-frequency minor alleles below a certain threshold (usually 0.1) along physical distance. In this work, these approaches are referred to as Derived Allele Frequency (DAF), and Minor Allele Frequency (MAF) threshold analyses, respectively.

The derived-alleles signature persists for a shorter period than the rare-allele signature discussed above because high-frequency derived alleles rapidly drift to fixation. This allows therefore the detection of selective events occurring in the last ~80,000 years of human history, predating the dispersal of modern humans (see **Figure 19**).

4.2.3 Differences between populations

Allele frequency variation between populations is largely determined by random genetic drift (see **section 3.2**). However, when a locus is subjected to positive selection in a geographically restricted population, the allele frequencies around the selected locus can change rapidly, leading to a high degree of population differentiation in the genome region subjected to selection (Lewontin and Krakauer 1973). Therefore, relatively large differences in allele frequencies between populations (at the selected allele itself or in surrounding variation) may reflect the action of local positive selection.

Commonly used statistics to detect this kind of signature include the F_{ST} fixation index, which is the proportion of the total genetic variance explained by the differences among populations (see **section 2.6**). F_{ST} can be calculated for haplotypes (when sequence data is available or genotype data is phased), as well as for single markers (e.g. SNPs) by means of a *locus by locus* analysis of molecular variance (AMOVA). Empirical studies in humans (Barbujani et al. 1997; Jorde et al. 2000) have demonstrated that most of the autosomal genetic variation (i.e., ~85%) occurs between individuals within the same population, and that smaller fractions of variation are due to differences between populations within the same continent (i.e., ~5%) and between different continents (i.e., ~10%). Overall, this apportionment of human variation results, on average, in a worldwide F_{ST} value of approximately 0.15 (see **Figure 23**), which has become an informal but useful reference to predict possible values reflecting unexpected levels of population differentiation.

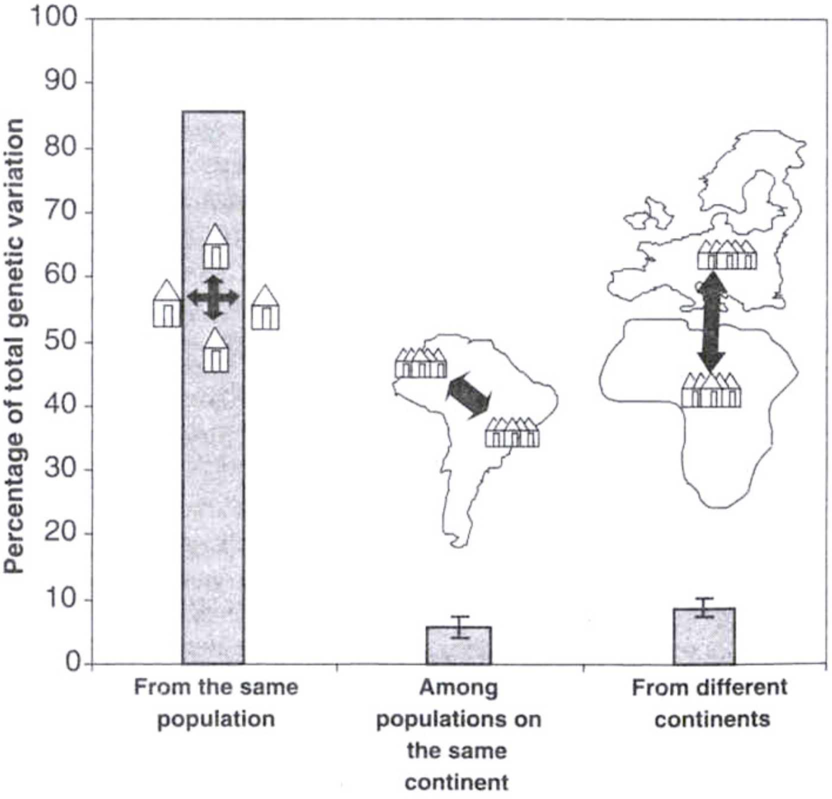


Figure 23: Apportionment of human genetic variation. The sum of the variation between populations from the same (~5%) and different (~10%) continents is roughly ~15%, which equals the expected worldwide F_{ST} value for human autosomes. Results from mtDNA and the Y chromosome may differ (adapted from (Boyd and Silk 2004)).

A statistical test of neutrality could be constructed by comparing the value of the F_{ST} statistic at a candidate locus with the distribution of values expected for a null neutral model of subdivided populations. In humans, however, is not easy to provide an appropriate null model of population structure; so that an empirical approach is usually taken in which the value of the F_{ST} statistic at the candidate locus is compared with the distribution of the statistic observed for a large set of independent loci.

Population differentiation mainly arises when populations are at least partially isolated reproductively. Hence, signatures in the human genome based on differences between populations, are likely to correspond to events that occurred after the major migrations out of Africa some 50,000 to 75,000 years ago (see **Figure 19**).

4.2.4 Long unbroken haplotypes

When a mutation appears, it arises on an existing background haplotype characterized by complete LD between the new mutation and all linked polymorphisms. Over time, new mutations and recombination reduce the size of this haplotype such that, on average, older mutations (which may be either common or rare) will be found on smaller haplotypes (i.e., with short-range LD between the mutation and linked sites). Younger (and thus typically rare) mutations are usually associated with long-range haplotypes. However, beneficial mutations favored by selection can rise in frequency rapidly enough to prevent recombination from breaking down the extension of LD in the region. Therefore, selective sweeps create a signature of an allele showing both high frequency and long-range associations (see **Figure 24**).

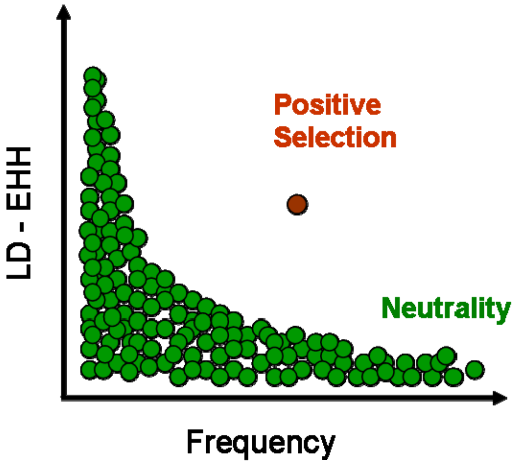


Figure 24: Relation between the extension of LD, as measured by EHH, and frequency. Values from regions under neutrality (green dots) tend to fall within the typical "L" shaped distribution of EHH, allowing values that contradict this tendency to stand out (red dot).

Measures for the extension of LD include the Extended Haplotype Homozygosity (EHH) and its many related statistics. One of the first and most popular statistical tests implementing EHH was introduced by Sabeti et al. (2002), which is known as the long-range haplotype (LRH) test. This test interrogates a particular *core* region and computes EHH at increasing distances from both sides of the core, at which EHH is expected to decrease more or less gradually depending on the LD extension of each haplotype (see **Figure 25**). It is the comparison of EHH, at a particular distance from the core, between the different haplotypes associated with the same core that results in relative EHH (REHH), which in turn is corrected for the frequency of the haplotype in the population. By analyzing many core haplotypes a distribution of REHH versus frequency can be obtained and contrasted against a background (empirical or simulated) distribution, so that statistical significance can be tested for outlier haplotypes (see **Figure 24**). An extension of the LRH test was introduced by Voight et al. (2006) with the integrated haplotype score (iHS) which basically integrates EHH under the curve of LD decay, eliminating the ambiguity of interrogating haplotype cores at particular distances.

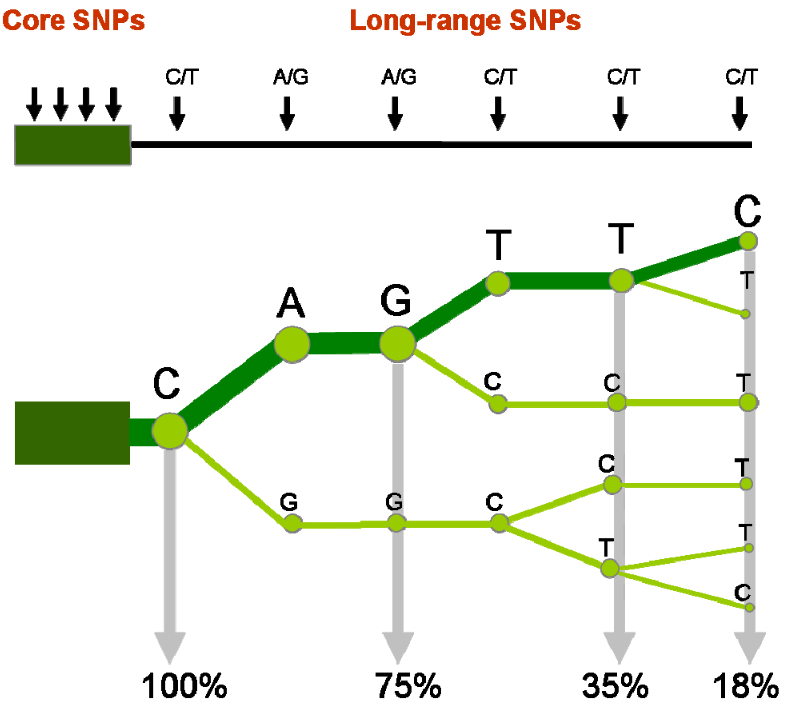


Figure 25: Cartoon of a one-sided bifurcation plot of core haplotypes. Each node (SNPs) represents an opportunity for haplotypes to bifurcate; EHH thus is the probability, at a given distance, that two random chosen chromosomes carrying the core haplotype of interest are identical by descent for the entire interval from the core region. Example values are shown for intuitive orientation only. Illustration originally crafted by Martin Sikora/Pompeu Fabra University, and slightly modified herein.

Both the LRH test and the iHS statistic are based on the comparison of EHH between alleles within a population, and thus, are powerful methods for identifying alleles that have been driven to intermediate frequencies by positive selection. However, such intrapopulation comparison has low power when the selected allele is at high frequency (see **Figure 26**), and becomes impossible when the selected allele is fixed. Therefore, this approach lacks power to detect selective sweeps that have resulted in near or complete fixation of an allele in a population, and hence may fail to detect a significant fraction of loci that have experienced local positive selection.

Overcoming this caveat, new alternative EHH-based statistics, such as XP-EHH (Sabeti et al. 2007) and Rsb (Tang et al. 2007), compare the decay of EHH of an individual SNP site, rather than that of an allele, between populations. For instance, Rsb also integrates the area under the curve of the LD decay into a single integrated EHH (iEHH). Then, the iEHH of a particular site is compared between different populations to obtain Rsb (relative iEHH of a site between populations). Extreme values suggest that positive selection has differentially favored one particular population increasing its iEHH at a given locus. Again, statistical significance can be assigned by comparing the observed values with a, preferentially genome-wide, background distribution of the statistic, allowing to control for the genomic EHH pattern of each population. This approach is particularly sensitive to detect fixed selected variants or partial selective sweeps near fixation (see **Figure 26**).

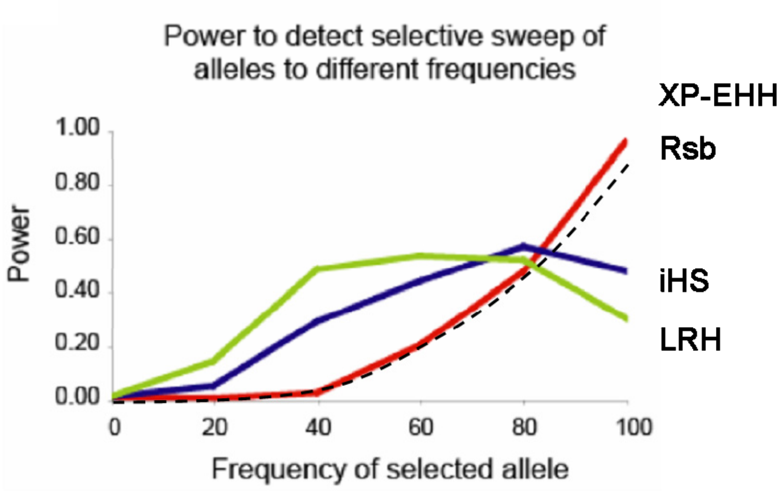
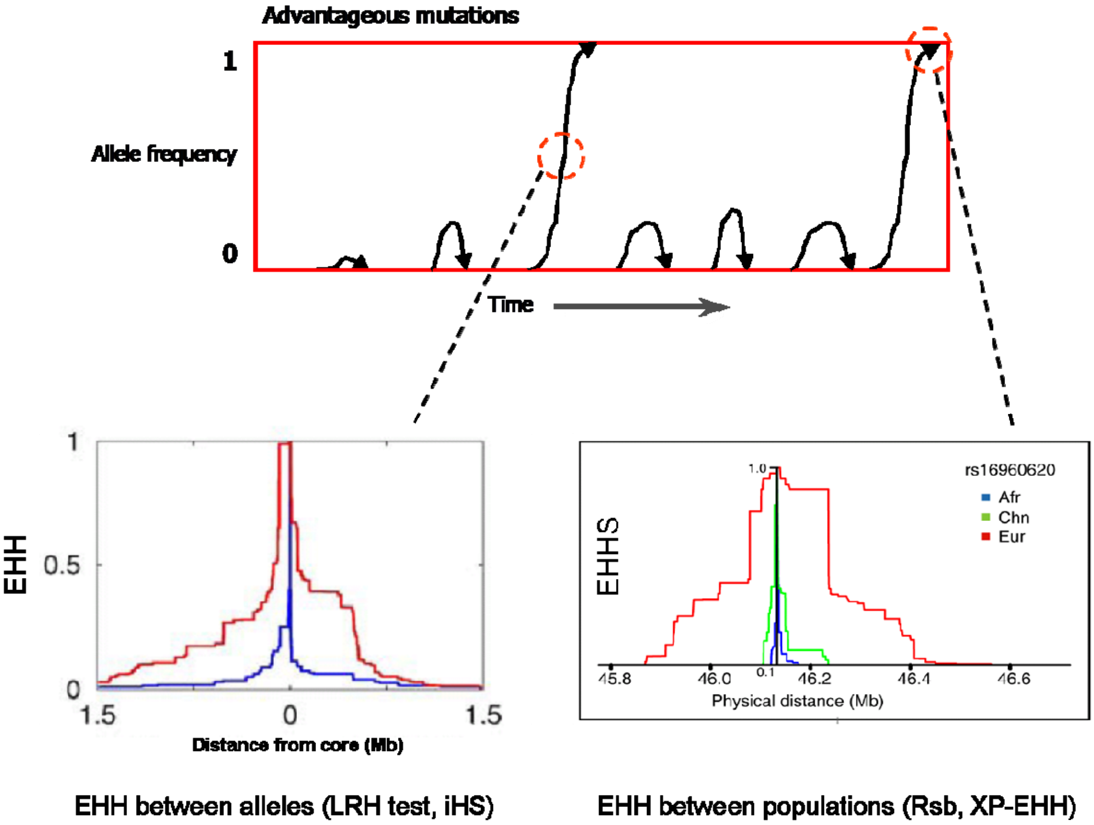


Figure 26: Estimated power for different EHH-based statistics. Rsb (not directly tested) is analogous to the XP-EHH statistic, so that it is expected to follow a similar pattern (dashed line) (adapted from (Sabeti et al. 2007)).

Capturing the complete decay of LD over distance usually requires the analysis of long stretches of DNA, which currently are more effectively and less costly covered with SNPs. Thus, selection tests designed to detect long-range haplotypes are mainly based on SNP data. Actually, an interrogated core region can consist of a single SNP (like the example in **Figure 27**), allowing the investigation of many markers over a densely covered candidate region.

Since we ignore beforehand the allele frequency of the positively selected variant, if any, an optimal approach is one in which different available statistics are combined in order to fully grasp the possible stages of selective sweeps imprinted in our genomes (see **Figure 27**).



EHH between alleles (LRH test, iHS)

EHH between populations (Rsb, XP-EHH)

Figure 27: Complementary EHH-based approaches to detect selected variants at different frequencies. Advantageous mutations may be captured at different stages of their way to fixation (schematically represented on *top*). When still segregating at intermediate frequencies they can be detected by statistics that compare EHH between alleles in a population (*bottom left*); whereas when already fixed (or nearly so) they are more likely to be detected by statistics comparing EHH between different populations (*bottom right*). The left panel shows the decay of EHH of the two alleles of a SNP in a European sample (adapted from (Voight et al. 2006)). The right panel shows the decay of EHH per site (EHHS) of a SNP in three different populations (adapted from (Tang et al. 2007)).

Finally, and maybe more important, is to consider that long-range haplotypes persist for relatively short periods of time (up to ~30,000 years), because recombination rapidly breaks down the haplotype. Therefore, tests based on this signature have widespread applicability in humans because local adaptive evolution is thought to have been particularly important over the past ~10,000 years, as humans changed from hunter-gatherers to an agriculture-based subsistence and encountered different pathogens and new environments in different regions of the world (see **Figure 19**).

4.3 Joint Polymorphism and Divergence Tests

So far, different selection tests have been discussed separately for those looking at differences between species and those looking at variation within them. However, methods combining information of both have also been developed.

Under neutrality, diversity within species and divergence between species should be proportional to each other, because both are due to mutation and drift. The neutral theory predicts that the ratio of divergence to polymorphism should be the same for all kinds of change. Thus, if this ratio is higher for nonsynonymous changes than for synonymous changes, it can be inferred that positive selection has established adaptive amino acid differences between the species.

The most widely used test following this approach is the McDonald–Kreitman test (McDonald and Kreitman 1991), which quantifies the comparison of the divergence to polymorphism ratio between different kinds of change. Sites are classed as polymorphic (*P*) or as divergent (*D*) positions correspondingly. Divergent positions that are nonsynonymous and synonymous are denoted D_n and D_s respectively (analogous to K_a and K_s described in **section 4.1.1**). Similarly, polymorphic positions are denoted P_n and P_s as appropriate. If both synonymous and nonsynonymous changes were neutral, the ratio of divergence to polymorphism should be the same for both: $D_n/P_n = D_s/P_s$. However, if $D_n/P_n >> D_s/P_s$, the action of positive selection is assumed. This can be tested through a simple contingency table (see **Table 1**). The McDonald–Kreitman test can be applied systematically across many genes to give an overall estimate of the fraction of divergence caused by selection.

The Hudson-Kreitman-Aguadé (HKA) test (Hudson et al. 1987) also compares within-species polymorphism and between-species divergence, but between two (or more) loci. If they are evolving under the neutral model, the levels of divergence and polymorphism should correlate at both loci. Thus in this case, the null hypothesis of neutrality can be tested at a chosen locus by comparing it with polymorphism and divergence data at a neutral control locus in the same two species (see **Table 1**).

Table 1: Contingency tables for both McDonald-Kretiman and HKA tests.

McDonald-Kreitman (MK) Test [a]

	Fixed	Polymorphic	↔
Synonymous	D_s = **17**	P_s = **42**	Ratio = 0.4
Nonsynonymous	D_n = **7**	P_n = **2**	Ratio = 3.5

Hudson-Kreitman-Aguadé (HKA) Test [b]

	Fixed	Polymorphic	
Locus 1	**18**/324 (5.6%)	**8**/79 (10.1%)	↕
Locus 2	**210**/4052 (5.2%)	**9**/414 (2.18%)	
	Roughly equal	4-fold difference	

[a] Data from McDonald and Kreitman (1991) on the *Adh* (Alcohol dehydrogenase) locus in *Drosophila* species (*D. melanogaster*, *D. simulans*, and the outgroup *D. yakuba*). Fisher's exact test gives a *p* value of 0.0073. Arrow indicates the direction of the ratio.

[b] Data from Hudson *et al.* (1987) on silent variation of the *Adh* locus and a 4-Kb 5' flanking region (locus 1 and 2 respectively) in *D. melanogaster* and *D. sechellia*. The resulting X^2 statistic gives a *p* value of 0.014. Arrow indicates the direction of the comparison of both divergence and polymorphism between locus 1 and 2.

Because it detects the increase of interspecific nonsynonymous substitutions (relative to intraspecific nonsynonymous polymorphisms), the signature of selection captured by the McDonald–Kreitman test has a similar time scale than the K_a and K_s test (i.e. many millions of years). Conversely, the HKA test can detect situations in which, while divergence is maintained roughly equal between two unlinked loci, large differences in polymorphism occur. Such signature is thus comparable to that captured by other statistical tests based on a reduction in genetic diversity, hence persisting over similar time scales (see **Figure 19** and **Table 2**).

Table 2: Some of the commonly used selection tests and the molecular signatures they detect.

Data	Test	Applicable to	Signature of selection	Signature's age
Interspecific Divergence	K_a/K_s (d_N/d_S) ratio tests	Sequence data	High proportion of function-altering mutations	Millions of years
Intraspecific Diversity	Tajima's D Fu and Li's D and D^* Fu and Li's F and F^*	Sequence data	Reduction in genetic diversity / excess of rare alleles	<250,000 years
	MAF threshold analysis	SNP data		
	Fay and Wu's H	Sequence data	Excess of high-frequency derived alleles	<80,000 years
	DAF threshold analysis	SNP data		
	F_{ST} fixation index and other F-statistics	Sequence data or SNP data	Extreme differences between populations	<50,000 to 75,000 years
	LRH test / iHS statistic XP-EHH / Rsb statistic	SNP data	Long-range haplotypes	<30,000 years
Divergence and Diversity	McDonald-Kreitman test	Sequence data	Higher divergence to polymorphism ratio in function-altering sites	Millions of years
	HKA test	Sequence data	Different correlation of polymorphism and divergence between loci	<250,000 years

An important consideration is that methods based on DNA sequence data take advantage of the whole spectrum of allele variation in the sample, whereas SNP-based genotyping studies may be biased depending on the strategy and the population used for discovering such SNPs. A frequency-specific *ascertainment bias* may result from the fact that the SNP discovery panel is often small, so that common SNPs are more likely to be discovered in comparison with rare SNPs (Clark et al. 2005).

Overall, the application of these tests (see **Table 2**) produces a set of candidate genes for further inquiry into the connection between genotypic and phenotypic variation. Many of them, either as individual genes or as a functional category of genes, have already been reported in a number of different species. The following section then reviews some of the main examples described in humans.

5 Evidences of Selection in the Human Genome

After discussing the theory of the different signatures of selection, one obviously wonders which genomic regions, and therefore which traits, have been influenced by natural selection in humans. Both negative and positive selection play an important role for conserving biological functions or allowing new ones to spread, respectively. However, much attention has focused on positive selection because it leaves a clear footprint of evolutionary adaptation at the molecular level.

An increasingly large body of evidence for positive selection in the human genome has been described, either by comparative studies of differences between species or by means of intraspecific studies of variation within humans. This section will provide a very partial list of the many cases reported to date, as it is only aimed at illustrating with a few examples these two approaches and at introducing their connection with the work presented here.

5.1 Comparative Studies

Most comparative studies focus on protein-coding genes, either following a candidate gene approach or screening large sets of coding sequences on different species. This allows the identification of loci undergoing positive selection on the human lineage leading from the ancestor of humans and chimpanzees to modern humans, which may underlie human-specific adaptations.

For instance, cognition and the ability to communicate through language are attributes that set humans apart from other species. The study of a family with severe deficiencies in language skills allowed the identification of a gene (i.e., *FOXP2*) that influences human speech (Fisher et al. 1998). *FOXP2* is a transcription factor that is highly conserved, mutations in which have been associated with speech and language disorders (Lai et al. 2001; MacDermot et al. 2005). Sequence comparison between species revealed only four amino-acid changes in the phylogeny of extant primates and other mammals (Enard et al. 2002), two of which occur on the human lineage (see **Figure 28**). This relative acceleration in the protein evolution of this gene along the human lineage led the authors to suggest the action of positive selection, which they supported with additional evidence (i.e., negative Tajimas'*D* and Fay and Wu's *H* values) compatible with a selective sweep occurring during the last ~200,000 years, that is, around the emergence of anatomically modern humans.

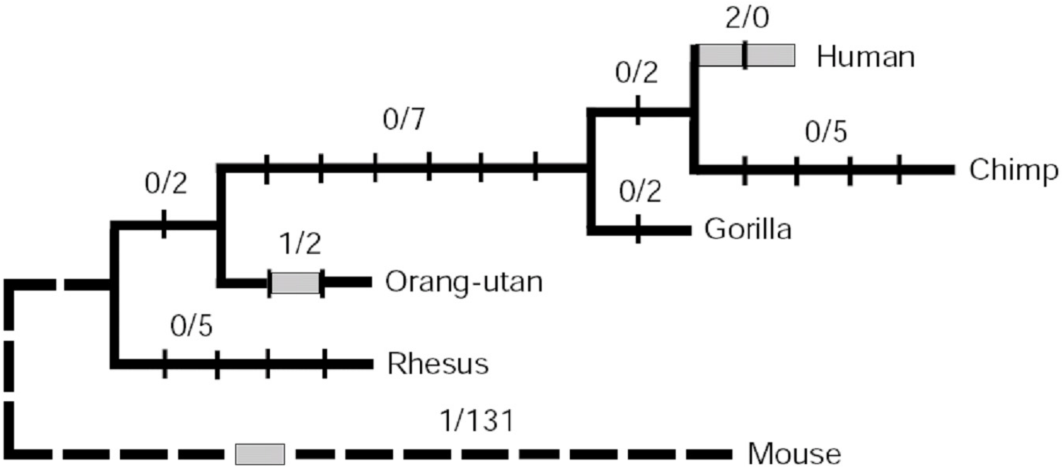


Figure 28: Nucleotide substitutions on the *FOXP2* gene as described by Enard and colleagues (2002). Bars represent mutations over the primate phylogeny and for each lineage the number of amino-acid changes (grey boxes) is indicated over the number of silent nucleotide changes (adapted from (Enard et al. 2002)).

Using ancient DNA techniques, Krause and colleagues (2007) sequenced the two human-specific derived alleles of *FOXP2* in Neanderthal DNA. Their results revealed that these changes are also present in the Neanderthal sequence, suggesting that they may have been fixed or segregating in the ancestral population prior to the human-Neanderthal split, which for nuclear genomic sequences is estimated to have occurred ~0.5 million years ago (Green et al. 2006; Noonan et al. 2006), well before the emergence of modern humans.

One interpretation is that the *FOXP2* gene may have been selected twice during human evolution, but it remains unclear whether the human selective sweep on this gene is actually associated with these two early amino-acid changes. Other functional variants or new splicing forms of the gene could be involved, but no evidences about such alternative explanations have been published so far.

Brain size is another distinct trait of *H. sapiens* (i.e., 1,350 cubic centimeters on average), which is proportionally larger than that of any other animal. Several genes have been described as specific regulators of brain size, mutations in which lead to microcephaly, a condition in which the brain is severely reduced in size (Bond et al. 2002; Jackson et al. 2002). For at least two of them, namely the abnormal spindle homologue microcephaly-associated gene (*ASPM*) and the *Microcephalin* gene (*MCPH1*), phylogenetic analyses have revealed signatures of positive selection in the lineage leading to humans (Evans et al. 2004; Kouprina et al. 2004; Wang and Su 2004). In two additional studies aimed at investigating whether positive selection has continued to operate on these genes after the emergence of anatomically modern humans, *ASPM* and *MCPH1* have been proposed to have unusually long haplotypes and extensive geographical variation within humans, which are both signatures of recent ongoing selection (Evans et al. 2005; Mekel-Bobrov et al. 2005). However, the consistency of their results remains controversial (Currat et al. 2006; Yu et al. 2007) and the demonstration that these patterns arose as a result of positive selection related to increased brain size remains unclear (Nielsen et al. 2007).

Other examples of individual genes that might underlie human-specific adaptations include genes involved in behavior and cognition, such as the monoamine oxidase A gene (*MAOA*) (Andres et al. 2004; Gilad et al. 2002), and genes associated with male reproduction, such as the sperm protamine 1 gene (*PRM1*). The latter, though, shares signals of rapid evolution with other higher primates but not with the rest of mammals (Rooney and Zhang 1999; Wyckoff et al. 2000).

Beyond the analysis of single genes, comparative studies have also provided insights into positive selection acting on gene families. The largest one in mammalian genomes is constituted by the olfactory receptor (OR) genes, which underlie the capacity for odor perception. Several studies have revealed that different evolutionary forces shape the OR gene repertoires of humans and that of great apes. For example, Gilad and colleagues (2003b) reported that humans have accumulated mutations that disrupt OR coding regions roughly 4-fold faster than any other primate species, suggesting a human-specific loss of OR genes. As a possible explanation the authors argued that humans do not rely on their sense of smell as much as apes. In spite of this, using a variant of the McDonald-Kreitman test, Gilad et al. (2003a) found suggestive evidence that positive selection acted on intact OR genes in humans but not in chimpanzees, perhaps in response to human-specific sensory needs. This could be caused by the larger difference in lifestyle between humans and apes than among other primates, leading to novel human olfactory functions. In such study, however, the signature of selection could not be precisely attributed to single OR genes but rather broadly to OR gene clusters. A functional characterization of any specific target of selection was thus not provided.

5.1.1 Genome scans for positive selection between species

With the increasing availability of complete genomes from different organisms over the past few years, comparative studies have been able to move from the analysis of single genes or a few of them, to the screening of the entire set of annotated genes on the whole genome between a variety of species.

In one pioneer study of such genome-wide scans, Clark et al. (2003) used a variant of the branch-site model by Yang and Nielsen (2002) to analyze 7,645 orthologous gene trios from human, chimpanzee, and mouse in order to infer positive selection on the human lineage. They identified 178 genes with more amino acid changes taking place on the human lineage than expected ($p < 0.01$), providing the first list of genes and functional categories about which further studies could formulate interesting hypotheses related to selected phenotypes during human evolution. Some of those genes, indeed, constituted the target of the initial analysis of this work (see **Materials and Methods** below). However, further improvements of the branch-site models (Zhang et al. 2005) suggested that previously available tests were unable to distinguish between relaxation of selective constraint and positive selection, and thus false detection of positively selected genes was frequent (Zhang 2004).

Using a codon-based model, Nielsen et al. (2005) performed a larger scale study on 13,731 human-chimpanzee orthologs. In total, 35 genes were found to have significantly ($p < 0.05$) elevated d_N/d_S ratios, implying the action of positive selection. However, such two-way comparison did not enable them to determine whether selection occurred in the human lineage, the chimp lineage of both. Additionally, they investigated intraspecific variation of the top 50 genes with the strongest evidence for selection by resequencing them in 20 European-Americans and 19 African-Americans. The authors found an excess of high-frequency derived alleles for nonsynonymous mutations in the top genes data set, and some of them, such as the olfactory receptor *OR5I1*, stood out as having particularly high levels of polymorphism according to the HKA test, supporting the presence of selective pressures acting on these genes.

Bustamante et al. (2005) performed a complementary study by analyzing 11,624 protein-coding genes that were sequenced in 20 European-Americans, 19 African-Americans, and one chimpanzee. As in the Gilad et al. study (2003a), they applied the MK-PRF method (a variant of the McDonald-Kreitman test based on the Poisson random field model) to compare levels of polymorphism within humans with the levels of divergence between chimpanzee and human for all informative loci. Overall, 304 out of 3,277 informative loci for positive selection (i.e., genes with at least four variable sites) showed significant ($p < 0.05$) evidence for adaptive evolution, whereas 813 out of 6,033 informative loci for purifying selection (i.e., genes with at least two variable sites) were classified as significantly negatively selected.

In general, these genome-wide studies allow the identification of distinct functional gene categories with an increased likelihood of having experienced positive selection on the human lineage. Mostly genes involved in immune-related functions, sensory perception (especially olfaction), and reproduction (especially spermatogenesis), among others.

Later genome scans have taken advantage of the availability of the complete sequence of the chimpanzee genome (Consortium 2005a). Besides the sequence itself, this study provided an initial comparison with the human genome. As for the analysis of protein-coding regions, the comparison focused on 13,454 pairs of human and chimpanzee genes with unambiguous 1:1 orthology. The authors assessed the rate of evolution for each gene by means of K_A/K_S and K_A/K_I ratios and observed 585 genes (4.4%) with $K_A/K_I > 1$, suggesting rapid evolution.

Arbiza and colleagues (2006a) compared a similar set of genes (i.e., 13, 198 orthologs) but in a larger number of species (i.e., human, chimp, mouse, rat, and dog) focusing on the distinction between genes involved in positive selection from those under relaxation of selective constraints. Using the improved branch-site models developed by Zhang et al. (2005), they tested for positive selection in human, in chimp, and in their ancestral lineage since the divergence from murids. They reported 108 human and 577 chimp positively selected genes according to the more stringent branch-site test for positive selection. Over-represented functional categories in human in relation to chimp include genes related to G-protein coupled receptors and sensory perception.

The most recent genome scans maintain the scale of the number of genes analyzed and the models by means of which the d_N/d_S ratio test is applied, but provide the inclusion of closer outgroups and/or larger phylogenies. For example, Bakewell et al. (2007) analyzed 13, 955 genes of humans and chimps using the macaque monkey as the outgroup. They reported 154 genes showing significant ($p < 0.05$) signals of positive selection in the human lineage and 233 genes in the chimp lineage. On the other hand, Nickel et al. (2008) created a database that contains the results of different tests for positive selection along the human lineage in 13,721 genes with orthologs in multiple mammalian species including the orangutan (i.e., human, chimp, orangutan, macaque, mouse, rat, rabbit, dog, cow, armadillo, elephant, tenrec and opossum). The results of the strict branch-site test include 244 and 152 genes within the 0.05 and 0.01 significance thresholds, respectively.

The usage of the orangutan and macaque genomes as outgroups improves the correct assignment of differences to the human and chimpanzee lineages. However, the identification of human-specific genetic changes that occurred during the last few hundred thousand years, that is, when fully anatomically and behaviorally modern humans appeared, will remain limited until a more closely related genome sequence, such as the Neanderthal genome, is available for comparison (Green et al. 2006).

The following table summarizes some of the main genome scans for selection performed by comparative studies involving the human genome.

Table 3: Genome-wide scans for selection in humans from comparative studies.

Study	Data	Species	Applied test	Year
Clark *et al.*	7,645 genes	Human Chimp Mouse	d_N/d_S test[b]	2003
Nielsen *et al.*	13,731 genes	Human Chimp	d_N/d_S test[c]	2005
	50 genes	Human (N=39)	Derived alleles HKA test	
Bustamante *et al.*	11,624 genes	Human (N=39) Chimp	MK-PRF method[d]	2005
Chimp genome project[a]	13,454 genes	Human Chimp	K_a/K_s ratio K_A/K_I ratio	2005
Arbiza *et al.*	13,198 genes	Human Chimp Mouse Rat Dog	d_N/d_S test Testing for both PS and RSC[e]	2006
Bakewell *et al.*	13,955 genes	Human Chimp Macaque	d_N/d_S test[f]	2007
Nickel *et al.*	13,721 genes	From 8 to 13 mammalian species[g]	d_N/d_S test[h]	2008

[a] The International Chimpanzee Sequencing and Analysis Consortium study.

[b] Using a modified branch-site model (Model 2 of the cited study).

[c] Using a codon-based model.

[d] This is a variant of the classic McDonald-Kreitman test.

[e] Tests for Positive Selection (PS) and Relaxation of Selective Constraints (RSC) based on branch-site models implemented in PAML.

[f] Using branch-site models implemented in PAML.

[g] Out of 13,721 genes, 12,905 had at least a portion of the sequence represented in at least eight of the following species: human, chimp, orangutan, macaque, mouse, rat, rabbit, dog, cow, armadillo, elephant, tenrec and opossum.

[h] Using all available models implemented in PAML v3.14b, including strict branch test, relaxed branch-site and strict branch-site tests.

5.2 Intraspecific Studies

While comparative studies have highlighted putative genetic changes underlying human-specific traits (e.g., language capacity, brain size, etc.), which are shared among all human populations owing to shared selective pressures during human speciation, intraspecific studies of human genetic variation have been able to identify a number of loci involved in recent selective sweeps (up to ~250,000 years), which may underlie population-specific adaptations depending on the geographic distribution of the selective forces acting within humans.

One of the best-understood examples of geographically restricted adaptation in humans is lactose tolerance in European populations. Lactose is the only nutritionally significant carbohydrate in milk, and most mammals, including humans, lose the capacity to metabolize it after weaning from breast milk mainly because lactase production stops. Accordingly, lactase nonpersistence appears to be the ancestral state in humans and is typical in most world populations (Swallow 2003). However, persistence of lactase into adulthood is common in populations where dairy products became a staple food (e.g., northern Europeans), allowing them to digest milk. Therefore, it has long been suspected that the lactase gene (*LCT*) had been the target of recent positive selection after the domestication of cattle around 10,000 years ago (Bersaglieri et al. 2004; Hollox et al. 2001; Myles et al. 2005). Indeed, Enattah et al. (2002) reported that differences in *LCT* expression in adults are caused by a C/T SNP located about 14 Kb upstream of the gene (−13910). Haplotypes carrying the allele associated with lactase persistence segregate at high frequencies in Europeans and extend over long stretches of the genome, that is, up to 1 cM or about 1 Mb (see **Figure 29**). Interestingly, Tishkoff et al. (2007) described in sub-Saharan African populations that use dairy farming, three novel SNPs (G/C −14010, T/G −13915, and C/G −13907) located also upstream of *LCT* that appear to increase its transcription in vitro. These variants are also associated with high-frequency haplotypes in regions of extended LD, indicating that recent positive selection has rapidly increased the frequency of several different *LCT* variants independently in populations from at least two parts of the world.

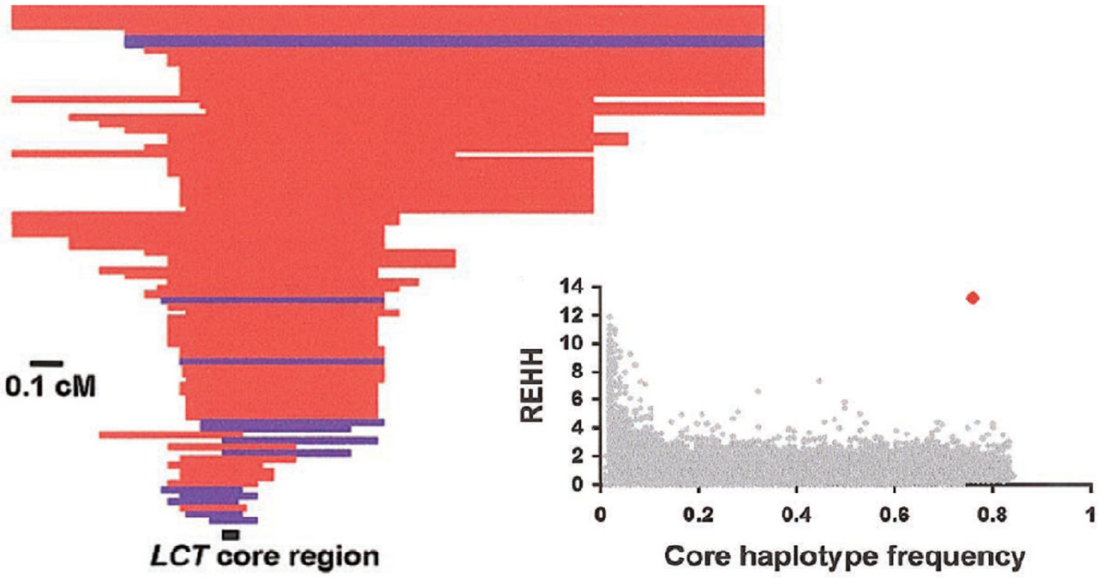


Figure 29: Extended haplotype homozygosity (EHH) around the *LCT* core region containing the −13910C/T variant (black bar) for a European-derived sample. *Left*: The extension of long-range core haplotypes is shown for each chromosome in cM. Core haplotypes containing the persistence-associated allele (−13910T) are shown in red, and those containing the nonpersistence–associated allele (−13910C) are shown in blue. *Right*: Haplotype frequency and relative EHH for the persistence-associated haplotype at *LCT* (red symbol), in comparison with 10,000 sets of simulated data (gray symbols). Both adapted from (Bersaglieri et al. 2004).

Infectious diseases are among the most important selective agents acting on the human genome. Malaria is of particular interest as numerous genes have shown evidence for selection mediated by some of the four types of the disease, being *Plasmodium falciparum* malaria the most deadly one. Most of the human genes that are thought to provide reduced risk from malarial infection are expressed in red blood cells, which are required by the parasite to develop and multiply. In addition to sickle-cell anemia (discussed in **section 3.4**), other resistance mechanisms have been documented. One of them is associated with deficiency of the enzyme glucose-6-phosphate dehydrogenase (*G6PD*), which is required for supplying energy to erythrocytes. Deficiency alleles of the *G6PD* gene cause a number of hemopathologies, but this is counterbalanced by the advantage of conferring resistance to malaria. Thus, the *G6PD* locus shows signatures of positive selection consisting of reduced variability and long-range haplotypes (Sabeti et al. 2002; Saunders et al. 2002; Tishkoff et al. 2001; Verrelli et al. 2002).

Variation at the Duffy blood group locus (*FY*) has also been associated with resistance to malaria infection, in this case to *P. vivax* malaria. Carriers of the Duffy null allele (*FY*O*) do not express a membrane protein on the red blood cells to which the parasite binds to invade the erythrocytes. This allele differs from the ancestral *FY*B* allele by a single noncoding mutation in the promoter of the *FY* gene that prevents transcription. As a result, *FY*O* homozygotes are resistant to this variety of malaria (Livingstone 1984). Allele frequencies at this locus illustrate an extreme case of population differentiation, as the *FY*O* allele is fixed or nearly so in most sub-Saharan Africa and is virtually absent elsewhere (see **Figure 30**). This is consistent with the notion that this allele underwent a complete selective sweep in sub-Saharan African populations, additionally supported by the presence of an excess of high-frequency derived alleles near the *FY* gene (Hamblin and Di Rienzo 2000; Hamblin et al. 2002).

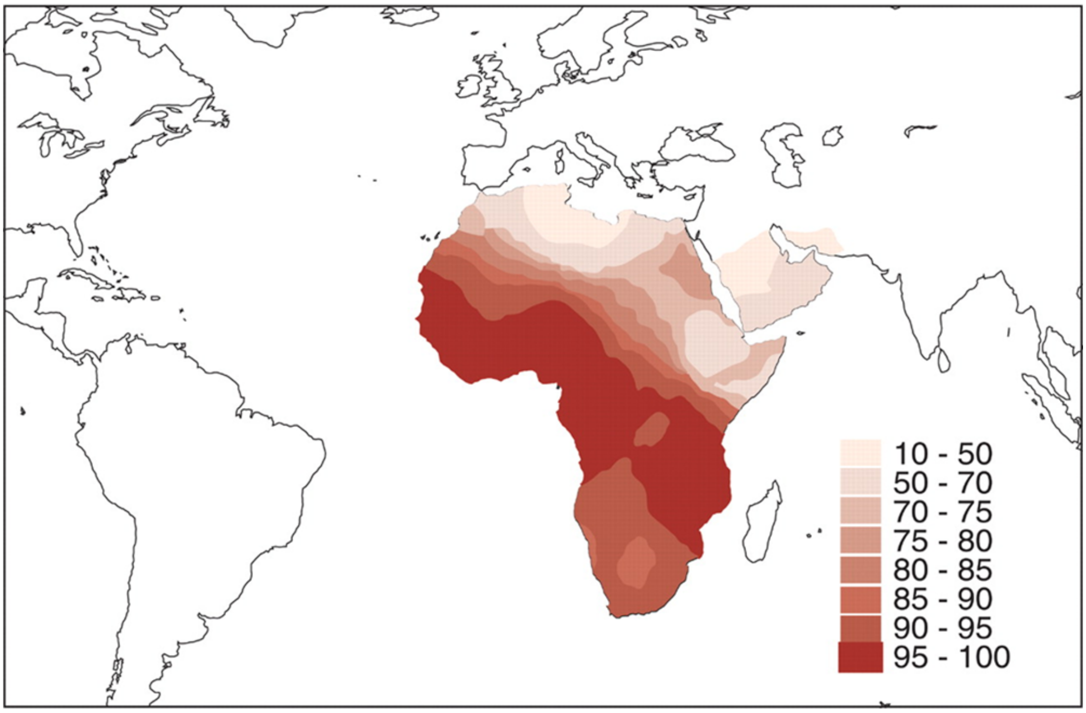


Figure 30: Geographical distribution of the *FY*O* allele, which confers resistance to *P. vivax* malaria, showing extreme population differences in allele frequency (color-coded scale). The *FY*O* allele is prevalent and even fixed in many African populations, but virtually absent outside Africa (adapted from (Sabeti et al. 2006)).

Skin pigmentation is a polygenic phenotype; the architecture of the genes involved has been largely dissected much better than for any other such complex character (Rees 2003 and references therein). Although it has traditionally been used to define human races, skin color actually varies more or less continuously world-wide (see **Figure 31** and **section 2.6** for a brief discussion about genetics and race). Dark skin is thought to be the ancestral phenotype, it protects against the damaging effects of UV radiation. However, UV light is required for synthesis of vitamin D, and in higher latitudes, where UV levels are lower, dark-skinned individuals suffer from vitamin D deficiency. Thus, it is plausible that pale skin has evolved in higher latitudes to avoid bone softening diseases derived from hypovitaminosis D. Human skin pigmentation has a complex genetic basis involving several genes. An early study examined variation at the melanocortin 1 receptor (*MC1R*) locus, which influences variation in skin and hair color, reporting evidences for different selective pressures between Africans and non-Africans (Harding et al. 2000). More recently, the human homolog of a gene (*SLC24A5*) that influences pigmentation in zebra fish was found to have an alanine to threonine amino-acid replacement at position 111 that explains about one-quarter of the difference in pigmentation between Europeans and Africans (Lamason et al. 2005). The light pigmentation allele is fixed or nearly fixed in Europeans and the *SLC24A5* gene region exhibits low diversity levels. Other signatures of selection for lighter pigmentation outside Africa, such as high F_{ST} values and long-range haplotypes, have been found in a number of candidate genes including *OCA2*, *MATP*, *DCT*, *ASIP*, *KITLG*, and *TYRP1* among others (Izagirre et al. 2006; Lao et al. 2007; Myles et al. 2007; Norton et al. 2007; Sulem et al. 2007).

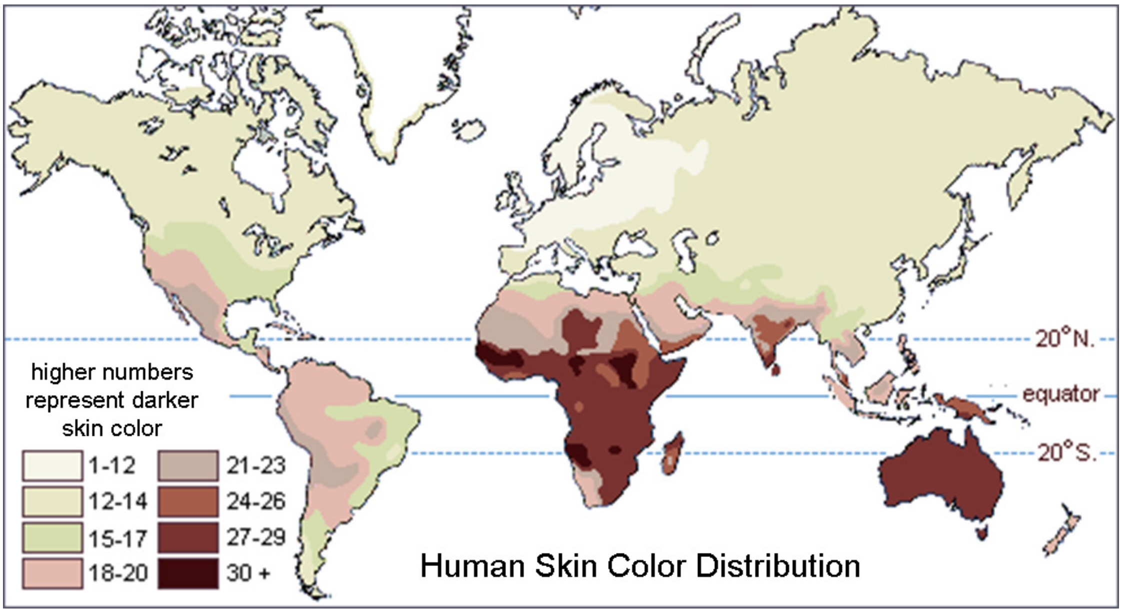

20°N.
equator
20°S.
higher numbers represent darker skin color
1-12
12-14
15-17
18-20
21-23
24-26
27-29
30 +
Human Skin Color Distribution

Figure 31: A world map showing that variation in human skin color is associated with levels of UV irradiation, which are higher near the equator. Data for native populations collected by R. Biasutti prior to 1940 (adapted from (Barsh 2003)).

Other examples of genes that have a signature for selection in different human populations include those involved in response to pathogens, such as *HLA* class I and class II genes (Hughes and Yeager 1998), resistance to human immunodeficiency virus, such as the *cis*-regulatory region of *CCR5* (Bamshad et al. 2002; Sabeti et al. 2005), drug metabolism, such as *CYP1A2* (Wooding et al. 2002), alcohol metabolism, such as *ADH1B* and *ALDH2* (Oota et al. 2004; Osier et al. 2002), dietary calcium uptake, such as *TRPV6* (Hughes et al. 2008), color vision, such as *OPN1LW* (Verrelli and Tishkoff 2004), and taste perception, such as the bitter-taste receptor *TAS2R16* (Soranzo et al. 2005).

5.2.1 Genome scans for recent positive selection within humans

Unlike candidate gene approaches, in which particular genes are identified *a priori* as candidates on the basis of functional information, whole genome scans for selection intend to detect loci under positive selection without such prior knowledge. Because the screening of polymorphisms is done genome-wide, they can be used equally well to detect selection in non-coding and protein-coding regions, but the results are usually interpreted in terms of predictions for annotated genes.

The vast majority of intraspecific genome scans for selection in humans are based on SNP data, although one of the first genome-wide studies used 332 STRs in Africans and Europeans to detect candidate regions influenced by local natural selection (Kayser et al. 2003). In a contemporary study, Akey et al. (2002) interrogated 26,500 SNPs in African-Americans, European-Americans and East Asians for signatures of selection and found 174 candidate genes based on F_{ST} values.

In 2005, the HapMap project (Consortium 2005b) and Perlegen Sciences (Hinds et al. 2005) provided two of the main catalogues of human genetic variation in which most of the later genome scans for selection were based on. The HapMap project initially provided polymorphism data for ~1 million SNPs in 270 individuals from three different human populations (see **section 2.4**), whereas the Perlegen dataset consisted of ~1.5 million SNPs genotyped in 71 individuals from three human populations of similar ancestry as HapMap (see **section 2.4**). Important differences between the two datasets include that HapMap individuals, with the exception of Utah residents of European ancestry, were sampled from native populations; whereas Perlegen samples are all from American citizens. The latter, though, is known to be less affected by ascertainment bias towards common variants (Clark et al. 2005). In the HapMap phase I publication, different signatures of selection were explored including population differentiation, site frequency spectrum distortions (low heterozygosity/rare alleles), and the LRH test. The analysis revealed 926 SNPs in 27 genes for which differentiation among populations was more extreme than for the *FY* locus. On the other hand, in the Perlegen scan F_{ST} and LD extension were only evaluated in terms of whole-genome patterns.

Over the past few years, a number of genome-wide studies, aimed at detecting different signatures of selection, have applied different statistical tests either to HapMap data, Perlegen data or both (see **Table 4**).

Genome scans applying tests based on site frequency spectrum include the Carlson et al. (2005) study and the Williamson et al. (2007) scan for recent adaptive evolution. The former used a sliding window analysis of Tajima's *D* across the human genome to identify Contiguous Regions of Tajima's *D* Reduction (CRTRs), whereas the latter introduced a composite-likelihood approach (i.e., the composite likelihood ratio (CLR) test) to localize recent complete selective sweeps at a fine scale (i.e., within 100 Kb windows). Both studies analyzed the Perlegen data set. Carlson et al. (2005) identified seven, 23, and 29 CRTRs in populations of African, European and Chinese descent, respectively. Williamson et al. (2007) identified 101 regions with strong evidence of recent positive selection, within which new and previously reported genes of biological interest were found, including clusters of olfactory receptors, pigmentation genes, heat shock genes and immune system genes. Of special interest for the work presented here, is the identification of a gene (*VPS37C*) as the closest target of the estimated position of the sweep within one of such top selected regions in the Chinese sample.

Genome scans for selection derived either from the HapMap project or Perlegen data, and focused on population differentiation have been mostly based on F_{ST} (see section 4.2.3). Weir et al. (2005) estimated F_{ST} values for autosomic SNPs using both Perlegen and Phase I HapMap data; Myles et al. (2008) introduced an algorithm that identifies genomic regions containing SNPs with extreme F_{ST} using Perlegen data; and Barreiro et al. (2008) used Phase II HapMap data to analyze the degree of population differentiation in a denser SNP dataset. Together, these studies have identified a fraction of loci that may contribute to large phenotypic differences between populations, many of which have likely experienced positive selection.

Finally, a number of EHH-based statistics have been used to scan the whole genome for signatures of recent positive selection. These include the integrated haplotype score (iHS), initially applied to Phase I HapMap data (Voight et al. 2006); the LRH test adapted for whole genome analysis (WGLRH test) applied to genotype data from Asian, Caucasian and African-American populations analyzed with a 100K SNP array

(Zhang et al. 2006); and the LD decay (LDD) test applied to both Perlegen and Phase I HapMap data by Wang et al. (2006). It is important to note that these scans made use of methods designed for detecting incomplete selective sweeps (i.e., in which the favored allele still segregates at intermediate frequencies). In agreement with this, the authors have found significant signatures of selection for known cases, such as the *LCT* and *G6PD* genes, as well as for new candidates, such as cytochrome P450 genes, among many others.

On the other hand, genome scans applying statistical tests designed to detect nearly or complete selective sweeps (i.e., in which the favored allele has reached fixation) have appeared more recently in the literature and are thought to be complementary to the above mentioned. These include the Tang et al. (2007) study, in which a novel between population EHH statistic (i.e., Rsb) was applied to interrogate both Perlegen and HapMap datasets. Kimura et al. (2007) reported an analogous method, the ratio of haplotype homozygosity between populations (rHH and rMHH), which they used to scan Phase I HapMap data only. As for the writing of this thesis, the last EHH-based genome-wide scan for selection appeared along the publication of the Phase II HapMap data (Frazer et al. 2007), which Sabeti and colleagues (2007) used to apply the Cross Population Extended Haplotype Homozygosity (XP-EHH) test, analogous to Rsb, to detect complete selective sweeps. Although they also exploited Phase II HapMap data to apply the aforementioned iHS and LRH test. Combining these evidences they found more than 300 candidate regions, and focusing on the twenty-two strongest candidates for selection, they scrutinize the regions to identify possible targets of selection. The authors highlight three pairs of genes with different biological functions that have apparently undergone positive selection in the same population: *LARGE* and *DMD*, both related to Lassa virus infection (Kunz et al. 2005a; Kunz et al. 2005b), in West Africa; *SLC24A5* and *SLC45A2*, both involved in skin pigmentation (Graf et al. 2005; Lamason et al. 2005), in Europe; and *EDAR* and *EDA2R*, both involved in development of hair follicles (Botchkarev and Fessing 2005), in Asia.

Table 4: Genome-wide scans for selection in humans from intraspecific studies.

Study	Data and samples	Explored signature	Applied test	Sensitive to detect	Year
Kayser *et al.*	332 STRs 23 Africans 48 Europeans	Population Differentiation	R_{ST} [a]	Geographically restricted sweeps	2003
Akey *et al.*	26,530 SNPs 42 Afr-American 42 Eur-American 42 East Asian		F_{ST}		2002
Perlegen[b]	1,586,383 SNPs 23 Afr-American 24 Eur-American 24 Han Chinese		F_{ST}		2005
HapMap[b]	1,007,329 SNPs[c] 90 YRI[d] 90 CEU[d] 45 CHB[d] 45 JPT[d]		F_{ST}		2005
		Long-range haplotypes	LRH test	Incomplete sweeps	
Weir *et al.*	Perlegen data Phase I HapMap	Population Differentiation	F_{ST}	Geographically restricted sweeps	2005
Myles *et al.*	Perlegen data		F_{ST}		2008
Barreiro *et al.*	Phase II HapMap[e]		F_{ST}		2008
Carlson *et al.*	Perlegen data	SFS distortions[f]	Tajima´s *D*[g]	Nearly completed or complete sweeps	2005
Williamson *et al.*	Perlegen data		CLR test		2007
Voight *et al.*	Phase I HapMap	Long-range haplotypes	iHS statistic	Incomplete sweeps	2006
Wang *et al.*	Perlegen data Phase I HapMap		LDD test		2006
Zhang *et al.*	116,204 SNPs 42 Afr-American 42 Caucasian 37 Asian		LRH test (WGLRH)		2006
Tang *et al.*	Perlegen data Phase I HapMap		Rsb statistic	Nearly completed or complete sweeps	2007
Kimura *et al.*	Phase I HapMap		rHH statistic		2007
Sabeti *et al.*	Phase II HapMap		XP-EHH statistic		2007
			iHS statistic LRH test	Incomplete sweeps	

[a] Analogous to the F_{ST} statistic for Short Tandem Repeats (STRs) data

[b] These studies are not primarily intended to perform a scan for selection, so the applied tests referred here concern the analysis of positive selection only.

[c] Refers to Phase I HapMap data

[d] Abbreviations refer to samples from the following populations: Yoruba in Ibadan, Nigeria; Utah residents, USA, from the Centre d'Etude du Polymorphisme Humain (CEPH) collection; Han Chinese in Beijing, China; and Japanese in Tokyo, Japan.

[e] Phase II HapMap data includes 3,107,620 SNPs genotyped in the same 270 individuals.

[f] Site Frequency Spectrum distortions (e.g. excess of rare alleles).

[g] A sliding window analysis was applied to identify Contiguous Regions of Tajima's *D* Reduction (CRTRs).

Together, genome scans for selection within humans and between species, have identified, as of late 2006, nearly 2,500 genes as putative targets of positive selection in humans (Biswas and Akey 2006), the majority of which had not been previously evaluated as candidates. Direct comparison among studies is difficult because of differences in experimental design, which explains the low concordance between results (Nielsen et al. 2007). Interestingly, selection signatures are frequently reported to be confined to a single population rather than shared among populations, reinforcing the importance of disentangling the selective pressures that acted on earlier stages of our evolutionary history (as revealed by comparative studies) from those acting on human populations after the emergence of modern humans or even after their out-of-Africa dispersal, about which intraspecific studies can give insights.

For instance, with the exception of the Nielsen et al. (2005) genome scan of human-chimpanzee orthologs with further resequencing of the top candidate genes in 39 human individuals, few studies, if any, have intentionally investigated intraspecific variation of several human genes with previous evidences of positive selection from comparative studies. In such study, the authors took advantage of the availability of a large set of human-chimp orthologs, but such limited comparison (i.e., between two species only) precluded them from identifying human-specific changes. In addition, although having the advantage of possessing resequencing data, this corresponded to as few as 39 individuals from two different populations and only two signatures of recent positive selection were analyzed in their candidate set of genes.

Further studies performed with a larger number of species with subsequent analysis of all detectable signatures of recent positive selection in a comprehensive diversity panel of human populations, would enable to address additional issues such as whether positive selection has continued to operate within humans on a candidate set of genes once they have acquired human-specific genetic changes.

In summary, the continued collection of genotype data from large-scale projects has been useful for conducting studies of single genes as well as complete genome scans to detect evidence for positive selection in humans. However, more focused studies to test specific hypothesis or to follow up on previously reported results will have an important place in reconstructing the overall history of selective pressures among human populations.

OBJECTIVES

The work presented in this thesis is aimed at the comprehension of the diversity and evolution of particular genes that might have played an important role during early human evolution. Two alternative outcomes can be envisaged, either these genes are still evolving adaptively within human populations, or they show no evidences of recent positive selection, probably as a result of a primary implication in defining human-specific traits prior to the emergence of anatomically modern humans.

The objective of this work, then, is to study the human genetic variation, with particular interest of detecting recent selective sweeps, in a set of genes undergoing positive selection on the evolutionary lineage leading to modern humans since the split from the chimpanzee lineage. For this, the project involved two main steps: first, the conduction of a genome-wide comparative study in order to detect the fraction of genes undergoing accelerated protein evolution on the human lineage, and second, the further investigation of intraspecific variation by genotyping a limited number of candidate genes in human populations distributed world-wide. The specific aims concerning the aforementioned steps are:

> **Specific Aim 1.** To identify the fraction of protein-coding genes of the genome with significant evidence of having been subject to positive selection on the human lineage by comparison with closely related species.

To this end, we made use of both previously reported lists of positively selected genes available at the time of the beginning of the study, and public databases to retrieve all annotated coding sequences in several species to perform our own calculations (the complete list of genes derived from the latter can be found in **Appendix 1**).

> **Specific Aim 2.** To analyze the global patterns of genetic variation within humans in a subset of candidate genes derived from Specific Aim 1 in order to scan for signatures of recent positive selection.

To accomplish this goal, we decided to SNP-genotype more than 1,000 individuals from 39 different human populations representing global human diversity. Then, in order to ensure that our analyses encompass a comprehensive time scale over which selective sweeps involving recent human history may have occurred, we decided to apply at least one method for testing each signature of selection detectable from polymorphic data (see methods for detecting selection from intraspecific diversity within **section 4.2** for review).

The results are presented separately for the main set of candidate genes (**Chapter 1**) and for two single candidate genes (**Chapters 2 and 3**), in which besides the SNP data from 39 populations, DNA resequencing data was also gathered from two and three additional populations, respectively (see Materials and Methods for details).

> **Specific Aim 3.** To evaluate, when possible, the functional impact of putatively selected variants in those candidate genes, if any, showing clear evidences of being involved in a recent selective sweep in human populations.

The main purpose here is to add valuable information about the possible phenotypic effects on protein structure and function of the genetic changes presumably favored by selection, which ultimately may help to elucidate the underlying selective pressure acting on these candidate genes. In **Chapter 2** of the Results section we give an example of this, and in **Chapter 1** we functionally characterize, to a more limited extent, a particular genomic region surrounding the strongest candidate for selection.

MATERIALS AND METHODS

Selection of Candidate Genes

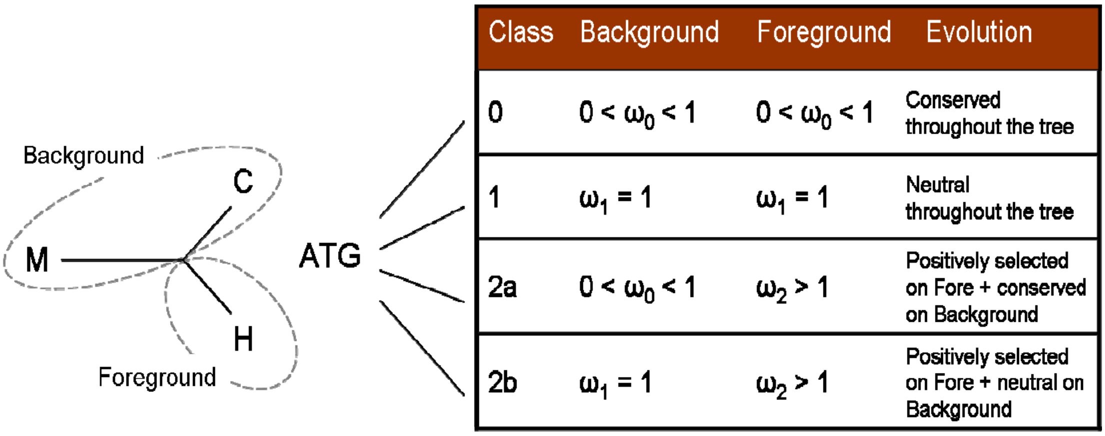
Three separate analyses were performed for choosing candidate genes based on different phylogenetic approaches, which provided a total of 30 genes that were genotyped in human samples.

In the first analysis, seven genes were selected from the Human-Chimp-Mouse orthologous gene trios reported by Clark et al. (2003) as positively selected in the human lineage. To infer non-neutral evolution in these genes the authors applied a modified branch-site model (referred to as Model 2 in their publication), based on the method described by Yang and Nielsen (2002). It is technically important to take into account that the human-chimp gene alignments were not constructed using the chimpanzee genome assembly (which was not available at that time), but by sequencing essentially all human exons in one male chimpanzee at Celera Genomics. For the inclusion of candidates in our study, we focused on genes from different functional categories to evaluate a wide range of possible selective pressures, and spanning no more than 300 Kb to enable dense SNP coverage. Additionally, we decided to include the *LCT* gene, for which there is sufficient evidence of a recent selective sweep detected from polymorphic data within humans (see **section 5.2**), so we could validate our own approach by genotyping this gene in our samples. **Table 5** summarizes the eight genes that were chosen in this first analysis.

For the second analysis, we used the publicly available database of alignments from the Clark et al. study to detect positive selection in the human branch following complementary models for Phylogenetic Analysis by Maximum Likelihood (PAML). Specifically, 7,645 Human-Chimp-Mouse alignments were used to compute d_N/d_S ratios according to the strict branch-based model of the codeml program implemented in the PAML package (Yang 1997). A likelihood ratio test (LRT) was constructed by comparing a model that allows positive selection on the (human) foreground lineage (i.e., Model 2 allowing free d_N/d_S ratios for branches), with a model that does not allow such scenario of positive selection (i.e., Model 3 with d_N/d_S ratio for the human branch fixed to 1). Preference was given to alignments with: i) d_N/d_S ratio > 1 in the human branch for Model 2, ii) d_N rate ≥ 0.0035 for the human branch in Model 2, and

iii) a known gene symbol annotated in the human genome. A total of 45 genes met these criteria, although none of them had a p value < 0.05 for the LRT test of Model 2 vs Model 3. Statistically significant signals from branch models are known to be rare at a genome-wide scale (Sabeti et al. 2006) as multiple selected changes in a gene are required before it will stand out against the background neutral substitution rate. This would be in agreement with the absence of significant values in our results. In spite of this, based on polymorphic information available in public databases such as HapMap and dbSNP, we were able to extract a representative subset of 11 candidate genes to be genotyped in human samples (see **Table 6**).

In the third analysis for selecting candidate genes, we took advantage of the availability of the complete Chimp genome sequence (Consortium 2005a), to construct new human-chimp gene alignments including not only Mouse, but also Rat and Dog genomes as outgroups of the phylogeny. For each species, all annotated coding sequences as of March 2006 were downloaded from the Ensembl database obtaining nearly 12,000 Human-Chimp-Mouse-Rat-Dog orthologous genes, out of which, 9,170 resulted to have a unique best-reciprocal match according to BLAST, and were therefore included to construct the alignments using ClustalW (Thompson et al. 1994). To test for positive selection in the human lineage, the LRT referred to as Test 2 of the improved branch-site model described by Zhang et al. (2005), was applied using the codeml program implemented in PAML v3.14. In this test, each codon is assigned to one of four assumed classes of sites. Site class 0 includes codons that are conserved throughout the tree ($0 < \omega_0 < 1$). Site class 1 includes codons evolving neutrally throughout the tree ($\omega_1 = 1$). Site classes 2a and 2b include codons that are conserved or neutral on the background branches, but evolving under positive selection on the foreground branch ($\omega_2 > 1$). These parameters constitute the alternative hypothesis of the test (see **Figure 32**). The null hypothesis is the same model but with $\omega_2 = 1$ fixed. The likelihood (λ) of the data fitting each of these hypotheses is computed separately for each gene. The LRT then compares both likelihood values, in which significant differences, as revealed by a χ^2 distribution, strongly suggest the action of positive selection on the foreground lineage (see **Figure 33**).



Class	Background	Foreground	Evolution
0	$0 < \omega_0 < 1$	$0 < \omega_0 < 1$	Conserved throughout the tree
1	$\omega_1 = 1$	$\omega_1 = 1$	Neutral throughout the tree
2a	$0 < \omega_0 < 1$	$\omega_2 > 1$	Positively selected on Fore + conserved on Background
2b	$\omega_1 = 1$	$\omega_2 > 1$	Positively selected on Fore + neutral on Background

Figure 32: Simplified explanation of the branch-site model implemented by Zhang *et al.* (2005). *Left*: Schematic representation of a Human (H)-Chimp (C)-Mouse (M) phylogeny indicating the background and foreground branches when testing for positive selection in the human lineage. *Right*: A hypothetical codon (ATG) and the four site classes to which it can be assigned. The parameters inside the table correspond to the alternative hypothesis of the LRT.

The LRT provided a total of 52 genes with significant ($p < 0.01$) differences between the null and alternative models and up to 88 at the 0.05 significance level (listed in **Appendix 1** of this thesis). However, a strict quality control was applied before the inclusion of candidate genes for further analysis in the study. In order to control for alignment errors, each significant candidate ($p < 0.05$) was manually rechecked, and those having large gaps or excessive mismatches were removed from the study. On the remaining candidates, preference was given to alignments with $\omega > 1$ for site classes 2a and 2b in the alternative model, and posterior probability of evolving under positive selection ≥ 95% for at least one codon, according to the Bayes Empirical Bayes (BEB) analysis implemented in PAML. A total of 15 alignments passed the quality control, out of which, only eleven had a reviewed gene symbol and known function in the human genome, thus, they were included for further investigation of intraspecific variation within humans. **Table 7** summarizes this set of candidate genes.

Table 5: Candidate genes selected in the first analysis: from Clark *et al.* (2003).

Symbol	Gene name	Chr.	Size (kb)	SNPs	Biological function	p-value
EYA4	Eyes absent homolog 4 (Drosophila)	6	288	41	Sensory perception (Vision)	0.00348
DIAPH1	Diaphanous homolog 1 (Drosophila)	5	103.9	24	Gametogenesis	0.02010
TECTA	Tectorin alpha precursor	11	88.1	25	Sensory perception (Hearing)	0.00002
GSTZ1	Glutathione transferase zeta 1	14	10.5	20	Amino acid metabolism	0.00090
ALDH6A1	Aldehyde dehydrogenase 6 family, member A1	14	24.3	15	Nucleic acid metabolism	0.02720
BCKDHA	Branched chain keto acid dehydrogenase E1, alpha polypeptide	19	27.2	17	Aminoacid metabolism	0.00464
FOXI1	Forkhead box I1	5	3.8	21	Sensory perception (Hearing)	0.00287
LCT	Lactase	2	49.3	25	Lactose metabolism	

Number of SNPs refers to successfully genotyped SNPs, functional classification is based on PANTHER biological processes, and P-values are as reported by Clark *et al* (2003) of their Model 2 for detecting positive selection in the human lineage (except for *LCT*).

Table 6: Candidate genes selected in the second analysis: branch model (PAML).

Symbol	Gene name	Chr.	Size	SNPs	Biological function	d_N	d_N/d_S
AIRE	Autoimmune regulator	21	12.3	16	Immune response	0,0056	19,18
CLDN8	Claudin 8	21	2.0	9	Cell structure and motility	0,0043	86,60
CST2	Cystatin SA	20	2.9	9	Protein metabolism and modification	0,0240	1,14
DACT1	Dapper homolog 1, antagonist of beta-catenin (xenopus)	14	10.1	12	Multicell. organismal development	0,0043	10,81
GIP	Gastric inhibitory polypeptide	17	10.0	9	Carbohydrate metabolism	0,0046	137,68
HCLS1	Hematopoietic cell-specific Lyn substrate 1	3	29.4	14	mRNA transcription regulation	0,0062	548,48
IL15RA	Interleukin 15 receptor, alpha	10	25.8	21	Immune response	0,0138	41,05
IL1RL2	Interleukin 1 receptor-like 2	2	52.3	22	Immune response	0,0047	2,73
MRPL35	Mitochondrial ribosomal protein L35	2	13.3	13	mRNA transcription regulation	0,0059	1,84
OR5I1	Olfactory receptor, family 5, subfamily I, member 1	11	0.9	11	Sensory perception (Olfaction)	0,0066	1,94
TMPRSS2	Transmembrane protease, serine 2	21	43.6	19	Protein metabolism and modification	0,0080	1,65

Size is indicated in Kb, number of SNPs refers to successfully genotyped SNPs, functional categories are based on PANTHER biological processes, and d_N and d_N/d_S values are for the human branch according to the applied branch model (see text for details).

Table 7: Candidate genes selected in the third analysis: branch-site model (PAML).

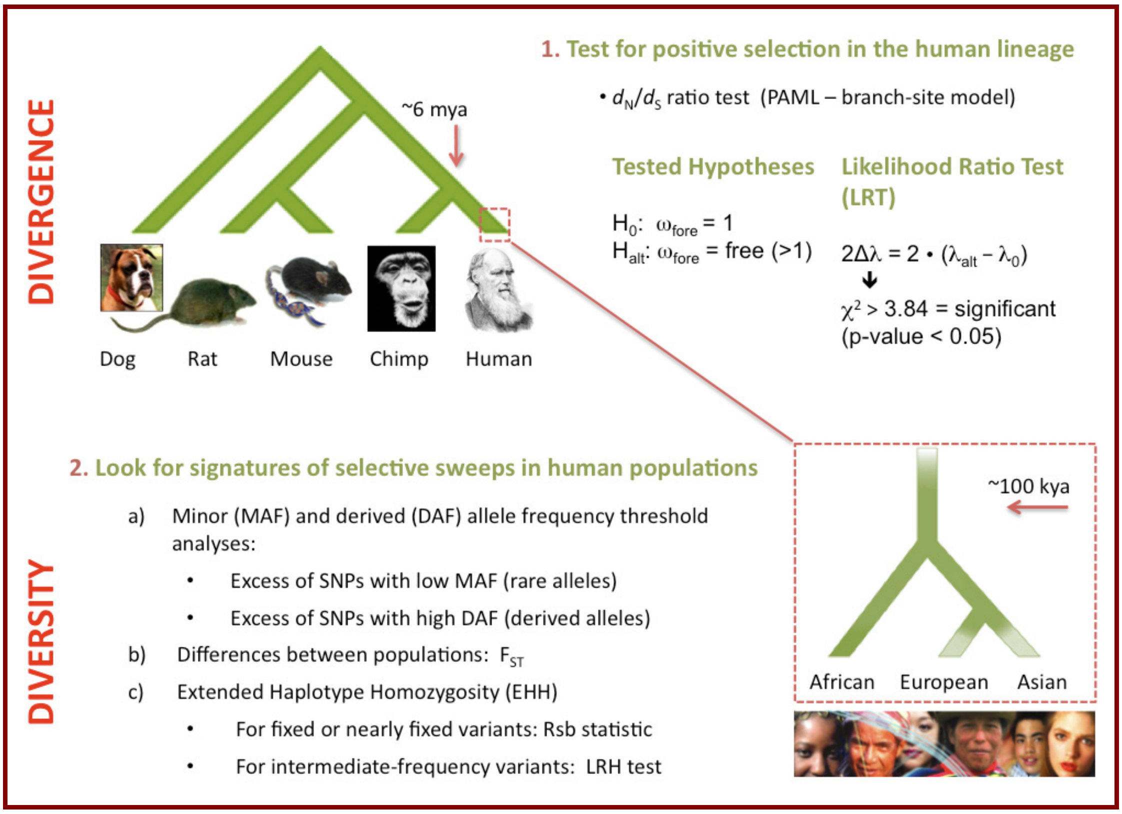
Symbol	Gene name	Chr.	Size	SNPs	Biological function	ω	2ΔL	p-val
GFRA3	GDNF family receptor alpha 3	5	22.2	20	Glycobiological activity	999	10.6	0.0012
PTGER4	Prostaglandin E receptor 4	5	13.8	20	Immune response (skin)	625	10.2	0.0014
HDHD3	Haloacid dehalogenase-like hydrolase domain containing 3	9	2.6	19	Glycobiological activity	999	10.1	0.0014
LHPP	Phospholysine phosphohistidine inorganic pyrophosphate phosphatase	10	152.3	26	Phosphate metabolism	999	9.8	0.0017
CA14	Carbonic anhydrase XIV	1	7.3	17	Carbon metabolism	999	8.7	0.0032
OR2A14	Olfactory receptor, family 2, subfamily A, member 14	7	0.9	16	Sensory perception (Olfaction)	137	7.1	0.0077
VPS37C	Vacuolar protein sorting 37C	11	31.2	22	Endosomal transport (including viral budding)	620	6.8	0.0091
MRC2	Mannose receptor, C type 2	17	64.9	23	Endocytosis (inlcuding binding of HIV)	88.5	6.5	0.0111
USP2	Ubiquitin specific peptidase 2	11	25.4	22	Cell cycle regulation and apoptosis	126	5.8	0.0161
OR5G1P	Olfactory receptor, family 5, subfamily G, member 1, pseudogene	11	0.8	17	Sensory perception (Olfaction)	217	5.7	0.0168
ADI1	Acireductone dioxygenase 1	2	21.6	21	Amino acid metabolism	82.5	5.0	0.0256

Notes: Size is in Kb, SNPs are successfully genotyped SNPs, functional categories are based on PANTHER biological processes, ω values are for site classes 2a and 2b in the human branch, 2ΔL is the twice difference of likelihoods between null and alternative models, and p-values are for the corresponding LRT (see text for details).

With the availability of the complete sequence of the Chimpanzee genome, we decided to consider as reliable analyses only those carried out on multiple sequence alignments based on the chimp genome assembly (CHIMP2.1, Mar 2006). Therefore, we intentionally reanalyzed the 19 candidate genes that were already genotyped in human samples (derived from the first and second analyses), by retrieving from Ensembl (Mar 2006) their corresponding orthologous sequences for three (Human-Chimp-Mouse) and five (Human-Chimp-Mouse-Rat-Dog) mammalian species. However, no significant results could be recovered for these genes after applying a number of different methods, including the initial branch model and the improved branch-site model in PAML; neither after retrieving K_a/K_s ratios reported by The Chimpanzee Sequencing and Analysis Consortium (2005) from the human-chimp genome-wide comparison.

In consequence, we could no longer treat these particular set of 19 genes as a single group of candidates sharing the common feature of undergoing accelerated protein evolution in the human lineage. Nonetheless, single genes showing interesting patterns of intraspecific variation, namely *OR5I1* and *FOXI1*, were further studied as individual candidates and their results are presented in **Chapter 2 and 3** respectively of the Results section of this thesis.

On the other hand, the remaining set of 11 genes (derived from the third analysis), was indeed considered as a highly consistent group of fast-evolving genes due to i) the reliable source of the alignments ii) the quality control described above, and iii) the replication of similar results in a number of independent studies (see **Discussion** for details). Therefore, they constitute the set of genes that better responds to the aims and scope of this work, so they are presented in the **first chapter** of results of this thesis. An overview summarizing the followed workflow for studying both interspecific divergence and intraspecific diversity in these genes is shown in **Figure 33**.



1. Test for positive selection in the human lineage

- d_N/d_S ratio test (PAML – branch-site model)

Tested Hypotheses

H_0: $\omega_{fore} = 1$
H_{alt}: ω_{fore} = free (>1)

Likelihood Ratio Test (LRT)

$2\Delta\lambda = 2 \cdot (\lambda_{alt} - \lambda_0)$
↓
$\chi^2 > 3.84$ = significant
(p-value < 0.05)

2. Look for signatures of selective sweeps in human populations

a) Minor (MAF) and derived (DAF) allele frequency threshold analyses:
- Excess of SNPs with low MAF (rare alleles)
- Excess of SNPs with high DAF (derived alleles)

b) Differences between populations: F_{ST}

c) Extended Haplotype Homozygosity (EHH)
- For fixed or nearly fixed variants: Rsb statistic
- For intermediate-frequency variants: LRH test

Figure 33: Workflow overview for the analysis of human genetic variation in fast-evolving genes. First (*top*), following a genome-wide approach, we performed a phylogenetic analysis by maximum likelihood between closely related species to test for positive selection in the human branch (see text for details). Second (*bottom*), following a candidate gene approach, we explored a number of signatures of recent selective sweeps detectable from intraspecific SNP data in human populations from different geographic regions of the world (shown picture is orientative only; see below for details about samples and populations included).

DNA Samples

For investigating global patterns of SNP variation in all candidate genes we analyzed the Human Genome Diversity Cell Line Panel (HGDP-CEPH) (Cann et al. 2002), which contains 1,064 DNA samples from individuals belonging to 51 different populations distributed worldwide (see **Figure 34**). This set of populations represents most of the complete human genetic diversity, as reported by Rosenberg et al. (2002). Unless otherwise stated, all population analyses presented throughout this thesis are based on the HGDP-CEPH diversity panel; however, the subset of DNA samples used differs from the originally reported by Cann et al. (2002) in two aspects: on the one hand the initial 51 populations were regrouped into 39 study populations based on geographic and ethnic criteria to maximize sample sizes; the resulting 39 populations were in turn grouped into 7 major continental regions for the purpose of analysis in a substantial part of the work (see **Table 8** and below for details). On the other hand, following recommendations reported by Rosenberg (2006), a number of atypical, duplicated and deducted first-degree related individuals were removed from the original panel, resulting in the so-called H971 standardized subset, whose sample sizes adjusted to our 39 study populations are given in **Table 8**.

Details about the regrouping of 51worldwide populations are as follows: Tuscans and Bergamese were grouped into North Italians; Uygur, Tu, and Xibo populations were combined as Northwest Chinese; Daur, Hezhen, Mongolian, and Oroqen populations as Northeast Chinese; and Dai, Lahu, Miaozu, Naxi, She, Tujia, and Yiku as South Chinese (see **Figure 35**). For the remaining populations, we conserved the original organization of the HGDP-CEPH diversity panel. The total number of populations was thus 39. The seven geographical regions in which they were further grouped are: Sub-Saharan Africa (SSAFR), Middle East-North Africa (MENA), Europe (EUR), Central-South Asia (CSASIA), East Asia (EASIA), Oceania (OCE) and America (AME) (see **Figure 34**).

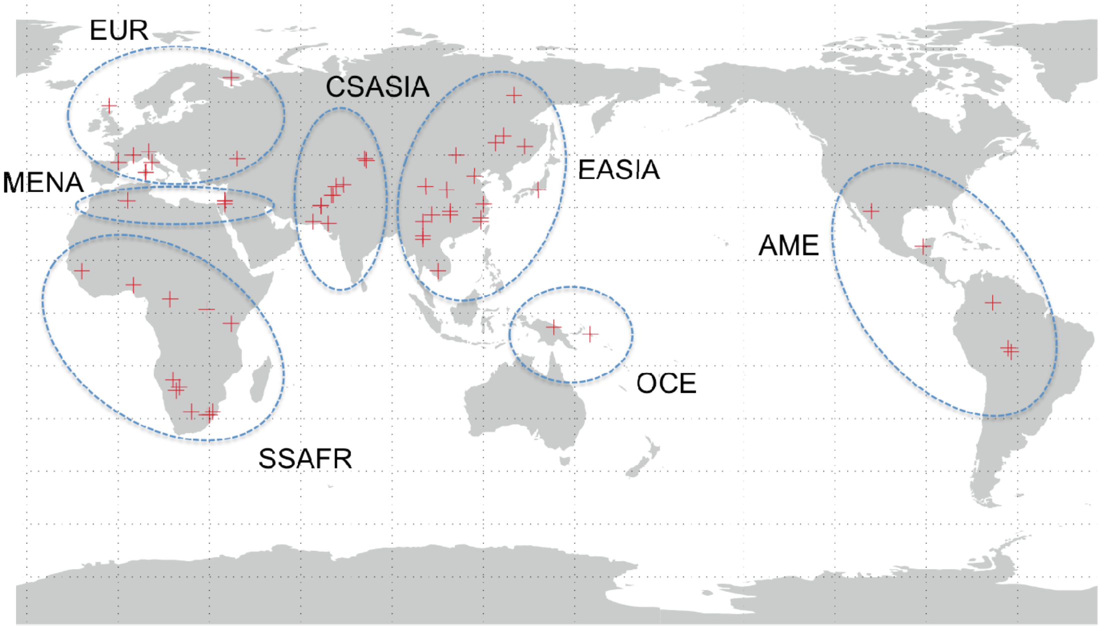


Figure 34: Geographic location of the 51 populations (red crosses) of the HGDP-CEPH diversity panel and the ir further grouping into 7 continental regions in which they were further grouped (dashed blue lines). Labels for continental regions are as in Table 8. World map modified from (Cann et al. 2002).

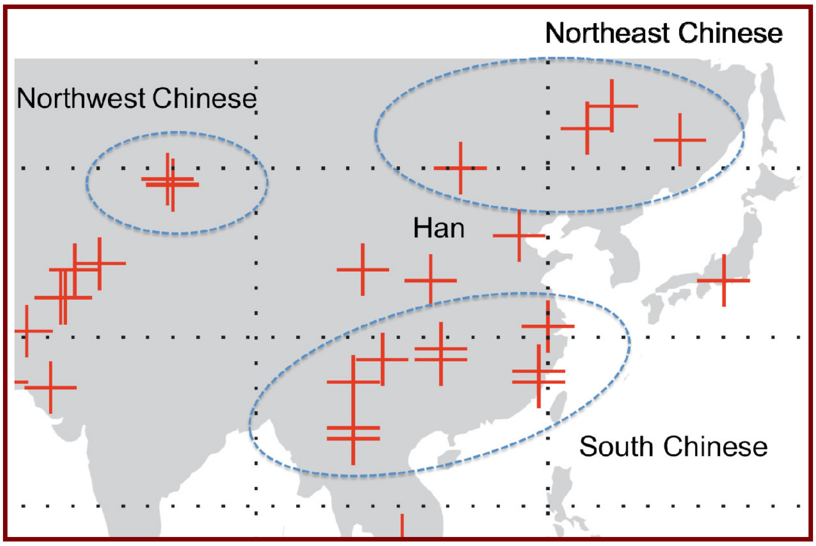


Figure 35: Partial view of the world map in Figure 34 showing the grouping of Chinese populations included in the HGDP-CEPH diversity panel. Red crosses indicate HGDP populations accounting for the original set of 51 populations, whereas dashed lines denote those combined to result in our 39 study populations. Non-Chinese populations are not labeled but shown for geographical reference.

Table 8: HGDP-CEPH diversity panel populations and sample sizes for the H971 subset.

Continental area	Country / region	HGDP population	Study population	N
Sub Saharan Africa (**SSAFR**)	South East Africa	Bantu	Bantu	19
	Central African Republic	Biaka Pygmies	Biaka Pygmies	27
	Democratic Rep. of Congo	Mbuti Pygmies	Mbuti Pygmies	13
	Senegal	Mandenka	Mandenka	24
	Namibia	San	San	6
	Nigeria	Yoruba	Yoruba	22
Middle East-North Africa (**MENA**)	Algeria (Mzab)	Mozabite	Mozabite	29
	Israel (Central)	Palestinian	Palestinian	50
	Israel (Negev)	Bedouin	Bedouin	47
	Israel (Carmel)	Druze	Druze	44
Europe (**EUR**)	France	French	French	28
	France	Basque	Basque	24
	Orkney Islands, UK	Orcadian	Orcadian	15
	Italy	Sardinian	Sardinian	28
	Italy	Bergamese	North Italian	21
	Italy	Tuscan		
	Russia Caucasus	Adygei	Adygei	17
	Russia	Russian	Russian	25
Central-South Asia (**CSASIA**)	Pakistan	Balochi	Balochi	24
	Pakistan	Brahui	Brahui	25
	Pakistan	Burusho	Burusho	25
	Pakistan	Hazara	Hazara	23
	Pakistan	Kalash	Kalash	24
	Pakistan	Makrani	Makrani	25
	Pakistan	Pathan	Pathan	24
	Pakistan	Sindhi	Sindhi	24
	North China	Uygur	Northwest Chinese	29
	North China	Xibo		
	North China	Tu		
East Asia (**EASIA**)	North China	Daur	Northeast Chinese	38
	North China	Hezhen		
	North China	Mongola		
	North China	Oroqen		
	South China	Dai	South Chinese	67
	South China	Lahu		
	South China	Miaozu		
	South China	Naxi		
	South China	She		
	South China	Tujia		
	South China	Yizu		
	China	Han	Han	44
	Siberia, Russia	Yakut	Yakut	25
	Cambodia	Cambodian	Cambodian	10
	Japan	Japanese	Japanese	29
Oceania (**OCE**)	Bougainville	Melanesian	Melanesian	13
	New Guinea	Papuan	Papuan	17
America (**AME**)	Mexico	Pima	Pima	14
	Mexico	Maya	Maya	22
	Colombia	Colombian	Colombian	7
	Brazil	Karitiana	Karitiana	14
	Brazil	Surui	Surui	9
Total		**51**	**39**	**971**

Besides the HGDP-CEPH diversity panel, additional smaller sets of populations were analyzed for gathering resequencing data in two specific candidate genes, namely *OR5I1* and *FOXI1*, presented in **Chapters 2 and 3** of the Results section, respectively. In the former we analyzed resequencing data for 20 European Americans and 19 African Americans available in the *OR5I1* gene (Nielsen et al. 2005) and provided by C. Bustamante (Cornell University); whereas in the latter we present results from resequencing the *FOXI1* gene in 20 European, 20 Asian, and 20 African samples obtained from Coriell Cell Repositories (Camden, NJ, USA). The ancestry of these individuals is the same as the CEU, ASN and YRI HapMap samples (see **Chapter 3's Materials and Methods** for details).

SNP Selection

Detailed information about the specific SNPs analyzed in each part of the project is given in Materials and Methods sections throughout the different Chapters of Results of this thesis. However, it is worth to describe here the common rationale for choosing SNPs that was applied to all candidate genes investigated in this work.

In order to capture most of the variation along each gene, including possible functional variants and neighbor regulatory elements, we densely covered all candidate region as follows: one SNP was selected every 5-10 Kb inside the gene and up to ± 30 Kb, slightly increasing marker density over the closest vicinity in both flanking regions. Then, one SNP was added every 40 Kb up to ± 200 Kb, resulting in ~400-Kb genotyped regions (see **Figure 36**). Preference was given to SNPs with a minor allele frequency (MAF) over 10%, which were compiled from HapMap (Release 20 Jan 2006) and dbSNP (Build 125 Oct 2005) databases. Additionally, most coding non-synonymous SNPs (CNS) and other functional SNPs were included regardless of their allele frequency or validation status. Predicted functional SNPs were identified using PupaSuite web-based SNP analysis tool (Conde et al. 2006). A total of 566 SNPs were successfully genotyped covering 30 candidate genes (see **Tables 5 to 7**).

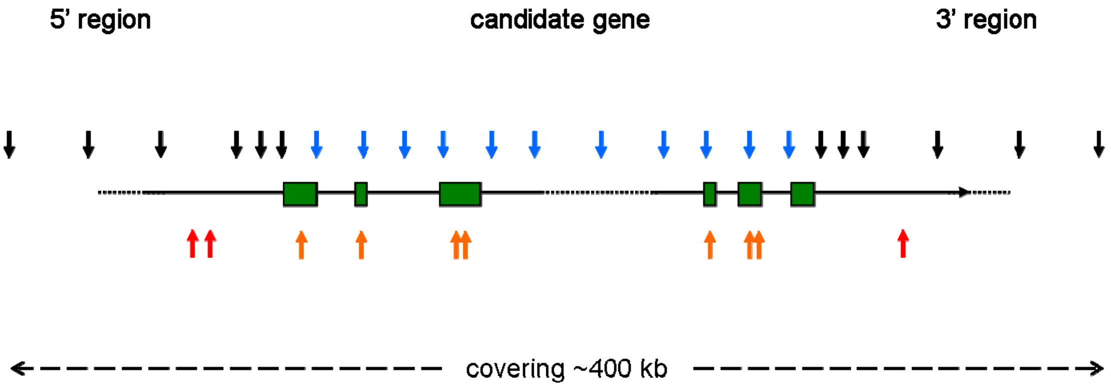


Figure 36: Schematic representation of a given candidate gene densely covered with SNPs according to the rationale applied in this work (see text for details). The horizontal arrow points the transcriptional direction of the gene, green squares represent exons, and dotted line between them a large intron. Blue arrows are SNPs inside the gene, whereas black arrows denote flanking SNPs, and red arrows represent coding and functional extra SNPs.

In order to capture the complete decay of LD around each candidate gene, and thus be able to successfully apply EHH-based methods to detect selection, the analyses on twelve 400-Kb candidate regions were extended up to 2 Mb using public genotype data for the Infinium Human Hap650Y BeadChip available in the same samples from the HGDP-CEPH diversity panel (Li et al. 2008). Genotypes were downloaded in bulk from the Stanford Human Genome Center website. Twelve gene-centered regions spanning ~2 Mb were delimited based on NCBI build 36.1 coordinates and extracted for further analysis. Although the array provides a substantial increase both in marker density and length coverage, it may not capture a substantial fraction of the functional variation in our candidate genes, so we decided to merge the SNPs from both 400-Kb and 2-Mb regions into a single mixed dataset (see **Table 9** for details). Genotypes from overlapping SNPs were considered only once after crosschecking for consistency. The average number of SNPs per candidate region increased to 460, and a total of 5,522 SNPs spanning ~24 Mb of the genome were analyzed.

Table 9: Intersection of our SNP data set with publicly available data in the same samples[a]

Gene region	Genotyped (400 Kb)	650K array (2 Mb)	Mixed dataset	Extra SNPs in (due to)	
				2 Mb (our SNPs)	400 kb (650K)
CA14	17	163	175	12	40
GFRA3	20	253	266	13	32
HDHD3	19	655	663	8	90
LHPP	26	642	655	13	162
MRC2	23	246	261	15	49
ADI1	21	663	679	16	84
OR2A14	16	357	371	14	61
OR5G1P	17	410	423	13	80
PTGER4	20	480	493	13	54
USP2	22	391	408	17	67
VPS37C	22	403	420	17	64
FOXI1	21	693	708	15	153
Total	**244**	**5356**	**5522**	**166**	**936**
Average	20	446	460	14	78

[a] The intersection of successfully genotyped samples (i.e., 961 and 1,043 respectively) resulted in 949 individuals in most of the cases. This subset of genes is limited to those in which the analysis was extended up to 2 Mb using publicly available genotype data from the 650K SNP array, namely the 11 fast-evolving genes (Chapter 1) and the *FOXI1* gene (Chapter 3).

SNP Genotyping

After applying the rationale for choosing SNPs described above, a total of 653 SNP markers were selected to genotype 1,064 DNA samples, thus an intermediate-throughput scale genotyping technology was chosen (i.e., SNPlex Genotyping System from Applied Biosystems). A total of 566 SNPs were successfully genotyped in at least 961 individuals of the H971 HGDP subset, resulting in a success rate of 86.7%.

The SNPlex technology is based on the Oligonucleotide Ligation Assay (OLA) (Landegren et al. 1988), which uses a set of three oligonucleotides, in combination with a thermostable Taq DNA ligase enzyme, to discriminate SNP alleles. The pair of allele-specific oligonucleotides differs by a 3'-terminal discriminatory nucleotide corresponding to the SNP alleles. The locus-specific oligonucleotide is designed to anneal immediately 3' to the target SNP. The oligonucleotides are allowed to hybridize to target DNA in the presence of the ligase, and when a pair of perfectly matched oligonucleotides is immediately juxtaposed to form a duplex with the target DNA, they are covalently ligated together. The number of ligation products is then increased to detectable levels by thermal cycling of the OLA reaction Finally, to discriminate allele-specific ligation products, their electrophoretic mobility is modified by adding fluorescently labeled probes, so they can be distinguished by running on a capillary sequencer. An overview of the genotyping pipeline is given in **Figure 37**.

The SNPlex genotyping platform can achieve relatively high levels of multiplexing, genotyping up to 48 SNPs in a single reaction (i.e., up to a 48-plex level). We designed multiple sets of OLA oligonucleotiedes following the SNPlex assay pipeline for optimizing multiplexing, which resulted in a total of 14 different pools of SNPlex probes. Allele separation was performed on an Applied Biosystems 3730 DNA analyzer. The system analysis software collected and provided raw data, and both quality metrics and allele calling were reviewed using GeneMapper Software 3.5 (see **Figure 38**).

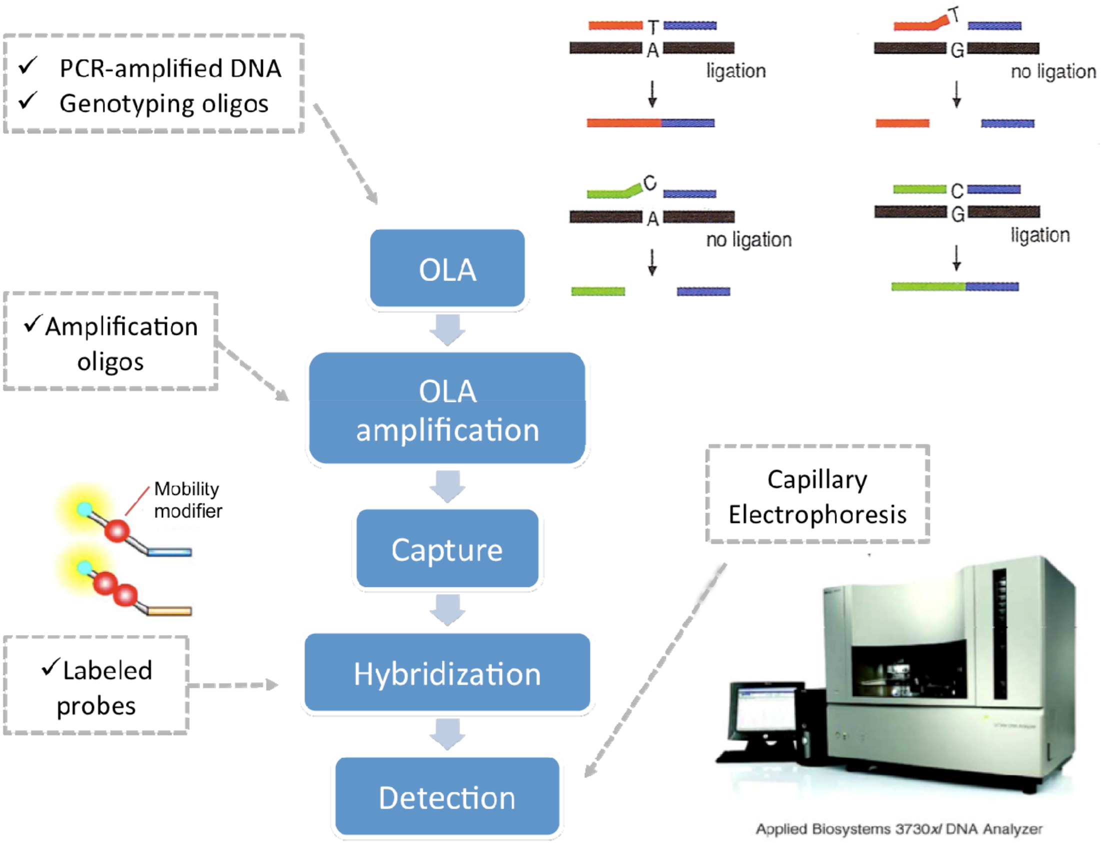


Figure 37: Genotyping pipeline of the Oligonucleotide Ligation Assay (OLA).*Top right*: A pair of allele-specific oligos (red/green) and a single locus-specific oligo (blue) are permitted to hybridize the target, which is heterozygous for the A/G SNP shown here (see text for details).

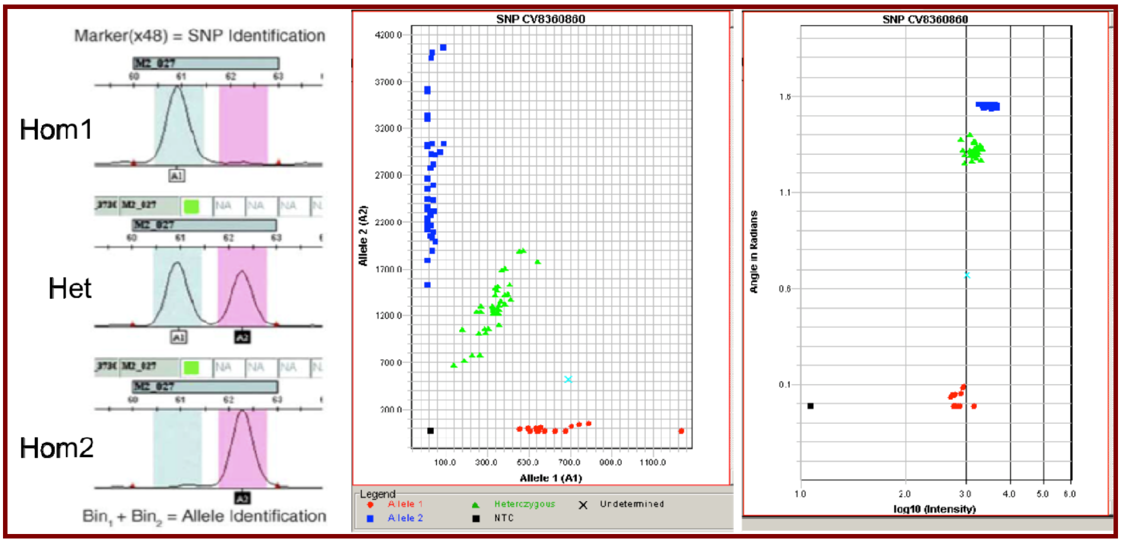


Figure 38: Allele calling (*left*) and genotype clustering with Cartesian (*middle*) and Polar (*right*) plotting using GeneMapper Analysis Software v3.5 (adapted from (De la Vega et al. 2005)).

RESULTS

Chapter 1: Interrogating fast-evolving genes for signatures of recent positive selection in worldwide human populations.

Andrés Moreno-Estrada, Kun Tang, Martin Sikora, Tomàs Marquès-Bonet, Ferran Casals, Arcadi Navarro, Francesc Calafell, Jaume Bertranpetit, Mark Stoneking, and Elena Bosch.

(Submitted to *Molecular Biology and Evolution*)

Interrogating fast-evolving genes for signatures of recent positive selection in worldwide human populations

To be submitted as Research Article

Andrés Moreno-Estrada[1], Kun Tang[2,3], Martin Sikora[1], Tomàs Marquès-Bonet[1,4], Ferran Casals[5], Arcadi Navarro[1,6], Francesc Calafell[1,7], Jaume Bertranpetit[1,7], Mark Stoneking[3] and Elena Bosch[1,7]

[1] Institut de Biologia Evolutiva (UPF-CSIC), Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona, Spain.

[2] CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China

[3] Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany.

[4] Department of Genome Sciences, University of Washington, Seattle, USA.

[5] Ste Justine Hospital Research Centre, Department of Pediatric, Faculty of Medicine, University of Montreal, Montreal, Quebec H3T 1C5, Canada

[6] Institució Catalana de Recerca i Estudis Avançats (ICREA) i UPF

[7] Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain.

Corresponding author: Elena Bosch, Institut de Biologia Evolutiva (UPF-CSIC), Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra,

C/ Dr. Aiguader 88, 08003 Barcelona, Spain.

Tel. +34 93 316 0841

Fax. +34 93 316 0901

E-mail: elena.bosch@upf.edu

Key words: Accelerated Evolution, Recent Positive Selection, SNP Data, Extended Haplotype Homozygosity, Population Differentiation, Human Genome Diversity Panel.

Running head: Recent selection on fast-evolving genes

Abstract

Different signatures of natural selection persist over varying time scales in our genome, revealing possible episodes of adaptative evolution during human history. Here, we identify genes showing signatures of ancestral positive selection in the human lineage and investigate whether those genes have been evolving adaptatively in extant human populations. Specifically, we compared more than 11,000 human genes with their orthologs in chimpanzee, mouse, rat and dog and applied a branch-site likelihood method to test for positive selection on the human lineage. Among the significant cases, a set of 11 genes were then further explored for signatures of recent positive selection using SNP data. We genotyped 223 SNPs in 39 worldwide populations from the HGDP Diversity panel and supplemented this information with available genotypes for up to 4,814 SNPs distributed along 2 Mb centered on each gene. After exploring the allele frequency spectrum, population differentiation and the maintainance of long unbroken haplotypes, we found signals of recent adaptative phenomena in only one of the 11 candidate gene regions. However, the signal of recent selection in this region may come from a different, neighbouring gene (*CD5*) rather than from the candidate gene itself (*VPS37C*). Our results show that positively-selected genes in the human lineage have not maintained their rapid evolutionary pace among human populations. Adaptation for human-specific and for population-specific traits thus involved different sets of genes.

Introduction

Identifying traits that have undergone positive selection during human evolution is essential to understand the adaptive events that have shaped our genomes. Whereas comparative genomics of closely-related species has shed light on the species-specific traits that set us apart from our closest living relatives, the genomic signature of recent adaptations can be directly detected from human population genetic data. The standard tool to detect signatures of selection in comparative data is the K_a/K_s ratio (also called d_N/d_S or ω in different analysis contexts) which expresses the ratio of nonsynonymous to synonymous substitutions in a given protein coding sequence. Genetic variants that were selected during the process of hominization are common to all humans and are detected by comparison with sequences from other primates. Moreover, methods have been developed to examine variation in the K_a/K_s ratio among lineages (Zhang, Kumar, and Nei 1997), among codon sites (Nielsen and Yang 1998; Yang et al. 2000), and to identify selection on individual codons along specific lineages (Zhang, Nielsen, and Yang 2005). Recently, several genome-wide efforts for identifying positively selected genes and/or functional categories enriched for such genes in the human and chimpanzee lineages have been conducted, yielding valuable insights into understanding human-specific traits (Clark et al. 2003; Bustamante et al. 2005; Nielsen et al. 2005; Arbiza, Dopazo, and Dopazo 2006).

The divergence that has accumulated in the human lineage since our separation from the chimpanzee occurred over the past 5 to 7 million years, and hence does not necessarily reflect recent selection that occurred after the origin of our species some 150,000 years ago. It is precisely within this evolutionary time scale that our ancestors dispersed from Africa to colonize most of the globe, and were challenged by many new selective pressures. Nonetheless, intraspecific diversity patterns within human populations may indeed reflect such modern adaptations, since distinctive signatures of recent selective sweeps (such as an overall reduction in genetic diversity, excess of high frequency derived alleles or long range haplotypes) stay imprinted in the genome for many tens to a few hundreds of thousands of years (Sabeti et al. 2006). Identifying genes affected by recent selective sweeps in human populations has gained great interest during the last few years, as they may help to explain population-specific adaptations. Along with the tremendous increase in the availability of SNPs in public databases (Hinds et al. 2005; Frazer et al. 2007) and the high-throughput methodologies that currently exist, new analytical methods aimed at detecting the footprint of selection from SNP data have been developed and widely applied in a number of candidate-gene (Sabeti et al. 2002; Bersaglieri et al. 2004; Hughes et al. 2008), path-related genes (Walsh et al. 2006; Han et al. 2007; Sikora et al. 2008) and genome-scanning studies (Akey et al. 2002; Kimura et al. 2007; Sabeti et al. 2007; Tang, Thornton, and Stoneking 2007; Williamson et al. 2007; Barreiro et al. 2008; Myles et al. 2008). Signatures of selection such as increased levels of population differentiation, unusual allele frequency spectra and elevated levels of linkage disequilibrium (LD) are usually examined and identified in comparison to a genome-wide empirical distribution or simulated data.

Recently, a remarkable amount of evidence for targets of recent selection in humans has been gained from a set of relatively new statistics especially design to detect long range haplotypes through the measure of the extended haplotype homozygosity (EHH), whose first implementation was introduced by the long range haplotype (LRH) test (Sabeti et al. 2002). These methods rely on the principle that positively selected alleles are expected to rise to high frequency rapidly enough that long range association with alleles at nearby loci will not have time to be erased by recombination. Different strategies have recently been developed in order to capture the extended LD around a putatively selected allele (or core haplotype) in a given population (Sabeti et al. 2002; Voight et al. 2006), and on particular alleles (Kimura et al. 2007; Sabeti et al. 2007; Hughes et al. 2008) or sites (Tang, Thornton, and Stoneking 2007) when comparing pairs of populations.

Here, we address whether the genes that were positively-selected during the evolution of the *Homo* lineage present any molecular signature of recent positive selection among human populations; that is, whether they continued to evolve at a fast, non-neutral pace. We first identified fastly-evolving human genes by comparing more than 11,000 coding sequences with their chimpanzee, mouse, rat and dog orthologs. For a subset of 11 significant cases, we analysed SNP data from a worldwide sample of 39 human populations belonging to the HGDP-CEPH diversity panel (Cann et al. 2002). Signals of recent positive selection were interrogated through population differentiation, allele frequency threshold analyses, and by applying two complementary EHH-based tests especially designed to detect both fixed (or nearly so) and intermediate frequency selected variants.

Materials and Methods

DNA samples

We analyzed the H971 standardized subset of the Human Genome Diversity Cell Line Panel (HGDP-CEPH) (Cann et al. 2002) recommended by Rosenberg (Rosenberg 2006), which contains no atypical, duplicated or deduced first-degree related individuals. In order to maximize sample sizes and after considering geographic and ethnic criteria, populations from the original panel were re-grouped into 39 population samples. In particular, we grouped Tuscans and Bergamese into North Italians; Dai, Lahu, Miaozu, Naxi, She, Tujia, and Yiku as South Chinese; Daur, Hezhen, Mongolian, and Oroqen as Northeast Chinese; and Tu, Uygur, and Xibo as Northwest Chinese. For some analyses, populations were further grouped into seven major geographical regions as in Moreno-Estrada et al. (2008): Sub-Saharan Africa (SSAFR), Middle East-North Africa (MENA), Europe (EUR), Central-South Asia (CSASIA), East Asia (EASIA), Oceania (OCE) and America (AME). Individual samples for which genotypes in any gene region analyzed failed for at least 50% of the SNPs were not considered. Two chimpanzee samples provided by the Barcelona Zoological Park were also genotyped and if their alleles matched one of the human states, were considered ancestral.

Selection of genes

We selected eleven genes (Table 1) that exhibited evidence of accelerated evolution in the human lineage after applying the following process. Around 11,000 human, chimpanzee, mouse, rat, and dog orthologous genes were retrieved from Ensembl (March 2006) and subsequently checked to have a unique best-reciprocal BLAST match in all five species. For the remaining 9,170 orthologous genes, we then performed multiple sequence alignments with ClustalW (Thompson, Higgins, and Gibson 1994) and applied a branch-site likelihood method (Zhang, Nielsen, and Yang

2005) to test for positive selection on the human branch in the underlying phylogeny of the five mammal species. A likelihood ratio test (LRT) was performed by comparing the likelihood values of two hypotheses, allowing variation in omega (ω) among lineages and sites at the same time. As the alternative hypothesis, we used the modified Branch-Site Model A as described in Zhang, Nielsen, and Yang (2005), in which ω is estimated for all branches of the phylogeny, allowing for sites with ω> 1 only in the human (foreground) branch. As the null hypothesis, we used the same Branch-Site Model A but with $\omega_2 = 1$ fixed in the human branch. The null and alternative hypotheses for this improved branch-site test 2 of positive selection (Zhang, Nielsen, and Yang 2005) and the corresponding Bayesian empirical inference of amino acid sites under positive selection (Yang, Wong, and Nielsen 2005) were performed using the codeml program implemented in the PAML package (Yang 1997). According to the LRT, 52 genes exhibited a signature of positive selection at the 0.01 level of significance and up to 88 at the 0.05 level. Significant cases were further explored for the presence of large gaps or excessive mismatches in the alignment. Additionally, the estimated ω value in the human foreground branch under the alternative model was checked to be greater than one, and we required at least one codon with a posterior probability of belonging to the site class of positive selection on the human lineage to be equal or greater than 95%; only 11 genes were left after applying these strict criteria.

SNPs

We selected one SNP every 5-10 kb inside each candidate gene and up to around 30 kb in both 5' and 3' flanking regions, plus additional SNPs every 40 kb up to around 200 kb in both flanking regions. Preference was given to SNPs with a minor allele frequency (MAF) over 10%, based on the HapMap (Release 20 Jan 2006) and dbSNP (Build 125 Oct 2005) databases. Additionally, most coding non-synonymous SNPs (CNS) within each candidate gene and other functional SNPs identified using the PupaSuite web-based SNP analysis tool (Conde et al. 2006) were included, regardless of their allele frequency or validation status. A total of 223 SNPs out of 270 (82.6%) were successfully genotyped using the SNPlex Genotyping System from Applied Biosystems following the manufacturer's standard protocol. Allele separation was performed on an Applied Biosystems 3730 analyzer and both quality metrics and allele calling were reviewed using GeneMapper Software 3.5. Data from the Infinium Human Hap650Y BeadChip genotyped in the HGDP-CEPH panel (Li et al. 2008) was downloaded in bulk from the Stanford Human Genome Center website. For each candidate region, SNP genotype data were then extracted for 2 Mb regions centered on each gene of interest and merged with our previously obtained genotypes in order to maximize marker density. For *CA14* it was not possible to obtain a centered 2 Mb region because the 650K Bead Chip lacked appropiate SNP coverage along ~500 Kb; we therefore obtained a slightly off-centered 2 Mb region for *CA14*. In total, 4,814 SNPs spanning ~22 Mb of the genome were analyzed (Table 1).

Data handling and analysis

Unless otherwise stated, data storage, quality control and data analysis were carried out using the SNPator web-based SNP data analysis platform (Morcillo-Suarez et al. 2008). For specific calculations and plotting purposes we used the R statistical software package (version 2.4.0, http://www.r-project.org).

Allele frequency analysis

Across every genomic region and population, we analyzed the distribution of the minor allele frequencies (MAF) and the derived allele frequencies (DAF) of the corresponding SNPs. The proportion of SNPs with allele frequencies higher or lower than a defined threshold (MAF < 0.10 for the MAF analysis and DAF > 0.80 for the DAF analysis) was calculated within sliding windows of 100 kb in size every 20 kb and plotted against distance over the full 2 Mb regions (see Figure 1). Thresholds were chosen to maximize sensitivity to selection as suggested by Walsh et al. (2006), and we required a minimum of 5 SNPs per window. Non-polymorphic SNPs in the overall 39 populations were not considered in the threshold analyses of minor and derived allele frequencies. However, SNPs fixed in a population but polymorphic elsewhere were counted as having MAF < 0.10 and DAF > 0.80 when applicable. The definition of minor allele was specific for each population rather than global and uniform across populations. We considered as outliers those regions in which multiple windows were found in the top 5% of the distribution of the respective proportions obtained independently for each population and considering all 2 Mb regions together. Given their small sample size (< 10 individuals), the San population was not included in the MAF threshold analysis. Ancestral states for all analyzed SNPs were obtained from the chimpanzee and/or the macaque genome sequences (panTro2, Mar. 2006 assembly and rheMac2, Jan. 2006 assembly, respectively). For 31 SNPs neither of the human alleles corresponded to the chimpanzee or macaque sequence and therefore these were not included the DAF threshold analysis. Genotyping of two chimpanzee samples confirmed the ancestral state of 192 SNPs.

F_{ST} calculation

The proportion of the variance explained by populational differences was measured through the molecular fixation index F_{ST}, by means of a locus by locus Analysis of Molecular Variance (AMOVA) (Excoffier, Smouse, and Quattro 1992) using the Arlequin software package version 3.11 (Excoffier, Laval, and Schneider 2005). Global F_{ST} values were obtained, taking into account all 39 worlwide populations studied as a single group. Empirical percentiles of global F_{ST} values were calculated based on the distribution of all F_{ST} values obtained over the eleven 2 Mb regions analyzed. We defined as outliers to be considered for further analysis those values within the top 5% of the empirical distribution.

Haplotype and long-range haplotype analysis

Haplotypes centered on each candidate gene were estimated using FastPHASE (Scheet and Stephens 2006). Linkage disequilibrium (LD) blocks were explored using Haploview version 4.1 (Barrett et al. 2005). The Long Range Haplotype (LRH) test (Sabeti et al. 2002) was carried out using the SWEEP software package (version 1.1). We defined cores as the longest non-overlapping core haplotypes with at least one SNP and not more than 20 SNPs. For each identified core haplotype, we calculated the EHH and the relative EHH (REHH) at a genetic distance of 0.3 cM in both directions and plotted these against the core haplotype frequency. Distributions of EHH and REHH values were obtained for all main geographical regions from the relevant populational phased haplotype data, considering together the eleven 2 Mb regions centered on the genes of interest. Core haplotypes were placed in 5% frequency bins and the respective EHH and REHH values were log-transformed for each bin in order to obtain approximately normally distributed values. Empirical p values for the LRH test were obtained by using the mean and standard deviation of the empirical distribution of the respective scores in each continental region. The LRH test was not performed in Oceania because the populations in the continent showed reduced background distribution. To account for multiple testing, we estimated the false positive discovery rate (pFDR) (Storey and Tibshirani 2003) and calculated the q value for the scores within each frequency bin using the package q value (version 1.1) for R. The q value for a particular p value is defined as the expected proportion of false positives among all significant p values when calling that p value significant. We used a q value cutoff of 0.05 for assigning significance.

In order to allow for multiple populational and/or continental comparisons of EHH, we slightly modified the method introduced by Tang, Thornton, and Stoneking (2007) based on the ln(Rsb) statistic. The integrated extended haplotype homozygosity of individual SNP sites (iEHHS or iES) was calculated for every SNP site and population directly from genotype data using a home-coded script (PopMX package by Tang K, Bauchet M, Theunert, unpublished data). Each iEHHS value was first normalized to the median of all values within each population, resulting in the EHHS' as following:

$$EHHS' = \frac{EHHS}{median(EHHS)}$$

EHHS' from individual population was then divided by the average EHHS' across populations weighted for population sizes as following:

$$XP - Rsb = \frac{EHHS'}{ave(EHHS')}$$

Where,

$$ave(EHHS') = \frac{\sum (EHHS'_i \cdot n_i)}{N}$$

Note that ave(EHHS') takes sample size into consideration. Here n_i refers to the number of individuals in a population *i* and *N* is the total number of samples used from the HGDP–CEPH panel. To estimate significance values for our results, we obtained a background distribution of XP-Rsb values for 642,690 genome-wide distributed SNPs using genotype data for the same panel generated elsewhere (Li et al. 2008). For each population and each SNP site, we obtained a p-value by ranking its XP-Rsb value across the whole genome in that population and determine its quantile. We then log transformed the p-values and plotted them against position within each 2 Mb region, searching for clusters of significant values inside or around our candidate genes at both population and continental levels. We calculated XP-Rsb for every SNP site and each population (vs. all other HGDP-CEPH populations) as well as for each main geographical region (versus the remaining regions represented in the panel).

Identification of functional variants

We limited the search of possible functional variants along the *VPS37C* genomic region to an LD block spanning ~420 kb, where the strongest signals of selection were concentrated. For this purpose, we explored *in silico* the functional relevance of all the genotyped SNPs in our mixed dataset as well as of all available HapMap SNPs within the same region (HapMap data Release 23a/phaseII). The PupaSuite web-based tool (Conde et al. 2006) was used to detect all SNPs with a potential phenotypic effect, including coding non-synonymous SNPs (CNS), SNPs disrupting miRNAs and their targets, as well as those SNPs located at triplexes or altering exonic splicing enhancers, exonic splicing silencers or transcription factor binding sites. The impact exerted by the amino acid substitution of each CNS was evaluated by means of Grantham's physicochemical distances (Grantham 1974), the damaging probabilities predicted by PolyPhen (Ramensky, Bork, and Sunyaev 2002) and by the codon-level selective constraints for prediction of functional altering mutations as estimated in PupasSuite (Arbiza et al. 2006). Haplotype extension of rs2229177 was explored on HapMap Phase II data using the Haplotter web-based application (Voight et al. 2006). We also downloaded HapMap data (release 22/phaseII) to look for tagSNPs in the *CD5* gene ±100 kb for both CEU and CHB+JPT samples using Tagger with the default parameters given by the authors (de

Bakker et al. 2005). Multi-species alignment of the *CD5* sequence was visualized within the Ensembl genome browser (release 50, July 2008) selecting all available sequences from 23 eutherian mammals. The amino acid sequence and structure information of the CD5 protein were obtained from the UniProt Knowledgebase (UniProtKB, entry P06127) and both the ModBase database (Pieper et al. 2004) and the Protein Data Bank (PDB, entry 1by2) (Berman et al. 2000), respectively.

Results

Selection of genes

Table 1 summarizes the eleven fast-evolving genes in the human lineage for which we analyzed SNP genotype data from 39 globally-distributed populations. Most of them had p-values smaller than 0.01 as determined by a likelihood ratio test of positive selection specifically in the human lineage on a phylogeny containing five mammal species (see Materials and Methods). None of the analyzed regions overlap, and although some map to the same chromosome, they can be treated as eleven independent genomic regions. Six of the selected genes had more than two codons putatively affected by selection-driven amino acid changes, and two cases (both of them olfactory receptors) showed up to seven codons putatively affected by selection, according to the posterior probability analysis of the Bayesian empirical inference. Interestingly, some of the positively selected codons contain non-synonymous polymorphic positions in extant human populations (rs6597801 in *LHPP*, rs2961160 and rs2961161 in *OR2A14,* rs754382 in *VPS37C,* and rs1943639 in *OR5G1P*). Moreover, seven of the 11 genes analyzed here were also inferred to have undergone positive selection in the human lineage when applying the same branch-site test of positive selection applied here but using human, chimpanzee and macaque gene trios (Bakewell, Shi, and Zhang 2007). Next, we explored these gene regions for different signatures of recent positive selection in worldwide human populations using SNP data.

Allele frequency threshold analyses

The distribution of minor and derived allele frequencies around a given genomic region may suggest particular selective pressures acting on it. In particular, an excess of high frequency derived alleles may indicate positive selection, whereas the presence of an excess of low frequency variants could reflect either purifying selection or a recent selective sweep.

In the MAF threshold analysis, for each population within each of the seven main geographical regions analyzed, we plotted the proportion of SNPs with MAF < 0.10 within multiple 100 kb sliding windows along 2 Mb regions centered on each candidate gene (Figures S1–S7). The general trend is characterized by a limited number of scattered outlier windows in different populations within each continental region along the 2 Mb regions analyzed. The American and Oceanian populations displayed many frequency fluctuations resulting in many consecutive windows with extreme proportions (either high or low) of SNPs with MAF < 0.10, a pattern which is probably due to their high levels of genetic drift and isolation. As for the genes of interest, only *VPS37C* and *OR2A14* concentrated an excess of rare alleles in East Asian, Oceanian and/or in American populations, respectively (Figures 1a and S5-S7).

In the DAF threshold analysis, for each population within each of the seven main geographical regions analyzed, we computed the proportion of SNPs with DAF > 0.8 within multiple 100 kb sliding windows along 2 Mb regions centered on each candidate gene (Figures S8-S14). Again, there are several clusters of windows with an excess of high-frequency derived alleles along the 2 Mb regions analyzed but just a few of them seem to involve the positively-selected genes. Within those, both the *VPS37C* and *USP2* genomic regions stand out for displaying in several populations a significant excess of high-frequency derived alleles, either in the genes or just nearby. The strongest signals for *VPS37C* are found in East Asia (Figure 1b) and in some Central South Asian populations (Figure S11). Without clearly including the *VPS37C* gene, the same pattern of an excess of high-frequency derived alleles is detectable in almost any population outside Sub-Saharan Africa, extending almost 400 kb from the 3’ flanking region. The signal for the *USP2* genomic region is found in populations from all main geographical regions except Oceania and America, and in all cases involves the 3’ region of the *USP2* gene but not the candidate itself.

Population differentiation

Local adaptation may cause unusually large allele frequency differences between populations at the selected loci, and consequently accentuate their levels of population differentiation. Here, we used F_{ST} to measure differentiation among all 39 worldwide populations for the 4,814 SNPs distributed across the eleven 2 Mb regions analyzed (Figure S15). None of the candidate genes have unusually high F_{ST} values, either in the gene or nearby, except *VPS37C* (Figure 1c). A total of 64 SNPs above the 95th percentile (23.3% of the top 5% values) were found in the 2 Mb region centered on *VPS37C.* The highest individual $\mathrm{F_{ST}}$ value was 0.4252 (rs17156025) and the average $\mathrm{F_{ST}}$ within the 64 highly differentiated SNPs was 0.2889.

Long unbroken haplotypes

Recent selective sweeps can produce a distinctive signature on the haplotype structure of chromosomes consisting of an allele (or haplotype) that has both high frequency and long-range associations with alleles at nearby loci (Sabeti et al. 2006). In order to try to detect such a signature in the candidate regions, we applied two complementary approaches based on the Extended Haplotype Homozygosity (EHH) measure. The first approach compares the EHH decay between the alleles (EHHA) of a site or core-haplotype within a given population and has strong power for identifying alleles that have been driven to intermediate frequencies during a recent selective sweep (Sabeti et al. 2002). In contrast, the second approach aims to detect nearly or recently completed local selective sweeps by comparing the EHH profile at individual SNP sites (EHHS) between populations (Tang, Thornton, and Stoneking 2007). As to the first approach, we applied the long range haplotype (LRH) test (Sabeti et al. 2002) by measuring for each core haplotype detected in our data the relative EHH (REHH) at a genetic distance of 0.3 cM in both directions from each core. Figure 2 shows the distributions of REHH values versus frequency for all of the populations analyzed within each main geographical region (except Oceania). Table 2 lists the corresponding significant core haplotypes after correction for multiple testing. Both Europe and Middle East-North Africa presented two high frequency core haplotypes as outliers (Figure 2). Three of them remained significant after multiple test correction (Table 2), but none of them mapped directly upon any of the candidate genes. However, the long range homozygosity associated with the significant core haplotype found in Middle East-North Africa is mantained near the *VPS37C* gene (Figure 3). The corresponding haplotype bifurcation plots for the two main haplotypes found in the core show unusual long range LD for the ACG core, given its frequency (0.821). Notably, the strongest signal for this significant core is reached at 0.42 cM where the REHH goes up to 46.4 (Figure S16). Central South Asia and East Asia showed three significant outliers but in low-frequency bins (Figure 2). Interestingly, two of those involved the candidate *LHPP* gene. Given their low frequency the maintenance of these haplotypes over the region is less clear when looking at the bifurcation and EHH decay patterns (data not shown); nonetheless they remained significant inside their frequency bins, which could reflect a partial ongoing selective sweep on the way to higher frequencies.

An obvious caveat of the previous analysis is that the intra-population comparison has low power when the selected allele variant is at high frequency, and becomes impossible when the variant is fixed. For this reason, in our second approach, we applied a slight modification of the ln(Rsb) statistic developed by Tang Thornton, and Stoneking (2007) designed to detect local selective sweeps by means of inter-population comparisons of EHH. Here we are analyzing 39 different populations and a minimum of seven groups when pooling populations into their main geographical regions, a number for which the ln(Rsb) statistic was not initially designed. To tackle this problem we modified the original formulation to XP-Rsb by comparing each individual population's iEHHS against a weighted cross-population average for each SNP position of the 2 Mb genomic regions (see details in Materials and Methods). Figure S17 shows the –log p-value of XP-Rsb along the eleven 2 Mb genomic regions analyzed for each main geographical region studied. Although there are some clusters of significant p-values, only two were located near any of the candidate genes. In particular, East Asia showed a significant EHH differentiation pattern when compared to the other geographic regions around the *GFRA3* and *VPS37C* gene regions (Figure 1d). In order to identify which population(s) within each geographic region might account for these signals, we also computed XP-Rsb between the 39 worldwide populations. Detailed results for the eleven full 2 Mb genomic regions are shown in Figures S18–S28; previously observed signals at the regional level could be attributed to specific populations, and some new signals were detected. For example, Cambodians, Han and Japanese are behind the EASIA signal previously observed in the *GFRA3* gene region. As for the *VPS37C* gene (Figure 1e), we found that significant XP-Rsb values were obtained in this gene region for three out of six East Asian populations, namely North East China, Han and Japanese, with Han Chinese accounting for most of the significant values. Not surprisingly, the signal is less significant than the one observed at the regional level, since the latter is decomposed into different individual population signals. New signals were also found for South Chinese in *HDHD3*, Pathan and Pima in *LHPP* and Japanese and Yakut in *OR5G1P*. Despite encompassing the genes of interest, these last cases have their highest significance values far outside them.

Insights on the VPS37C genomic region

Out of the eleven 2 Mb regions centered on our candidate genes, *VPS37C* consistently exhibits significant signatures of positive selection, especially in Asians. Most of the signals extend along more than 0.5 Mb and comprise several genes besides *VPS37C* (see top part of Figure 1). In order to identify which allelic variants could be responsible for the observed pattern, we first characterized the haplotype composition in this 0.5 Mb region, and then searched for variants with functional relevance on the putatively selected haplotype. In particular, we focused our analysis on a ~420 Kb region of relatively strong linkage disequilibrium, delimited by two hotspots of recombination, containing 54 SNPs. The haplotype frequency distribution across the 39 worldwide populations analyzed for this narrowed *VPS37C* region (Figure 4) reveals a total of 692 different haplotypes. While up to 80% of the Sub-Saharan African haplotypes were found in single chromosomes and most of the Eurasian populations had 10-20% unique haplotypes, one particular 54-SNP based haplotype stands out as having relatively high frequencies in North West China (59 %) and most East Asian (67 %) and American populations (60 %).

We functionally characterized not only the 54 SNPs contained in the analyzed dataset, but also all available SNPs in HapMap (Frazer et al. 2007) within the same region. In order to explain the observed significant pattern of selection any potentially causative genetic variant should be: i) functionally relevant, ii) particularly frequent in Asians but not elsewhere and iii) embedded within the extended haplotype in the major allele state. A total of 12 coding non-synonymous SNPs (CNS) affecting seven different genes were found in the target region, most of which were not typed in our mixed dataset but in HapMap. Table 3 summarizes their allele frequencies in the HapMap populations, the amino acid replacements they involve, and their inferred functional effects as predicted by different methods (see Materials and Methods). Only one CNS (rs2229177 in *CD5*) showed high frequencies in Asians with intermediate frequencies elsewhere, and a relevant functional effect (predicted as possibly damaging by PolyPhen and pathological by PupaSuite). Since this SNP was not genotyped in the HGDP panel, we explored Haplotter Phase II data for rs2229177 and confirmed that the derived state (T) sits in a long unbroken haplotype that is maintained at very high frequencies for approximately 400 kb in the Asian sample (data not shown). Moreover, taking advantage of the strong linkage disequilibrium in the region, we looked for tagSNPs capturing rs2229177 variation in the HapMap populations. Several SNPs tag rs222917 with $r^2 = 1.0$, four of which (i.e. rs4245224, rs10897141, rs610777 and rs628831) were typed in the HGDP panel. All four major alleles (as for Asian populations) are embedded within the main 54 SNP based haplotype found in Asians. Moreover, two of them (i.e. rs4245224 and rs628831) reproduced exactly the same derived allele frequencies of rs2229177 from the four HapMap populations in our equivalent samples from the HGDP-CEPH panel (i.e. Yoruba, French, Han and Japanese) which allows us to infer the possible worldwide frequency distribution of this CNS. When searching for other SNP functional categories (see Materials and methods for details), we compiled a total of 62 additional SNPs with potential phenotypic effects (data not shown) but only one, rs1787904, appeared to be differentiated at high frequencies in Asians (CHB: 0.988, JPT: 1, YRI: 0.542 and CEU: 0.567) and linked to the putatively selected haplotype (data not shown). Although this substitution maps within an intron of the *VPS37C* gene, it is located in a triplex sequence (Goni, de la Cruz, and Orozco 2004; Conde et al. 2006) that is within 10 Kb from the 3' end of the *CD5* gene, and hence could modify *CD5* expression.

Discussion

We have addressed the question of whether fast-evolving genes in the human lineage are still evolving adaptatively within extant human populations. To do so, we looked for signatures of recent adaptation in worldwide human populations along eleven 2 Mb genomic regions, each centered on a gene that we identified as fast-evolving on the human lineage. Most of the candidate gene regions did not show clear evidence of undergoing recent selection within the worldwide human diversity panel. The absence of recent signatures of selection on most positively-selected genes may imply that once they had acquired a specific human function, they became constrained functionally against further change. Only one genomic region, *VPS37C,* showed significant signals across all the signatures of selection we explored. The observed pattern in this region is consistent with the action of recent positive selection in East Asian populations. Despite the strong evidence in favour of a selective sweep occurring along this genomic region, it is difficult to pinpoint the source, since there were different clusters of significant signals across a ~ 0.5 Mb region. While there are highly differentiated loci throughout the candidate region, some signals (such as those displayed by the MAF and DAF threshold analyses or the presence of significant core haplotypes in the LRH test) do appear to be concentrated from the vicinity of the *VPS37C* gene up to ~ 400 Kb upstream. On the contrary, the highest concentration of significant p-values in the XP-Rsb analysis starts at *VPS37C* but only extends 100 Kb downstream. Despite the limitations of these methods to accurately locate the target of selection, the observed pattern does suggest that it might be somewhere in or around the *VPS37C* gene and that part of the extended signal is due to high LD.

In agreement with these results, a previous genome-wide scan reported the *VPS37C* gene region among the 101 regions with the strongest evidence for a recent selective sweep in Chinese (Williamson et al. 2007). This study used a composite likelihood ratio test, which provides fine-scale estimates of the position of the selected site, and which for this region was mapped to a 200-SNP window centered on the *VPS37C* gene. As suggested by the same authors (Williamson et al. 2007), since the VPS37C protein is recruited by HIV and other viruses to promote viral budding from infected cells (Stuchell et al. 2004; Eastman et al. 2005), it might play an important role in pathogen interactions. However, the identification of *VPS37C* as the actual gene responsible for the signal of selection in this region remains to be confirmed, and all known genes within ±100 Kb are not rejected as alternative candidates.

In a detailed analysis of the haplotype composition of the region we identified a paticular 54-SNP based haplotype at relatively high frequencies in Asia and America. This haplotype spans a ~420 Kb block, and encompasses all of the different signals observed in this genomic region. The functional characterization of all known allelic variants linked to this putatively selected haplotype suggested two candidates, a non-synonymous coding SNP located in the last exon of *CD5* (rs2229177) and a substitution altering a triplex-forming target sequence within the 3' end regulatory region (rs1787904). The *CD5* gene is located just 18 Kb from *VPS37C* and codes for a 495-amino-acid-long transmembrane receptor expressed in the T-cell surface. Topologically, it comprises a large extracellular domain (amino acids 25-372) containing three repeats of a cysteine-rich region (SRCR domains), followed by a single-pass transmembrane domain (amino acids 373-402) and a short cytoplasmic region (amino acids 403-495). The aforementioned CNS (rs2229177) leads to an Ala-Val substitution at position 471 in the cytoplasmatic part of the protein, which has been reported as essential for the function of the receptor (Pena-Rossi et al. 1999; Bhandoola et al. 2002). Alanine is encoded by the ancestral state (C) while Valine (encoded by T) is the derived state, which characterises the major form of the protein in Asian and American populations. In contrast with the much more variable extracellular region, this cytoplasmatic part of the receptor is highly conserved across species, based on multiple sequence alignments. Berland and Wortis (2002) reported that only 5 out of the 96 amino acids of this region differ among five mammalian sequences (human, mouse, sheep, bovine and rat). Moreover, all other eutherian mammals (23 species compared) conserved the ancestral state at the polymorphic rs2229177 position. Despite the availability of several 3D-structure models for this receptor, none of them includes the cytoplasmatic region where the A471V substitution is located. The lack of a complete experimental template prevents any conclusive prediction of the structural or functional impact of this substitution.

The *CD5* gene codes for a glycoprotein that acts as a transmembrane receptor in regulating T-cell proliferation. Specifically, CD5 functions as a negative regulator of T-Cell Receptor (TCR) signaling during intrathymic T cell development. Experimental studies have reported that CD5 mediated down-regulation does not require the CD5 extracellular domain and, consequently, does not involve CD5 binding of an extracellular ligand (Bhandoola et al. 2002). In contrast, the cytoplasmic portion of the molecule is required to act as an inhibitory receptor (Pena-Rossi et al. 1999). It has been pointed out that autoimmune disorders may result from the disruption of inhibitory receptors, particularly in their conserved intracellular motifs which are responsible for transducing signals to distinct pathways (Ravetch and Lanier 2000). Additional evidence for a functional role for rs2229177 (A471V) comes from a genetic association study in which it was shown that homozygosity for the ancestral A allele in A471V is associated with a poorer prognosis in patients of chronic lymphocytic leukemia (CLL) (Sellick et al. 2008). Given the function of the *CD5* gene and its role in the immune system physiopathology, it is tempting to speculate a possible protective effect for the putatively selected haplotype in Asians, although the exact mechanism by which the silencing of such a regulatory receptor would have been favoured by selection remains elusive. Additionally, other unknown variation linked to the same haplotype cannot be discarded as the actual functional variant responsible for the observed signals of selection.

Our results also demonstrate how both EHH-based approaches complement each other, as predicted by their estimated power to detect selection depending on the frequency of the selected allele in the population (Sabeti et al. 2007). Here, we found signals for the LRH test in populations where the putatively selected haplotype is segregating at intermediate frequencies (i.e. Middle East-North Africa), while for XP-Rsb we found evidence for selection involving the same haplotype in East Asia, where it has nearly reached fixation. Finally, we illustrate a case in which the ancient selective event of *VPS37C* during early human evolution and its apparent recent selective sweep are actually two independent phenomena. These results show that for the genes with the clearest signs of adaptation in the human lineage, their non-neutral evolution ended with the advent of the human species. Population-specific adaptation in humans is an independent process, involving different sets of genes than those that participated in defining our species.

Acknowledgments

This research was funded by grants BFU2005-00243 and BFU2008-010406/BMC awarded by Dirección General de Investigación, Ministerio de Educación y Ciencia (Spain) and by the Direcció General de Recerca, Generalitat de Catalunya (2005SGR00608). A Moreno-Estrada was supported by a fellowship from the Consejo Nacional de Ciencia y Tecnología (CONACyT), México; and received a travel grant from the Boehringer Ingelheim Foundation, Germany. SNP genotyping services were provided by the Spanish "Centro Nacional de Genotipado" (www.cegen.org).

Electronic Database Information

The Uniform Resource Locators (URLs) for data presented herein are as follows:

Stanford HGDP SNP Genotyping Data: http://shgc.stanford.edu/hgdp/index.html

SNPator web application: http://bioinformatica.cegen.upf.es

R statistical software package: http://www.r-project.org

SWEEP software package: http://www.broad.mit.edu/mpg/sweep/index.html

PolyPhen: http://genetics.bwh.harvard.edu/pph

SNPeffect: http://snpeffect.vib.be

Haplotter: http://hg-wen.uchicago.edu/selection/haplotter.htm

PupaSuite: http://pupasuite.bioinfo.cipf.es/

Tagger: http://www.broad.mit.edu/mpg/tagger/server.html

UniProtKB: http://www.uniprot.org/

Protein Data Bank: http://www.pdb.org/pdb/home/home.do

ModBase database: http://modbase.compbio.ucsf.edu/modbase-cgi/index.cgi

HapMap Genome Browser: http://www.hapmap.org/

References

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, Dopazo J, Dopazo H. 2006. Positive selection, relaxation, and acceleration in the evolution of the human and chimp genome. PLoS Comput Biol. 2:e38.

Arbiza L, Duchi S, Montaner D, Burguet J, Pantoja-Uceda D, Pineda-Lucena A, Dopazo J, Dopazo H. 2006. Selective pressures at a codon-level predict deleterious mutations in human disease genes. J Mol Biol. 358:1390-1404.

Bakewell MA, Shi P, Zhang J. 2007. More genes underwent positive selection in chimpanzee evolution than in human evolution. Proc Natl Acad Sci USA 104:7489-7494.

Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L. 2008. Natural selection has driven population differentiation in modern humans. Nat Genet. 40:340-345.

Barrett JC, Fry B, Maller J, Daly MJ. 2005. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21:263-265.

Berland R, Wortis HH. 2002. Origins and functions of B-1 cells with notes on the role of CD5. Annu Rev Immunol. 20:253-300.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. Nucleic Acids Res. 28:235-242.

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

Bhandoola A, Bosselut R, Yu Q, Cowan ML, Feigenbaum L, Love PE, Singer A. 2002. CD5-mediated inhibition of TCR signaling during intrathymic selection and development does not require the CD5 extracellular domain. Eur J Immunol. 32:1811-1817.

Bustamante CD, Fledel-Alon A, Williamson S, et al. (14 co-authors). 2005. Natural selection on protein-coding genes in the human genome. Nature 437:1153-1157.

Cann HM, de Toma C, Cazes L, et al. (41 co-authors). 2002. A human genome diversity cell line panel. Science 296:261-262.

Clark AG, Glanowski S, Nielsen R, et al. (17 co-authors). 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science 302:1960-1963.

Conde L, Vaquerizas JM, Dopazo H, Arbiza L, Reumers J, Rousseau F, Schymkowitz J, Dopazo J. 2006. PupaSuite: finding functional single nucleotide polymorphisms for large-scale genotyping purposes. Nucleic Acids Res. 34:W621-625.

de Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D. 2005. Efficiency and power in genetic association studies. Nat Genet. 37:1217-1223.

Eastman SW, Martin-Serrano J, Chung W, Zang T, Bieniasz PD. 2005. Identification of human VPS37C, a component of endosomal sorting complex required for transport-I important for viral budding. J Biol Chem. 280:628-636.

Excoffier L, Laval G, Schneider S. 2005. Arlequin ver. 3.0: An integrated software package for population genetics data analysis. Evol Bioinform Online 1:47-50.

Excoffier L, Smouse PE, Quattro JM. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131:479-491.

Frazer KA, Ballinger DG, Cox DR, et al. (233 co-authors). 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851-861.

Goni JR, de la Cruz X, Orozco M. 2004. Triplex-forming oligonucleotide target sequences in the human genome. Nucleic Acids Res. 32:354-360.

Grantham R. 1974. Amino acid difference formula to help explain protein evolution. Science 185:862-864.

Han Y, Gu S, Oota H, Osier MV, Pakstis AJ, Speed WC, Kidd JR, Kidd KK. 2007. Evidence of positive selection on a class I ADH locus. Am J Hum Genet. 80:441-456.

Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR. 2005. Whole-genome patterns of common DNA variation in three human populations. Science 307:1072-1079.

Hughes DA, Tang K, Strotmann R, Schoneberg T, Prenen J, Nilius B, Stoneking M. 2008. Parallel selection on TRPV6 in human populations. PLoS ONE 3:e1686.

Kimura R, Fujimoto A, Tokunaga K, Ohashi J. 2007. A practical genome scan for population-specific strong selective sweeps that have reached fixation. PLoS ONE 2:e286.

Li JZ, Absher DM, Tang H, et al. (11 co-authors). 2008. Worldwide human relationships inferred from genome-wide patterns of variation. Science 319:1100-1104.

Morcillo-Suarez C, Alegre J, Sangros R, et al. (17 co-authors). 2008. SNP analysis to results (SNPator): a web-based environment oriented to statistical genomics analyses upon SNP data. Bioinformatics 24:1643-1644.

Moreno-Estrada A, Casals F, Ramirez-Soriano A, Oliva B, Calafell F, Bertranpetit J, Bosch E. 2008. Signatures of selection in the human olfactory receptor OR5I1 gene. Mol Biol Evol. 25:144-154.

Myles S, Tang K, Somel M, Green RE, Kelso J, Stoneking M. 2008. Identification and analysis of genomic regions with large between-population differentiation in humans. Ann Hum Genet. 72:99-110.

Nielsen R, Bustamante C, Clark AG, et al. (12 co-authors). 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol. 3:e170.

Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929-936.

Pena-Rossi C, Zuckerman LA, Strong J, Kwan J, Ferris W, Chan S, Tarakhovsky A, Beyers AD, Killeen N. 1999. Negative regulation of CD4 lineage development and responses by CD5. J Immunol. 163:6494-6501.

Pieper U, Eswar N, Braberg H, et al. (15 co-authors). 2004. MODBASE, a database of annotated comparative protein structure models, and associated resources. Nucleic Acids Res. 32:D217-222.

Ramensky V, Bork P, Sunyaev S. 2002. Human non-synonymous SNPs: server and survey. Nucleic Acids Res. 30:3894-3900.

Ravetch JV, Lanier LL. 2000. Immune inhibitory receptors. Science 290:84-89.

Rosenberg NA. 2006. Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. Ann Hum Genet. 70:841-847.

Sabeti PC, Reich DE, Higgins JM, et al. (17 co-authors). 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832-837.

Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES. 2006. Positive natural selection in the human lineage. Science 312:1614-1620.

Sabeti PC, Varilly P, Fry B, et al. (244 co-authors). 2007. Genome-wide detection and characterization of positive selection in human populations. Nature 449:913-918.

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

Sellick GS, Wade R, Richards S, Oscier DG, Catovsky D, Houlston RS. 2008. Scan of 977 nonsynonymous SNPs in CLL4 trial patients for the identification of genetic variants influencing prognosis. Blood 111:1625-1633.

Sikora M, Ferrer-Admetlla A, Mayor A, Bertranpetit J, Casals F. 2008. Evolutionary analysis of genes of two pathways involved in placental malaria infection. Hum Genet. 123:343-357.

Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. Proc Natl Acad Sci USA 100:9440-9445.

Stuchell MD, Garrus JE, Muller B, Stray KM, Ghaffarian S, McKinnon R, Krausslich HG, Morham SG, Sundquist WI. 2004. The human endosomal sorting complex required for transport (ESCRT-I) and its role in HIV-1 budding. J Biol Chem. 279:36059-36071.

Tang K, Thornton KR, Stoneking M. 2007. A New Approach for Using Genome Scans to Detect Recent Positive Selection in the Human Genome. PLoS Biol. 5:e171.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673-4680.

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

Walsh, EC, Sabeti P, Hutcheson HB, et al. (16 co-authors). 2006. Searching for signals of evolutionary selection in 168 genes related to immune function. Hum Genet. 119:92-102.

Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, Nielsen R. 2007. Localizing recent adaptive evolution in the human genome. PLoS Genet. 3:e90.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 13:555-556.

Yang Z, Nielsen R, Goldman N, Pedersen AM. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431-449.

Yang Z, Wong WS, Nielsen R. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. Mol Biol Evol. 22:1107-1118.

Zhang J, Kumar S, Nei M. 1997. Small-sample tests of episodic adaptive evolution: a case study of primate lysozymes. Mol Biol Evol. 14:1335-1338.

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol. 22:2472-2479.

Table 1. Summary of analyzed regions

Gene[a]	Chr	Size (kb)	Biological function	LRT[b] P value	Positively selected codons[c]	SNP coverage	
						400 kb[d]	2 Mb[e]
GFRA3	5	22.2	Glycosyl-phosphatidyl inositol receptor	0.0012	304	20	266
PTGER4	5	13.8	Skin immune responses	0.0014	8, 235	20	493
HDHD3	9	2.6	Glycobiological function	0.0014	28	19	663
LHPP	10	152.3	Hydrolase activity	0.0017	69, 78, **94** (rs6597801)	26	655
CA14	1	7.3	Hydration of carbon dioxide	0.0032	131, 153, 204	17	175
OR2A14	7	0.9	Olfactory receptor	0.0077	4, 14, 75, **132** (rs2961160), 157, **163** (rs2961161), 256	16	371
VPS37C	11	31.2	Endosomal transport important for viral budding	0.0091	70, **197** (rs754382), 248, 275	22	420
MRC2	17	64.9	Binding and transmission of HIV	0.0111	158	23	261
USP2	11	25.4	Cell cycle regulation and apoptosis	0.0161	140	22	408
OR5G1P	11	0.8	Olfactory receptor	0.0168	4, 43, 48, 54, 108, **183** (rs1943639), 193	17	423
ADI1	2	21.6	Suppression of tumor cell invasion in tissues	0.0256	62	21	679

[a] Gene of interest

[b] Likelihood Ratio Test

[c] According to the Bayes Empirical Bayes analysis, all codons with posterior probability greater than 95% for being positively selected in the human lineage are listed. Codons involving polymorphic positions within human populations are in bold and the corresponding SNP ID presented in brackets

[d] Number of genotyped SNPs covering a 400 kb region centered on the respective gene

[e] Total number of analyzed SNPs in 2 Mb around each gene including both our genotyped SNPs and those publicly available from the 650K SNP array typed on the HGDP panel (Li et al. 2008).

Table 2. Core haplotypes with significant REHH values across the eleven 2 Mb genomic regions analyzed

Genomic region[a]	Geographical region	Genes in core region [b]	Distance (cM)[c]	Core haplotype	REHH	Frequency	*P* value	*q* value
VPS37C	MENA	*C11orf79, C11orf66, SYT7*	0.33	ACG [d]	24.87	0.821	2×10^{-4}	0.0124
VPS37C	EUR	*C11orf11, C11orf9, C11orf10, FEN1*	0.30	A[e]	34.52	0.735	0.8×10^{-4}	0.0080
PTGER4	EUR		0.26	A[f]	18.71	0.605	1.6×10^{-4}	0.0342
OR2A14	CSASIA	*TPK1*	0.30	TCT[g]	22.79	0.197	0.6×10^{-4}	0.0191
LHPP	CSASIA	*LHPP*	0.35	CGTC[h]	27.12	0.183	0.2×10^{-4}	0.0153
LHPP	EASIA	*LHPP, FAM53B*	-0.30	AGGAGGGA[i]	25.76	0.114	0.8×10^{-4}	0.0487

[a] Genomic regions are identified with the name of the candidate gene they contain

[b] Genes within ±100 kb around the core are considered

[c] Genetic distance (cM) from the core at which the signal has been captured. (–) indicates downstream direction, otherwise upstream

[d] rs3019187, rs2957858, rs12295977

[e] rs174534

[f] rs1876142

[g] rs6946827, rs17287011, rs990282

[h] rs7917600, rs7070581, rs4962607, rs11245137

[i] rs12411439, rs3781458, rs1006368, rs3781453, rs17152175, rs7099298, rs3781452, rs6597848

Table 3. Functional characterization and HapMap frequencies for the coding non-synonymous SNPs present in the 54-SNP *VPS37C* region

SNP[a]	Position[b]	Gene	Alleles[c]	Derived allele frequencies				Aa	Aa	Grantham	PolyPhen		Phenotypic
				YRI	CEU	CHB	JPT	change	pos	Distance[d]	prediction[e]	score[f]	effect[g]
rs2241002	60643489	*CD5*	C/**T**	0.300	0.186	0.044	0.045	P/L	224	98	PR D	2.547	Pathological
rs637186	60649182	*CD5*	G/**A**	0	0.083	0	0	H/R	461	29	benign	0.4	
rs2229177	60649811	*CD5*	C/**T**	0.475	0.576	0.988	1	A/V	471	64	PO D	1.542	Pathological
rs4297482	60656155	*VPS37C*	A/**C**	0	0	0	0	S/A	261	99	benign	1.323	Pathological
rs754382	60656343	*VPS37C*	C/**T**	0.058	0.300	0.011	0	L/S	198	145	benign	1.441	
rs3750982	60783066	*VWCE*	G/**C**	0	0	0	0	P/R	842	103	PR D	2.074	
rs2260655	60865550	*DAK*	G/**A**	0.342	1	0.989	0.966	A/T	185	58	benign	0.668	Pathological
rs11605407	60875103	*CYBASC3*	A/**C**	0	0	0	0	V/G	214	109	PO D	2.64	Pathological
rs11557691	60935017	*CPSF7*	G/**T**	0	0	0	0	D/Y	464	160	benign	N/A	Pathological
rs1064377	60939747	*CPSF7*	G/**C**	0	0	0	0	A/P	351	27	benign	1.452	Pathological
rs11230707	61010417	*C11orf66*	C/**A**	0.142	0	0	0	T/N	238	65	PO D	1.536	
rs12787061	61013931	*C11orf66*	G/**C**	0.008	0.018	0	0	S/T	382	58	benign	0.335	

[a] For the two SNPs in bold, genotype data for the HGDP panel are also available

[b] SNP positions are based on NCBI build 36

[c] Observed alleles are indicated as ancestral/**derived**

[d] Mean chemical distance for the corresponding amino acid pair

[e] PR D, probably damaging; PO D, possibly damaging.

[f] Ratio of the likelihood of a given amino acid occurring at a particular position to the likelihood of this amino acid occurring at any position

[g] Phenotypic effect of nonsynonymous coding SNPs as predicted by selective pressures estimated at the codon level. Pathological effects imply residues with $\omega < 0.1$.

Legends to Figures

Figure 1: Summary of signatures of selection in the *VPS37C* region. Depicted on top is the gene track along 2 Mb centered on *VPS37C* (highlighted in yellow) and below is a summary of the results of the different types of analyses in this region (*VPS37C* is delimited by the vertical bars in each plot). a and b, Proportion of SNPs with MAF < 0.10 and DAF > 0.80, respectively, within 100 kb sliding windows separated by 20 kb steps in East Asian populations. Solid dots represent values above the 95th percentile for each population, whereas open dots are values below the 95th percentile. c, F_{ST} values between all 39 populations for each SNP over the region. Solid dots represent the top 5% values of the overall F_{ST} distribution obtained across all analyzed regions. d and e, -log p-values of the XP-Rsb statistic for the seven continental regions and the six populations within EASIA, respectively. Horizontal dashed lines indicate statistical significance at the 0.05 level. Solid dots represent the lowest 5% p-values of the genome-wide Rsb distribution within each population (see methods for details).

Figure 2: Distribution of REHH against frequency for populations within six of the main geographical regions studied. Core haplotypes within ±100 kb of the candidate genes (black dots) are plotted over the background distribution of cores from the eleven full 2 Mb regions (gray dots) analyzed. REHH is shown at a distance of about 0.3 cM for all populations. Dashed lines indicate 0.95, 0.99 and 0.999 percentiles of REHH considering all cores. Cores that remained significant after multiple test correction (q < 0.05) are indicated with a black open diamond.

Figure 3: Bifurcation plots and EHH decay over physical distance for the two main haplotypes observed at the significant 3-SNP core of the *VPS37C* genomic region in MENA. On top, boxes represent genes, vertical gray lines are SNPs, vertical blue lines denote those constituting the core and vertical red lines indicate non-synonymous SNPs. Underlined SNPs represent other cores within the region. Gene

symbols are shown (from right to left) for the gene containing the core, the gene of interest in the region and the gene with the putatively selected variant in the region.

Figure 4: Worldwide frequency distribution of non-unique haplotypes based on 54 SNPs in the *VPS37C* region. Light blue segments represent the putatively selected haplotype whereas other haplotypes are indicated in other colours.

Figure 1.

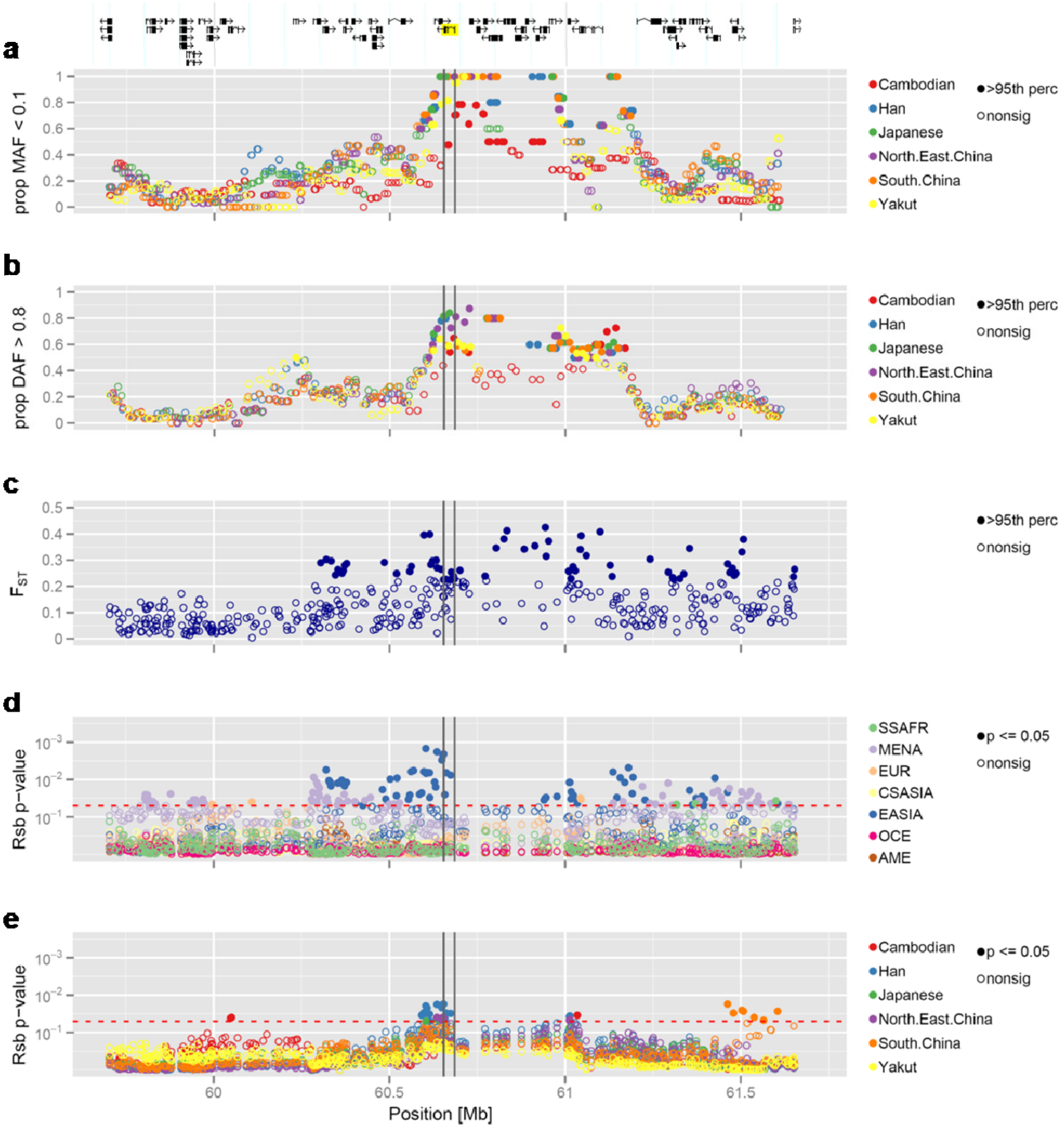

a
prop MAF < 0.1
1
0.8
0.6
0.4
0.2
0
●Cambodian
●Han
●Japanese
●North.East.China
●South.China
●Yakut
●>95th perc
○nonsig
b
prop DAF > 0.8
c
F_{ST}
0.5
0.4
0.3
0.2
0.1
0
d
Rsb p−value
10^{-3}
10^{-2}
10^{-1}
●SSAFR
●MENA
●EUR
●CSASIA
●EASIA
●OCE
●AME
●p <= 0.05
○nonsig
e
60
60.5
61
61.5
Position [Mb]

Figure 2.

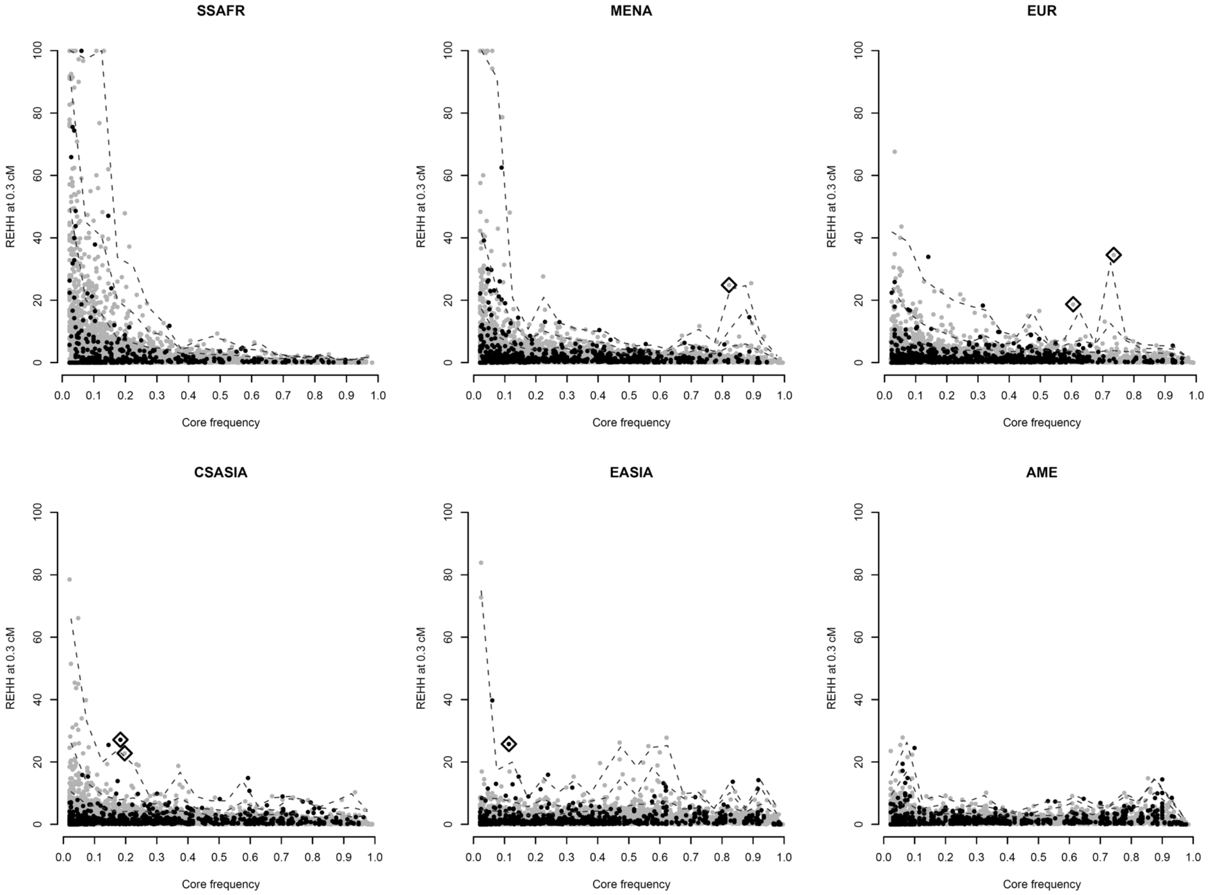

SSAFR
MENA
EUR
CSASIA
EASIA
AME
REHH at 0.3 cM
Core frequency
0
20
40
60
80
100
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

Figure 3.

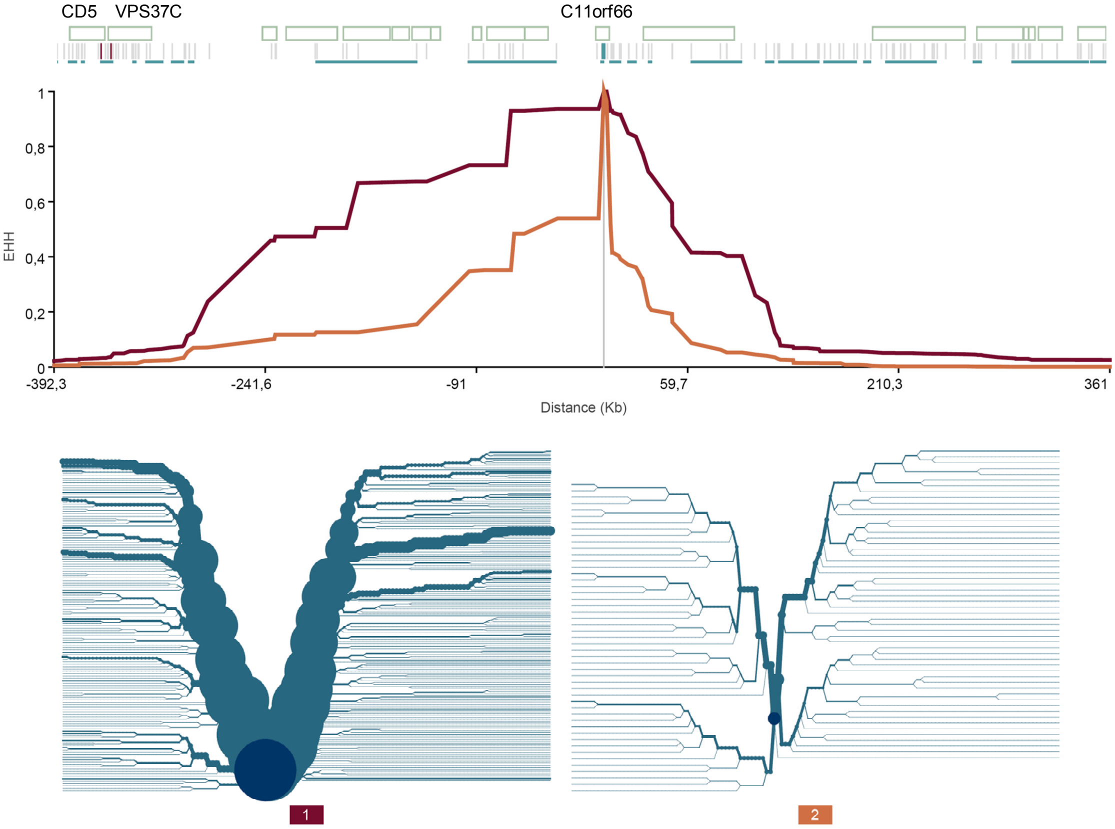
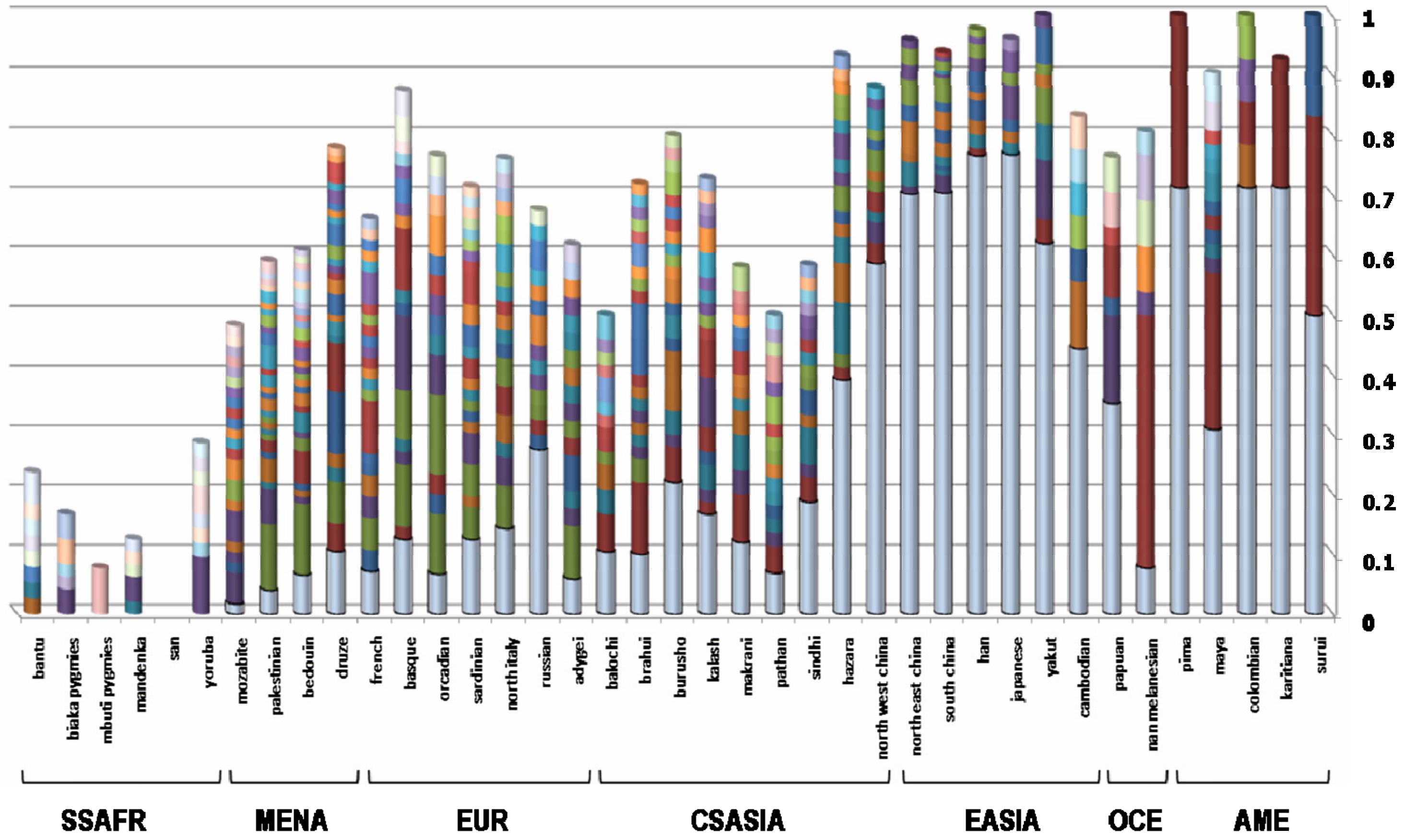

CD5
VPS37C
C11orf66
EHH
1
0,8
0,6
0,4
0,2
0
-392,3
-241,6
-91
59,7
210,3
361
Distance (Kb)
1
2

Figure 4.


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
bantu
biaka pygmies
mbuti pygmies
mandenka
san
yoruba
mozabite
palestinian
bedouin
druze
french
basque
orcadian
sardinian
north italy
russian
adygei
balochi
brahui
burusho
kalash
makrani
pathan
sindhi
hazara
north west china
northeast china
south china
han
japanese
yakut
cambodian
papuan
nan melanesian
pima
maya
colombian
karitiana
surui
SSAFR
MENA
EUR
CSASIA
EASIA
OCE
AME

Chapter 2: Signatures of Selection in the Human Olfactory Receptor OR5I1 Gene

Andrés Moreno-Estrada, Ferran Casals, Anna Ramírez-Soriano, Baldo Oliva, Francesc Calafell, Jaume Bertranpetit, and Elena Bosch.

Molecular Biology and Evolution. 2008. January; 25(1): 144-54.

Moreno-Estrada A, Casals F, Ramírez-Soriano A, Oliva B, Calafell F, Bertranpetit J, Bosch E.
Signatures of selection in the human olfactory receptor OR5I1 gene.
Mol Biol Evol. 2008 Jan;25(1):144-54. Epub 2007 Nov 2.

Chapter 3: African signatures of recent climate-related natural selection in human FOXI1

Andrés Moreno-Estrada, Estel Aparicio, Martin Sikora, Anna Ramírez-Soriano, Francesc Calafell, and Elena Bosch.

(*Manuscript in preparation*)

African signatures of recent climate-related natural selection in human *FOXI1*

Provisional list of authors: A Moreno-Estrada[1], E Aparicio[1], M Sikora[1], A Ramírez-Soriano[1], F Calafell[1,2], E Bosch[1,2]

[1] Institut de Biologia Evolutiva (UPF-CSIC), Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Parc de Recerca Biomèdica de Barcelona, Barcelona, Catalonia, Spain

[2] Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública (CIBERESP), Barcelona, Catalonia, Spain

Corresponding author:

Elena Bosch
Institut de Biologia Evolutiva (UPF-CSIC)
CEXS-UPF-PRBB
C/ Dr. Aiguader 88
08003 Barcelona
Spain

Tel: +34 93 3160841
Fax: +34 93 3160901
Email: elena.bosch@upf.edu

Key words: *FOXI1*, recent positive selection, human diversity

Running title: Human worldwide variation in *FOXI1*

Introduction

FOXI1 is a family member of the forkhead-box (FOX) transcription factors which are characterized by the FOX domain, a ~100 amino acid monomeric DNA-binding domain (Katoh and Katoh 2004). This forkhead motif is also known as the winged helix due to the butterfly-like appearance of the loops in the protein structure of the domain (Lehmann et al. 2003). Different mutations in the human *FOXI1* gene and in its regulatory binding site on *SLC26A4* (also known as pendrin) have been shown to compromise the transcription of this anion transporter gene on patients with Pendred Syndrome and nonsyndromic enlargement of the vestibular aqueduct (Yang et al. 2007). Similarly, Foxi1 null mutant mice had previously been found to lack the chloride/bicarbonate transporter pendrin leading to deafness and expansion of the endolimphatic compartment in inner ear (Hulander et al. 2003). Foxi1 has also been recognised as a key factor necessary for correct patterning of distal nephron epithelium and for adequate acid-base homeostasis in the kidney causing distal renal tubular acidosis in Foxi1$^{-/-}$ mice (Blomqvist et al. 2004). The *Slc4a9* promoter encoding for the anion exchanger 4 (AE4), which is expressed in type B intercalated cells within the renal collecting duct epithelium, was also identified as a direct target of Foxi1 (Kurth et al. 2006). Moreover, Foxi1 has been reported to be a crucial activator of the B1-subunit of the vacuolar H+ -ATPase proton pump (*ATP6V1B1*) as well as for pendrin (encoded by *SLC26A4*) and carbonic anhydrase II expression in the epididymal cells, which are required for a correct mice sperm maturation (Blomqvist et al. 2006). Overall, these findings led to the hypothesis that mutations in the human *FOXI1* gene might prove to cause a sensorineural deafness syndrome with distal renal tubular acidosis and male infertility (Blomqvist et al. 2006). Recently, the role of Foxi1 has been extended as a master regulator of the expression of vacuolar H+-ATPase proton pump subunits A1, B1, E2 and a4 at three different specific locations in mouse: FORE cells in endolympahtic epithelium, intercalated cells in the kidney and narrow and clear cells of epididymis (Vidarsson et al. 2009).

Thus, *FOXI1* seems to be involved in the correct function of at least three different organs: inner ear, testis, and kidney. All three escenarios offer a possibility for natural selection. Using orthologous gene trios in humans, chimpanzee and mouse, *FOXI1* was suggested as example of gene involved in hearing that appeared to be under human-specific selection (Clark et al. 2003). However, the availability of additional *FOXI1*orthologous sequences from a variety of species including other primates permits the revision of such an initial finding. Genes involved in sensory perception have long been recognised to show accelerated evolution and/or being under positive selection on the human lineage (Clark et al. 2003; Arbiza, Dopazo, and Dopazo 2006) whereas reproduction was only revealed as biological function showing an excess of positively selected genes when considering together the genomes of humans and chimpanzees (Nielsen et al. 2005). Within humans, adaptation to different climates may have entailed adapting kidney function to different dehydration levels. For instance, a thrifty genotype in water retention has been suggested to have been selected in the grueling trans-Atlantic voyage of the slave ships, resulting in a higher prevalence of hypertension in African-Americans (Wilson and Grim 1991). Moreover, variants influencing salt homeostasis in the *CYP3A5* and *AGT* loci have been shown to be targets of a selective pressure varying in intensity in correlation with latitude (Thompson et al. 2004). It is tempting to especulate, therefore, that *FOXI1* may have been the target of local adaptation in humans. In order to test that hypothesis, we evaluated the patterns of nucleotide variation and looked for signals of positive natural selection in the human *FOXI1* gene by resequencing 20 Europeans, 20 Asians and 20 Yorubans and by assaying SNP variation in the Human Genome Diversity Panel (HGDP-CEPH; (Cann et al. 2002)).

Materials and Methods

Population samples

Sixty DNA population samples for resequencing were obtained from Coriell Cell Repositories (Camden, NJ, USA). These consisted of 20 Europeans (Utah residents with ancestry from Northern and Western Europe), 20 Asians (10 Japanese from Tokyo and 10 Chinese from Beijing) and 20 Yorubans (from Ibadan in Nigeria). Coriell Repositories numbers for these samples are as follows: European (NA06994, NA07000, NA07345, NA07357, NA11829, NA11830, NA11839, NA11840, NA11992, NA11993, NA12003, NA12004, NA12043, NA12044, NA12056, NA12057, NA12750, NA12751, NA12812, NA12813), Han Chinese (NA18576, NA18577, NA18579, NA18582, NA18593, NA18623, NA18624, NA18632, NA18635, NA18636), Japanese (NA18940, NA18942, NA18943, NA18944, NA18948, NA18949, NA18951, NA18956, NA18970, NA18973), Yoruban (NA18501, NA18502, NA18507, NA18508, NA18855, NA18856, NA18861, NA18862, NA19127, NA19128, NA19137, NA19138, NA19171, NA19172, NA19203, NA19204, NA19206, NA19207, NA19209, NA19210).

SNP genotypes were obtained for the Human Genome Diversity Panel (HGDP-CEPH), which contains 1,064 DNA samples from individuals representing 51 populations globally distributed (Cann et al. 2002). In all statistical analyses atypical, duplicated individuals and deduced first-degree relatives have been removed by using the H971 subset recommended by Rosenberg (2006). In order to maximize sample sizes, genotyped samples were re-grouped into 39 populations based on geographic and ethnic criteria as in Gardner et al. (2006). For part of the analysis, populations were further grouped into seven geographical regions: Sub-Saharan Africa (SSAFR), Middle East-North Africa (MENA), Europe (EUR), Central-South Asia (CSASIA), East Asia (EASIA), Oceania (OCE) and America (AME).

Resequencing of the FOXI1 gene

Two overlapping fragments of 2100 bp and 2099 bp, respectively, covering the entire *FOXI1* genomic sequence were amplified by use of the following primers: FO-F1 5'-GTCATTAGTGGGGACCTGAG-3', FO-R1 5'-GGCATGAGCATTAAGGAGTT-3'; FO-F2 5'-GCCAGGACTCAAGTCTGTCT-3', FO-R2 5'-GCACCACATGTTTGTTTGTT-3'. PCRs were performed in a total volume of 25 μl, containing 0.2 mM dNTPs, 1.5 mM $MgCl_2$, 0.5 μM of each primer, 1X buffer, 0.05 U Taq polymerase (Ecogen), and 10 ng genomic DNA. PCR conditions were as follows: 3 min at 94ºC, 30 cycles of 94ºC for 30 secs , 55ºC for 30 secs and 72ºC for 3 min; and a final step of extension of 5 min at 72ºC. PCR products were purified by use of the MultiScreen PCRμ96 chemistry (Millipore) according to the manufacturer's protocol. Amplified PCR products were sequenced using the Big Dye Terminator chemistry ver.3.1 (Applied Biosystems) and the following sequencing primers: FO-S1F 5'-CAACCCCTACCTCTGGTTC -3', FO-S2R 5'-

AGAGCCGAGTAGGAATAGGG-3', FO-S3F 5'-GTCCCCAAGGAAGACTCAC-3', FO-S4F 5'-TTGGTGAATGAATGACTGGA-3', FO-S5F 5'-CTCAAAGGAACCCCAACTC-3, FO-S6F 5'-GAGTCCCAGAGTTTTCCTGA-3', FO-S7F 5'-TTTACCCCCTTTCACTTTTG-3', FO-S8F 5'-TGACCTTCAACTCCTTCTCC-3', FO-S9F 5'-CAGAGCAGCACTAACAGTGG-3' and FO-S10F 5'-GCTACCACTCAAGGAAGGAA-3'. Extension products were purified by use of the Montage SEQ96 Cleanup kit (Millipore) and run on an ABI 3730 XL sequencer. Sequence analysis and contig assembly for each sample were performed with the Seqman module within the DNASTAR Lasergene software ver. 7.1.0 and visually inspected at least twice. All polymorphic sites were checked manually and heterozygote positions were confirmed by reamplifying and resequencing the SNP site from the same or opposite strand.

SNP genotyping data

We genotyped a total of 21 SNPs covering 400 kb centered on the *FOXI1* gene region. SNPs were selected every 5-10 kb within the gene and up to around 30 kb in both 5' and 3' end flanking regions; from this point, an extra SNP was then added every 40 kb up to around 200 kb in both flanking regions. Preference was given to SNPs with a minor allele frequency (MAF) over 10%, which were compiled from HapMap (Release 7 May 2004) and dbSNP (Build 121 June 2004) databases. SNPs were typed using the SNPlex Genotyping System from Applied Biosystems within a larger set of SNPs covering additional genes as described elsewhere (Moreno-Estrada et al. 2008). Illumina HumanHap650K Beadchip genotypes on the HGDP-CEPH panel were downloaded from the Stanford Human Genome Center website. From this publicly available data we extracted a SNP genotype set centered on *FOXI1* and extending up to 2 Mb (with a total of 693 SNPs), which was complemented by 15 of the previously obtained SNPs.

Statistical analysis

In order to test for positive selection on the human lineage we initially applied the improved branch-site test 2 of positive selection (Zhang, Nielsen, and Yang 2005) in a phylogeny containing five mammal species. The human reference sequence for *FOXI1* and its ortholog sequences in chimpanzee, mouse, rat and dog had been previously extracted from Ensembl (Gene IDs ENSG00000168269, ENSPTRG00000017514, ENSMUSG00000047861, ENSRNOG00000006293, ENSCAFG00000016968) and aligned with ClustalW (Thompson, Higgins, and Gibson 1994). Calculations for the corresponding null and alternative hypotheses were performed using the codeml program implemented in the PAML package (Yang 1997). Results for additional likelihood ratio tests of positive selection on the human lineage considering different multispecies alignments were obtained from the Human PAML browser (Nickel, Tefft, and Adams 2008). In that case, supplementary Table 1 summarises the likelihood ratio test results for the branch test (Model H versus Model H null) and for the strict branch positive site test of positive selection (Model A versus Model A null).

Arlequin (Schneider, 2000) was used to calculate F_{ST} values between the 39 populations studied with a locus by locus Analysis of Molecular Variance (AMOVA) (Excoffier, Smouse, and Quattro 1992). Haplotypes were inferred from genotype sequencing data using the Bayesian statistical method in program PHASE 2.1 (Stephens, Smith, and Donnelly 2001) using the default parameter set with 1,000 iterations. Haplotype estimation from the 400 Kb and 2 Mb SNP genotype data sets was performed with FastPHASE (Scheet et al., 2008). The Network 4.5.0.1 software package (http://www.fluxus-engineering.com) was used to construct the minimum mutation network by means of the median-joining algorithm (Bandelt et al. 1995; Bandelt, Forster, and Rohl 1999). The ancestral states for the *FOXI1* polymorphic positions were inferred from the previously aligned ortholog sequences in chimpanzee, mouse and rat but adding macaque (ENSMMUG00000014124) in the alignment.

Nucleotide diversity statistics and analysis of population polymorphic sites were carried out with the *FOXI1* resequencing data and the DnaSP software ver. 4.00 (Rozas et al. 2003). Departures from neutrality were tested by means of the Tajima's D, Fu and Li's F, F*, D and D* and Fay and Wu's H tests. In order to obtain realistic distributions for the statistics and thus evaluate evidence for natural selection, we performed 10,000 coalescent simulations using Cosi version 1.1 (Schaffner et al. 2005). As some demographic effects (such as population expansions) and positive selection have similar effects over genealogies (Charlesworth, Morgan, and Charlesworth 1993) those simulations include the ad-hoc human demographic calibration described in Schaffner et al. (2005) and provided with the Cosi source code. We have assumed an infinite-sites model, we have fixed S to the number of segregating sites found, and the length of simulated sequence has been set to 4007 bp (length of sequence analyzed). As for recombination, we have used the recombination rates estimated for the region (Myers et al. 2005). The critical value for each statistic has been obtained from the empirical distribution of the corresponding neutral model with a significance level of 0.05. For the whole human sample, DnaSP software ver. 4.00 (Rozas et al. 2003) was used to produce coalescent neutral simulations with a constant population size.

For every population we analyzed the distribution of the minor allele frequencies (MAF) and the derived allele frequencies (DAF) of the corresponding SNPs along 2 Mb centered on *FOXI1*. Ancestral alleles for the genotyped SNPs were those recovered from the chimpanzee and/or the macaque genome sequences (panTro2, Mar. 2006 assembly and rheMac2, Jan. 2006 assembly, respectively). The proportion of SNPs with allele frequencies higher or lower than a defined threshold (MAF < 0.10 for the MAF analysis and DAF > 0.80 for the DAF analysis) was calculated within sliding windows of 100 Kb in size every 20 Kb and plotted against distance. Thresholds were chosen to maximize sensitivity to selection as suggested by Walsh et al. (2006), and we required a minimum of 5 SNPs per window. Non-polymorphic SNPs in the overall 39 populations were not considered in the threshold analyses of minor and derived allele frequencies. However, SNPs fixed in a population but polymorphic elsewhere were counted as having MAF < 0.10 and DAF > 0.80 when applicable. The definition of minor allele was specific for each population rather than global and uniform across populations.

Unusual long range haplotypes along the 2 Mb region centered on *FOXI1* were explored by applying two complementary EHH-based approaches especially designed to detect both intermediate and fixed (or nearly so) frequency selected variants. The Long Range Haplotype (LRH) test (Sabeti et al. 2002) was carried out using the SWEEP software package (version 1.1) defining cores as the longest non-overlapping core haplotypes with at least one SNP and not more than 20 SNPs. In order to obtain sufficient data for a background distribution of EHH values we performed the analysis of the *FOXI1* region together with phased haplotype data for eleven 2 Mb regions studied elsewhere (Moreno-Estrada et al. submitted). For each identified core haplotype, we calculated the EHH and the relative EHH (REHH) at a 0.04 and 0.02 marker breakdown from the core for each main continental region separately. Core haplotypes were placed in 5% frequency bins and the respective EHH and REHH values were log-transformed for each bin in order to obtain approximately normally distributed values. Empirical p values for the LRH test were obtained by using the mean and standard deviation of the empirical distribution of the respective scores in each continental region. To account for multiple testing, we estimated the false positive discovery rate (pFDR) (Storey and Tibshirani 2003) and calculated the q value for the scores within each frequency bin using the package q value (version 1.1) for R. The q value for a particular p value is defined as the expected proportion of false positives among all significant p values when calling that p value significant. We used a q value cutoff of 0.05 for assigning significance. In order to allow for multiple populational comparisons of EHH, we computed XP-Rsb for every SNP site and population *versus* all other HGDP-CEPH populations directly from genotype data as described in Moreno-Estrada et al. (submitted). In order to assess significance, we obtained a p-value for each population and each SNP site along the complete 2Mb *FOXI1* region, which is basically the rank of its XP-Rsb value with respect that calculated for 642,690 genome-wide distributed SNPs generated elsewhere in the same panel (Li et al. 2008). We then log transformed the p-values and plotted them against position, searching for clusters of significant values inside or around the *FOXI1* gene.

Results

FOXI1 divergence patterns

The recent increase of new sequenced genomes available lead us to reavaluate the initial evidence of nonneutral evolution on *FOXI1* in the human lineage when using three-species sequence alignments (Clark et al. 2003). When considering a phylogeny of five mamals, the improved branch-site test 2 of positive selection (Zhang, Nielsen, and Yang 2005) rejected the hypothesis of positive selection at a subset of sites in the human branch (p=0.7039). We also investigated the patterns of FOXI1 protein evolution from several multiple sequence alignments, including additional non-human primates besides chimpanzee, using two maximum likelihood methods (see Materials and Methods) to especifically test for: (i) a d_N/d_S ratio on the human branch significantly different from 1, and (ii) codon sites with a d_N/d_S ratio significantly different from 1 in the human lineage. In none of the comparisons *FOXI1* presented overall significantly accelerated amino acid substitution rates or particular codon sites undergoing positive selection in humans (Supplementary Table 1).

Patterns of FOXI1 sequence variation

Polymorphic variation at the *FOXI1* gene was investigated by sequencing 4,007 bp, encompassing its two exons, most of its untranslated regions, and its corresponding intronic region, in 20 Yorubans, 20 Asians and 20 Europeans. We found 22 sequence variations: 21 substitutions and one deletion/insertion polymorphism (Table 1 and Figure 1). Among them, six were singletons: one specific of Yorubas, three specific of Europeans and two specific of Asians. One of the Asian-specific singletons implied the only non-synonymous substitution observed (rs3828625), an Asn to Ser replacement at amino acid position 362 for isoform a (or amino acid position 267 for isoform b). The remaining five polymorphims lying in the exons implied all synonymous changes. Seventeen of the 22 sequence variations detected here have been already reported in dbSNP build 129. Excluding singleton variants, we identified a total of 15 haplotypes (Table 2). Half of the analysed chromosomes belonged to either Ht-01 or Ht-04, found in all the populations studied. We find three specific haplotypes for Africans, three for Europeans but no haplotype was specific of Asians. In order to visualize the phylogenetic relationships among the identified haplotypes we constructed a median-joining network considering a putative human ancestral haplotype estimated from chimpanzee data (Supplementary Figure 1). The reticulated pattern observed in the network points to the action of recurrent mutation or recombination. However, given that often more than one nucleotide position is involved, the latter seems more likely as a mechanism for producing new sequence variation.

Summary statistics of genetic diversity and neutrality tests for *FOXI1* are reported in Table 3. The three studied populations showed similar levels of nucleotide diversity, which are slightly higher than the average obtained for the 322 resequenced genes included in SeatleSNPs database (with an average π value of 9.02×10^{-4} in the African-American sample and of 6.85×10^{-4} in the European CEPH population). Notably, the Yoruban population showed similar numbers of segregating sites, of total number of different haplotypes as well as similar values of nucleotide and haplotype diversity values than the European sample. In order to investigate the possible genetic footprint of selection we performed several neutrality tests on each of these three populations and on the whole human sample (Table 3). Significance was estimated by means of coalescent simulations under each inferred past demography (Schaffner et al. 2005) or in the case of the whole human sample by considering constant population size. None of the neutrality statistics displayed significant departures from neutrality ($P<0.05$) in any case.

Patterns of SNP variation in FOXI1

We also explored the pattern of SNP variation along a 2 Mb genomic region centered on *FOXI1* in 39 worldwide human populations covering most of the human genetic diversity. In particular, we examined for an excess of low frequency variants, for the presence of high frequency derived alleles, as well as for unusual long range haplotypes. These possible signatures of natural selection persist in the human genome at varying time scales, and, therefore, provide different overlapping windows in our evolutionary history where to explore adaptation (Sabeti et al. 2006). For each population within each of the seven main geographical regions, we plotted separately within multiple 100 Kb slidding windows along the 2 Mb region centered on *FOXI1*: (i) the proportion of SNPs with MAF < 0.10 (Supplementary Figure S2) and (ii) the proportion of SNPs with DAF > 0.80 (Supplementary Figure S3). In both analyses, no particular pattern emerged around or surrounding the *FOXI1* gene in any of the populations studied. In order to search for unusual long haplotypes we applied the XP-Rsb statistic (Moreno-Estrada et al. (submitted) and the long range haplotype test (LRH test, (Sabeti et al. 2002)). Whereas XP-Rsb compares for each particular SNP site the integrated Extended Haplotype Homozigosity (EHH) in a population to that averaged across multiple populations, the LRH test looks for core haplotypes that have a combination of high population frequency and high EHH in relation to other core haplotypes in the same population. The genome-wide significance of XP-Rsb at every SNP site and population across each main geographical region analysed was plotted against distance along the 2 Mb region centered in *FOXI1* (Figure 2). Interestingly, several significant windows are detected along the *FOXI1* gene position in the case of the Yoruba ($p< 0.01$) and Mandeka ($p< 0.05$) African populations. Similarly, up to five different core haplotypes close to *FOXI1* clearly stand out with relatively high frequencies (> 0.50) and significant REHH values in the Yoruba population (see Table 4) when compared to additional eleven 2 Mb regions analyzed elsewhere (see Materials and Methods). The EHH breakdown for the most extreme significant core haplotype near *FOXI1* can be seen in Figure 3. The signal is detected at both 0.02 and 0.04 marker breakdown from the core and is only observed in one direction, coinciding towards *FOXI1* position. An additional picture of the unusual EHH breakdown around *FOXI1* in Yorubans is presented in Supplementary Figure 4.

We next explored which may have been the functional variation under the recent selection signal detected in Africa by both EHH based methods. To that effect, we first explored in HapMap (Frazer et al. 2007) the Linkage Disequilibrium (LD) pattern around *FOXI1* in the Yoruban population and found an LD block containing *FOXI1* in a region of ~70 Kb between two hotspots of recombination. As shown in Figure 5, a similar LD block structure is detected in the Yoruban population of the HGDP-CEPH panel. Next, we screened all known SNP variation for a putative functional effect using PupaSuite (Conde et al. 2006; Reumers et al. 2008) around ~140 Kb containing such block. Out of 646 we detected 19 SNPs as functionally relevant polymorphisms (Supplementary Table 2). Since we have found that the footprint of natural selection seems to be evident in Africans, we can narrow down that candidate list by considering first the functional SNPs with the largest allele frequency difference between Africans and non-Africans. In four cases allele frequencies seem to be significantly different between Africans and non-Africans. All of them are putativaly functional SNPs affecting exonic splicing enhancers either for the *DOCK2* gene (rs6555882 and rs1045176) or for the *FOXI1* gene (rs2277944 and rs6873124). While rs2277944 and rs6873124 are less than 3.6 Kb from the haplotype block previously described in Africa and rs1045176 displays some linkage disequilibrium with it, rs6555882 maps further away. Therefore, rs2277944, rs6873124 and rs1045176 (or unknown functional variants in linkage diusequilibrium with them) are the best candidates to represent the functional variation at the base of this selection event.

Finally, in order to explore a possible connection between climate and *FOXI1* we checked for correlation between allele frequencies and absolute latitude for the SNPs in the significant core haplotypes and in rs2277944 (worldwide allele frequency distributions were not available for the rest of the suggested candidate functional SNPs). All core haplotypes contained at least one SNP with significant correlation with absolute latitude ($p<0.05$); rs2277944 was also significantly correlated with latitude ($p<0.01$). The most significant correlation (Pearson's r = -0.536, $p<0.01$ after Bonferroni correction) was for allele A at rs7736379 in core ACCC (see table 4), implying that its frequency decreases towards higher latitudes (Supplementary Figure S5).

Discussion

FOXI1 is involved in ear, testis, and kidney. Genes involved in fertility and/or sensory perception may have been driven by accelerated evolution in humans, but, contrary to Clark et al. (2003), we do not find such acceleration on the *FOXI1* gene. It is recognised that acuracy and power of likelihood ratio tests for detecting positive selection improve with the number of species used on such approaches (Anisimova, Bielawski, and Yang 2001). In accordance with that, we used the Human PAML browser (Nickel, Tefft, and Adams 2008) in order to perform both branch and branch sites tests of positive selection on multiple species comparisons containing several mammals and additional primates to human and chimpanzee. None of the tests performed indicated positive selection for the *FOXI1* gene in the human lineage.

Variation in the *FOXI1* gene sequence was compatible with neutral evolution in samples of Africans, East Asians and Europeans. However, patterns of extended linkage disequilibrium in the *FOXI1* gene region suggested positive selection in the African samples of the HGDP-CEPH diversity panel (Cann et al. 2002). The core haplotypes contained SNPs with allele frequencies that correlated negatively with latitude, suggesting a role for climate in the adaptation mediated by *FOXI1*. Patterns of linkage disequilibrium are suggested to pinpoint recent selective events, while the accumulation of variation by mutation requires some time before sequence-based neutrality statistics become significant (Ramirez-Soriano et al. 2008). The finding of no significant deviation from neutral evolution with statistics such as Tajima's D, or Fay and Wu's H has been previously recognized to be consistent with the low power of these traditional tests to detect recent selective sweeps as for example in the case of the *G6PD* locus (Sabeti et al. 2002). Therefore, our results suggest that climate may have driven selection on *FOXI1*. Of the diverse functions of these gene, those related to the kidney seem to be the most logical candidates to climate adaptation through water homeostasis and prevention of dehydratation. If that were the case, *FOXI1* would join the ranks of the growing number of known genes that allowed adaptation of humans to the diversity of climates we encountered during our past expansion such as *FABP2*, *RAPTOR* and *SLC24A5* (Hancock et al. 2008). Our evolutionary approach has pointed directions for functional analyses to test how genetic variation in the *FOXI1* gene region can result on phenotypic differences among human populations probably in relation to water homeostasis.

Acknowledgements

We specially thank M Vallès, R Anglada, and S Plaza for their technical advise and support. This research was funded by grant BFU2005-00243 awarded by Dirección General de Investigación, Ministerio de Educación y Ciencia (Spain), by grant BFU2008-01046 awarded by Subdirección General de Proyectos de Investigación, Ministerio de Ciencia e Innovación (Spain), and by the Direcció General de Recerca , Generalitat de Catalunya (2005SGR00608). A Moreno-Estrada is supported by a CONACYT fellowship from the Mexican government (grant

179339). SNP genotyping services were provided by the Spanish "Centro Nacional de Genotipado" (www.cegen.org).

References

Anisimova, M., J. P. Bielawski, and Z. Yang. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol Biol Evol **18**:1585-1592.

Arbiza, L., J. Dopazo, and H. Dopazo. 2006. Positive selection, relaxation, and acceleration in the evolution of the human and chimp genome. PLoS Comput Biol **2**:e38.

Bandelt, H. J., P. Forster, and A. Rohl. 1999. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol **16**:37-48.

Bandelt, H. J., P. Forster, B. C. Sykes, and M. B. Richards. 1995. Mitochondrial portraits of human populations using median networks. Genetics **141**:743-753.

Blomqvist, S. R., H. Vidarsson, S. Fitzgerald, B. R. Johansson, A. Ollerstam, R. Brown, A. E. Persson, G. G. Bergstrom, and S. Enerback. 2004. Distal renal tubular acidosis in mice that lack the forkhead transcription factor Foxi1. J Clin Invest **113**:1560-1570.

Blomqvist, S. R., H. Vidarsson, O. Soder, and S. Enerback. 2006. Epididymal expression of the forkhead transcription factor Foxi1 is required for male fertility. Embo J **25**:4131-4141.

Cann, H. M., C. de Toma, L. Cazes, M. F. Legrand, V. Morel, L. Piouffre, J. Bodmer, W. F. Bodmer, B. Bonne-Tamir, A. Cambon-Thomsen, Z. Chen, J. Chu, C. Carcassi, L. Contu, R. Du, L. Excoffier, G. B. Ferrara, J. S. Friedlaender, H. Groot, D. Gurwitz, T. Jenkins, R. J. Herrera, X. Huang, J. Kidd, K. K. Kidd, A. Langaney, A. A. Lin, S. Q. Mehdi, P. Parham, A. Piazza, M. P. Pistillo, Y. Qian, Q. Shu, J. Xu, S. Zhu, J. L. Weber, H. T. Greely, M. W. Feldman, G. Thomas, J. Dausset, and L. L. Cavalli-Sforza. 2002. A human genome diversity cell line panel. Science **296**:261-262.

Charlesworth, B., M. T. Morgan, and D. Charlesworth. 1993. The effect of deleterious mutations on neutral molecular variation. Genetics **134**:1289-1303.

Clark, A. G., S. Glanowski, R. Nielsen, P. D. Thomas, A. Kejariwal, M. A. Todd, D. M. Tanenbaum, D. Civello, F. Lu, B. Murphy, S. Ferriera, G. Wang, X. Zheng, T. J. White, J. J. Sninsky, M. D. Adams, and M. Cargill. 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science **302**:1960-1963.

Conde, L., J. M. Vaquerizas, H. Dopazo, L. Arbiza, J. Reumers, F. Rousseau, J. Schymkowitz, and J. Dopazo. 2006. PupaSuite: finding functional single nucleotide polymorphisms for large-scale genotyping purposes. Nucleic Acids Res **34**:W621-625.

Excoffier, L., P. E. Smouse, and J. M. Quattro. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics **131**:479-491.

Frazer, K. A.D. G. BallingerD. R. CoxD. A. HindsL. L. StuveR. A. GibbsJ. W. BelmontA. BoudreauP. HardenbolS. M. LealS. PasternakD. A. WheelerT. D. WillisF. YuH. YangC. ZengY. GaoH. HuW. HuC. LiW. LinS. LiuH. PanX. TangJ. WangW. WangJ. YuB. ZhangQ. ZhangH. ZhaoH. ZhaoJ. ZhouS. B. GabrielR. BarryB. BlumenstielA. CamargoM. DefeliceM. FaggartM. GoyetteS. GuptaJ. MooreH. NguyenR. C. OnofrioM. ParkinJ. RoyE. StahlE. WinchesterL. ZiaugraD. AltshulerY. ShenZ. YaoW. HuangX. ChuY. HeL. JinY. LiuY. ShenW. SunH. WangY. WangY. WangX. XiongL. XuM. M. WayeS. K. TsuiH. XueJ. T. WongL. M. GalverJ. B. FanK. GundersonS. S. MurrayA. R. OliphantM. S. CheeA. MontpetitF. ChagnonV. FerrettiM. LeboeufJ. F. OlivierM. S. PhillipsS. RoumyC. SalleeA. VernerT. J. HudsonP. Y. KwokD. CaiD. C. KoboldtR. D. MillerL. PawlikowskaP. Taillon-MillerM. XiaoL. C. TsuiW. MakY. Q. SongP. K. TamY. NakamuraT. KawaguchiT. KitamotoT. MorizonoA. NagashimaY. OhnishiA. SekineT. TanakaT. TsunodaP. DeloukasC. P. BirdM. DelgadoE. T. DermitzakisR. GwilliamS. HuntJ. MorrisonD. PowellB. E. StrangerP. WhittakerD. R. BentleyM. J. DalyP. I. de BakkerJ. BarrettY. R. ChretienJ. MallerS. McCarrollN. PattersonI. Pe'erA. PriceS. PurcellD. J. RichterP. SabetiR. SaxenaS. F. SchaffnerP. C. ShamP. VarillyD. AltshulerL. D. SteinL. KrishnanA. V. SmithM. K. Tello-RuizG. A. ThorissonA. ChakravartiP. E. ChenD. J. CutlerC. S. KashukS. LinG. R. AbecasisW. GuanY. LiH. M. MunroZ. S. QinD. J. ThomasG. McVeanA. AutonL. BottoloN. CardinS. EyheramendyC. FreemanJ. MarchiniS. MyersC. SpencerM. StephensP. DonnellyL. R. CardonG. ClarkeD. M. EvansA. P. MorrisB. S. WeirT. TsunodaJ. C. MullikinS. T. SherryM. FeoloA. SkolH. ZhangC. ZengH. ZhaoI. MatsudaY. FukushimaD. R. MacerE. SudaC. N. RotimiC. A. AdebamowoI. AjayiT. AniagwuP. A. MarshallC. NkwodimmahC. D. RoyalM. F. LeppertM. DixonA. PeifferR. QiuA. KentK. KatoN. NiikawaI. F. AdewoleB. M. KnoppersM. W. FosterE. W. ClaytonJ. WatkinR. A. GibbsJ. W. BelmontD. MuznyL. NazarethE. SodergrenG. M. WeinstockD. A. WheelerI. YakubS. B. GabrielR. C. OnofrioD. J. RichterL. ZiaugraB. W. BirrenM. J. DalyD. AltshulerR. K. WilsonL. L. FultonJ. RogersJ. BurtonN. P. CarterC. M. CleeM. GriffithsM. C. JonesK. McLayR. W. PlumbM. T. RossS. K. SimsD. L. WilleyZ. ChenH. HanL. KangM. GodboutJ. C. WallenburgP. L'ArchevequeG. BellemareK. SaekiH. WangD. AnH. FuQ. LiZ. WangR. WangA. L. HoldenL. D. BrooksJ. E. McEwenM. S. GuyerV. O. WangJ. L. PetersonM. ShiJ. SpiegelL. M. SungL. F. ZachariaF. S. CollinsK. KennedyR. Jamieson, and J. Stewart. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature **449**:851-861.

Gardner, M., A. Gonzalez-Neira, O. Lao, F. Calafell, J. Bertranpetit, and D. Comas. 2006. Extreme population differences across Neuregulin 1 gene, with implications for association studies. Mol Psychiatry **11**:66-75.

Hancock, A. M., D. B. Witonsky, A. S. Gordon, G. Eshel, J. K. Pritchard, G. Coop, and A. Di Rienzo. 2008. Adaptations to climate in candidate genes for common metabolic disorders. PLoS Genet **4**:e32.

Hulander, M., A. E. Kiernan, S. R. Blomqvist, P. Carlsson, E. J. Samuelsson, B. R. Johansson, K. P. Steel, and S. Enerback. 2003. Lack of pendrin expression leads to deafness and expansion of the endolymphatic compartment in inner ears of Foxi1 null mutant mice. Development **130**:2013-2025.

Katoh, M., and M. Katoh. 2004. Human FOX gene family (Review). Int J Oncol **25**:1495-1500.

Kurth, I., M. Hentschke, S. Hentschke, U. Borgmeyer, A. Gal, and C. A. Hubner. 2006. The forkhead transcription factor Foxi1 directly activates the AE4 promoter. Biochem J **393**:277-283.

Lehmann, O. J., J. C. Sowden, P. Carlsson, T. Jordan, and S. S. Bhattacharya. 2003. Fox's in development and disease. Trends Genet **19**:339-344.

Li, J. Z., D. M. Absher, H. Tang, A. M. Southwick, A. M. Casto, S. Ramachandran, H. M. Cann, G. S. Barsh, M. Feldman, L. L. Cavalli-Sforza, and R. M. Myers. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. Science **319**:1100-1104.

Moreno-Estrada, A., F. Casals, A. Ramirez-Soriano, B. Oliva, F. Calafell, J. Bertranpetit, and E. Bosch. 2008. Signatures of selection in the human olfactory receptor OR5I1 gene. Mol Biol Evol **25**:144-154.

Myers, S., L. Bottolo, C. Freeman, G. McVean, and P. Donnelly. 2005. A fine-scale map of recombination rates and hotspots across the human genome. Science **310**:321-324.

Nickel, G. C., D. Tefft, and M. D. Adams. 2008. Human PAML browser: a database of positive selection on human genes using phylogenetic methods. Nucleic Acids Res **36**:D800-808.

Nielsen, R., C. Bustamante, A. G. Clark, S. Glanowski, T. B. Sackton, M. J. Hubisz, A. Fledel-Alon, D. M. Tanenbaum, D. Civello, T. J. White, J. S. J, M. D. Adams, and M. Cargill. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol **3**:e170.

Ramirez-Soriano, A., S. E. Ramos-Onsins, J. Rozas, F. Calafell, and A. Navarro. 2008. Statistical power analysis of neutrality tests under demographic expansions, contractions and bottlenecks with recombination. Genetics **179**:555-567.

Reumers, J., L. Conde, I. Medina, S. Maurer-Stroh, J. Van Durme, J. Dopazo, F. Rousseau, and J. Schymkowitz. 2008. Joint annotation of coding and non-coding single nucleotide polymorphisms and mutations in the SNPeffect and PupaSuite databases. Nucleic Acids Res **36**:D825-829.

Rosenberg, N. A. 2006. Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. Ann Hum Genet **70**:841-847.

Rozas, J., J. C. Sanchez-DelBarrio, X. Messeguer, and R. Rozas. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics **19**:2496-2497.

Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. Levine, D. J. Richter, S. F. Schaffner, S. B. Gabriel, J. V. Platko, N. J. Patterson, G. J. McDonald, H. C. Ackerman, S. J. Campbell, D. Altshuler, R. Cooper, D. Kwiatkowski, R. Ward, and E. S. Lander. 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature **419**:832-837.

Sabeti, P. C., S. F. Schaffner, B. Fry, J. Lohmueller, P. Varilly, O. Shamovsky, A. Palma, T. S. Mikkelsen, D. Altshuler, and E. S. Lander. 2006. Positive natural selection in the human lineage. Science **312**:1614-1620.

Schaffner, S. F., C. Foo, S. Gabriel, D. Reich, M. J. Daly, and D. Altshuler. 2005. Calibrating a coalescent simulation of human genome sequence variation. Genome Res **15**:1576-1583.

Stephens, M., N. J. Smith, and P. Donnelly. 2001. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet **68**:978-989.

Storey, J. D., and R. Tibshirani. 2003. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A **100**:9440-9445.

Thompson, E. E., H. Kuttab-Boulos, D. Witonsky, L. Yang, B. A. Roe, and A. Di Rienzo. 2004. CYP3A variation and the evolution of salt-sensitivity variants. Am J Hum Genet **75**:1059-1069.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res **22**:4673-4680.

Vidarsson, H., R. Westergren, M. Heglind, S. R. Blomqvist, S. Breton, and S. Enerback. 2009. The forkhead transcription factor Foxi1 is a master regulator of vacuolar H-ATPase proton pump subunits in the inner ear, kidney and epididymis. PLoS ONE **4**:e4471.

Wilson, T. W., and C. E. Grim. 1991. Biohistory of slavery and blood pressure differences in blacks today. A hypothesis. Hypertension **17**:I122-128.

Yang, T., H. Vidarsson, S. Rodrigo-Blomqvist, S. S. Rosengren, S. Enerback, and R. J. Smith. 2007. Transcriptional control of SLC26A4 is involved in Pendred syndrome and nonsyndromic enlargement of vestibular aqueduct (DFNB4). Am J Hum Genet **80**:1055-1063.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci **13**:555-556.

Zhang, J., R. Nielsen, and Z. Yang. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol **22**:2472-2479.

Table 1. Summary of *FOXI1* polymorphisms

				Derived Allele Frequency			
Number	Position[a]	Nucleotide change	Ancestral allele	Yoruban (2N=40)	European (2N=40)	Asian (2N=40)	F_{ST}
1	169465818	A/G	G	0.575	0.200	0.350	0.1227
2	169466184	C/T	T	0.225	0.275	0.425	0.0256
3	169466367	C/G	C	0.550	0.225	0.325	0.0921
4	169466686	C/T	T	0.050	0.000	0.250	0.1616
5	169466751	A/G	A	0.875	0.550	0.775	0.1131
6	169466837	A/G	A	0.875	0.550	0.775	0.1131
7	169467111	A/G	A	0.025	0.100	0.000	0.0424
8	169467123	C/T	T	0.225	0.275	0.425	0.0256
9	169467572	A/G	G	0.000	0.025	0.000	0.0000
10	169468070	A/G	G	0.050	0.000	0.000	0.0256
11	169468100	C/T	C	0.100	0.100	0.000	0.0283
12	169468141	A/G	A	0.000	0.000	0.025	0.0000
13	169468312	A/C	C	0.525	0.200	0.325	0.0909
14	169468336	A/T	A	0.050	0.075	0.025	-0.012
15	169468524	C/G	G	0.000	0.000	0.025	0.0000
16	169468633	A/G	G	0.850	0.900	1.000	0.0507
17	169468728	A/G	A	0.825	0.700	0.950	0.0815
18	169468749	A/G	A	0.000	0.025	0.000	0.0000
19	169468769	A/T	T	0.000	0.025	0.000	0.0000
20	169469034	A/T	T	0.025	0.000	0.000	0.0000
21	169469117	C/T	C	0.000	0.075	0.000	0.0513
22	169469123	-/T	-	0.525	0.200	0.325	0.0909

[a] Positions are based on NCBI build 36.3

Table 2. Summary of human *FOXI1* haplotypes

	1	2	3	4	5	6	7	8	10	11	13	14	16	17	21	22				
Anc	G	T	C	T	A	A	A	T	G	C	C	A	G	A	C	-	Europe	Africa	Asia	Total
Ht-01	A	.	G	.	G	G	.	.	.	.	A	.	A	G	.	T	8	19	3	30
Ht-02	A	.	G	C	G	G	.	.	.	.	A	.	A	G	.	T	0	2	10	12
Ht-03	A	C	.	.	G	G	.	C	.	.	.	.	A	G	.	.	0	1	1	2
Ht-04	.	C	.	.	G	G	.	C	.	.	.	.	A	G	.	.	10	8	16	34
Ht-05	.	C	.	.	G	G	.	C	.	.	.	.	A	.	.	.	1	0	0	1
Ht-06	.	.	.	.	G	G	.	.	.	.	.	T	A	G	.	.	3	2	1	6
Ht-07	.	.	.	.	.	.	.	.	.	.	.	.	A	G	.	.	7	1	7	15
Ht-08	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	6	1	2	9
Ht-09	.	.	G	.	.	.	.	.	.	.	.	.	A	.	.	.	1	0	0	1
Ht-10	.	.	.	.	G	G	.	.	.	.	.	.	.	.	.	.	0	2	0	2
Ht-11	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	0	1	0	1
Ht-12	.	.	.	.	.	.	.	.	A	T	.	.	.	.	.	.	0	1	0	1
Ht-13	.	.	.	.	.	.	G	.	.	T	.	.	.	.	.	.	1	1	0	2
Ht-14	.	.	.	.	.	.	G	.	.	T	.	.	.	.	T	.	3	0	0	3
Ht-15	A	.	G	.	G	G	.	.	A	T	.	.	.	.	.	.	0	1	0	1

Note: Each polymorphic variant is displayed below the corresponding ancestral position. Ancestral-like alleles are indicated with dots.

Table 3. Population sequence variation and neutrality test statistics for *FOXI1*

Population	2N[a]	S[b]	π[c]	K[d]	H [e]	Tajima's D	Fu and Li's D*	Fu and Li's D	Fu and Li's F*	Fu and Li's F	Fu's Fs	Fay and Wu's H
European	40	17	0.0012±0.0001	12	0.874±0.026	0.485	0.431	0.433	0.530	0.543	-0.421	1.600
Asian	40	13	0.0010±0.0001	8	0.768±0.045	1.134	0.065	0.030	0.486	0.485	1.969	-1.923
Yoruban	40	16	0.0011±0.0001	12	0.741±0.063	0.487	0.763	0.799	0.793	0.831	-0.664	-2.928
All	120	22	0.0012±0.0001	19	0.837±0.019	0.360	-0.808	-0.863	-0.430	-0.458	-1.417	-0.410

Note: All neutrality statistics were not significant at $P < 0.05$ (see methods)

[a] Number of chromosomes
[b] Number of segregating sites
[c] Nucleotide diversity per base pairs
[d] Total number of haplotypes
[e] Haplotype diversity

Table 4. Core haplotypes with significant REHH values in Africa involving *FOXI1*

	REHH	Frequency	Distance (bp) [a]	Distance (cM) [b]	Core haplotype	Genes in core region [i]	*P* value	*q* value
H=0.02	19.58	0.528	80.443	0.056	ACCC[c]	*DOCK2, FOXI1*	0.8×10^{-5}	0.0038
H=0.04	16.93	0.528	61.086	0.026	ACCC[d]	*DOCK2, FOXI1*	0.3×10^{-5}	0.0016
	12.08	0.590	-64.129	-0.058	AGC[e]	*FOXI1*	0.6×10^{-4}	0.0213
	8.32	0.675	-41.241	-0.051	AG[f]	*DOCK2, FOXI1*	0.8×10^{-4}	0.0247
	10.42	0.528	55.531	0.028	A[g]	*DOCK2, FOXI1*	0.8×10^{-4}	0.0191
	10.65	0.561	-54.967	-0.053	CTG[h]	*DOCK2, FOXI1*	1.2×10^{-4}	0.0215

[a] Physical distance (bp) from the core at which the signal has been captured. (–) indicates downstream direction. otherwise upstream
[b] Genetic distance (cM) from the core at which the signal has been captured. (–) indicates downstream direction. otherwise upstream
[c] rs7736379, rs6872596, rs4449553, rs17562083
[d] rs7736379, rs6872596, rs4449553, rs17562083
[e] rs7729440, rs2879278, rs11134616
[f] rs4867919, rs11134612
[g] rs1501644
[h] rs7709558, rs12515896, rs6861611
[i] Genes within ±100 kb around the core are considered

Legends to figures

Figure 1. Genomic structure of the human *FOXI1* gene and position of detected polymorphisms. The polymorphism indicated in bold type is coding nonsynonymous. The two transcript variants produced by alternative splicing and encoding different isoforms are shown.

Figure 2. Distribution of populational –log p-values of XP-Rsb grouped across main geographical regions. The genome-wide significance of XP-Rsb at every SNP site and population is plotted against distance. Gray rectangles indicate the location of the gene of interest in each region. Dotted and dashed lines show 0.05 and 0.01 significance levels respectively. Values above the latter are additionally represented with solid color circles while open circles indicate values below the 0.01 significance level.

Figure 3. Decay of linkage disequilibrium around *FOXI1* in Africa. (A). Genes and SNPs on the region. Boxes represent genes, vertical gray lines are SNPs, vertical blue lines denote those constituting the core and the vertical red line indicate a non-synonymous SNP. Underlined SNPs represent other cores within the region. (B). Breakdown of EHH over physical distance. (C). Haplotype Bifurcation Plots.

Figure 4. Pattern of linkage disequilibrium in the Yoruba population (CEPH-HGDP) around *FOXI1* (positions 169,401,507 – 169,544,856 on chromosome 5, NCBI build 36.3).

Figure 1.

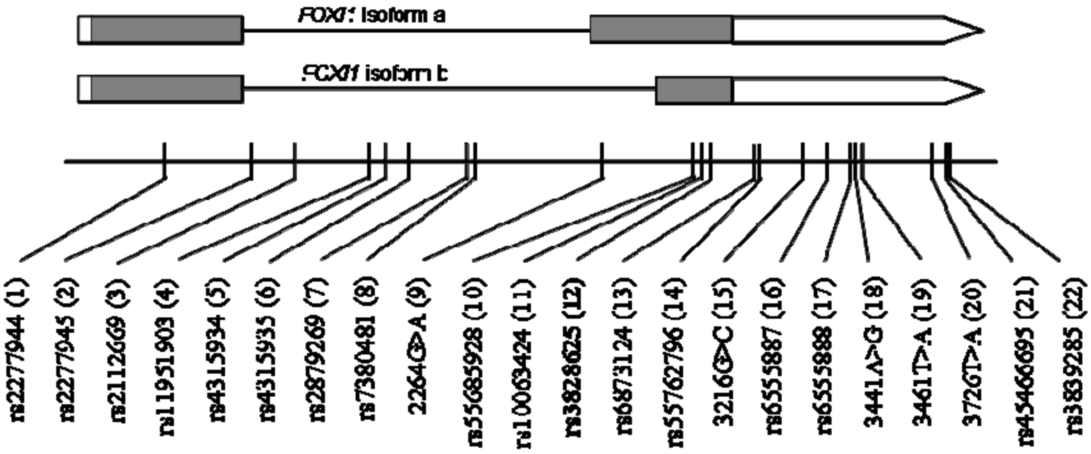

FOXI1 isoform a
FOXI1 isoform b
rs2277944 (1)
rs2277945 (2)
rs2112669 (3)
rs11951903 (4)
rs4315934 (5)
rs4315935 (6)
rs2879269 (7)
rs7380481 (8)
2264G>A (9)
rs55685928 (10)
rs10063424 (11)
rs3828625 (12)
rs6873124 (13)
rs55762796 (14)
3216G>C (15)
rs6555887 (16)
rs6555888 (17)
3441A>G (18)
3461T>A (19)
3726T>A (20)
rs45466695 (21)
rs3839285 (22)

Figure 2.

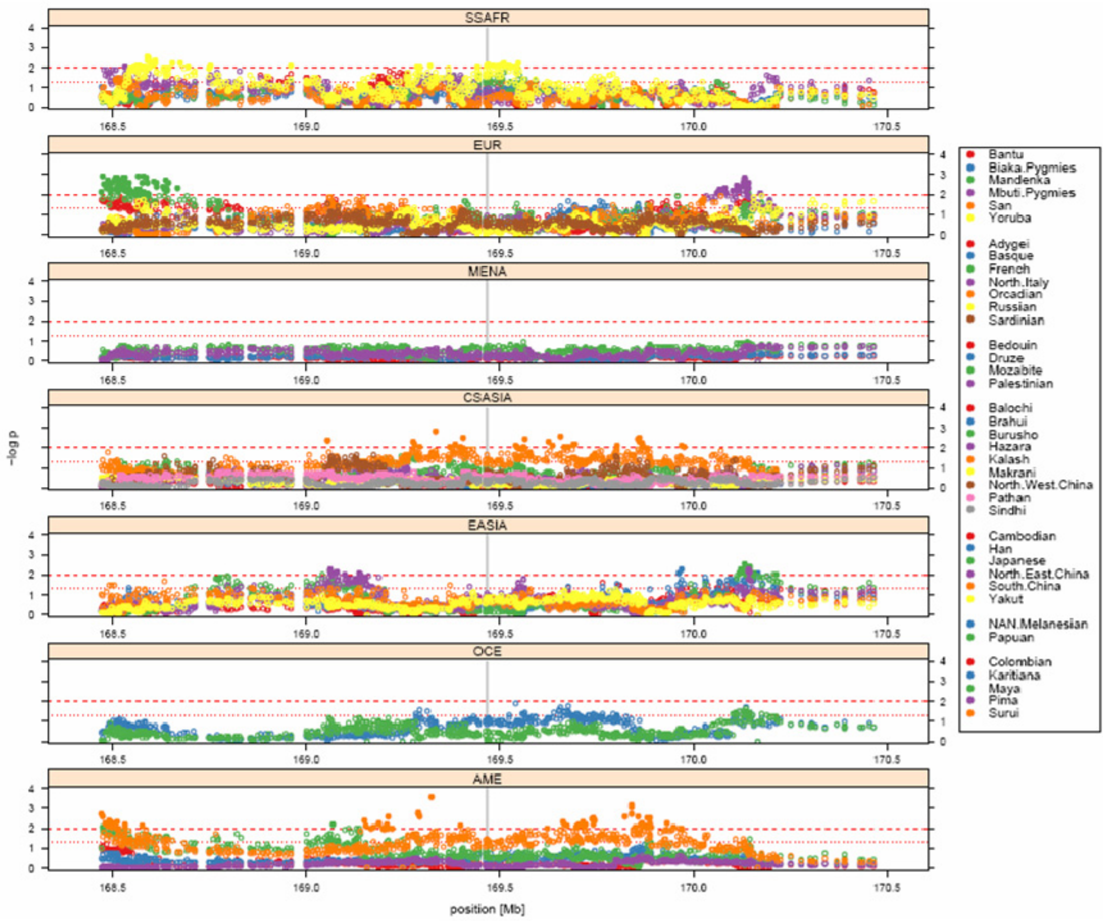

SSAFR
EUR
MENA
CSASIA
EASIA
OCE
AME
0
1
2
3
4
168.5
169.0
169.5
170.0
170.5
-log p
position [Mb]
Bantu
Biaka.Pygmies
Mandenka
Mbuti.Pygmies
San
Yoruba
Adygei
Basque
French
North.Italy
Orcadian
Russian
Sardinian
Bedouin
Druze
Mozabite
Palestinian
Balochi
Brahui
Burusho
Hazara
Kalash
Makrani
North.West.China
Pathan
Sindhi
Cambodian
Han
Japanese
North.East.China
South.China
Yakut
NAN.Melanesian
Papuan
Colombian
Karitiana
Maya
Pima
Surui

Figure 3.

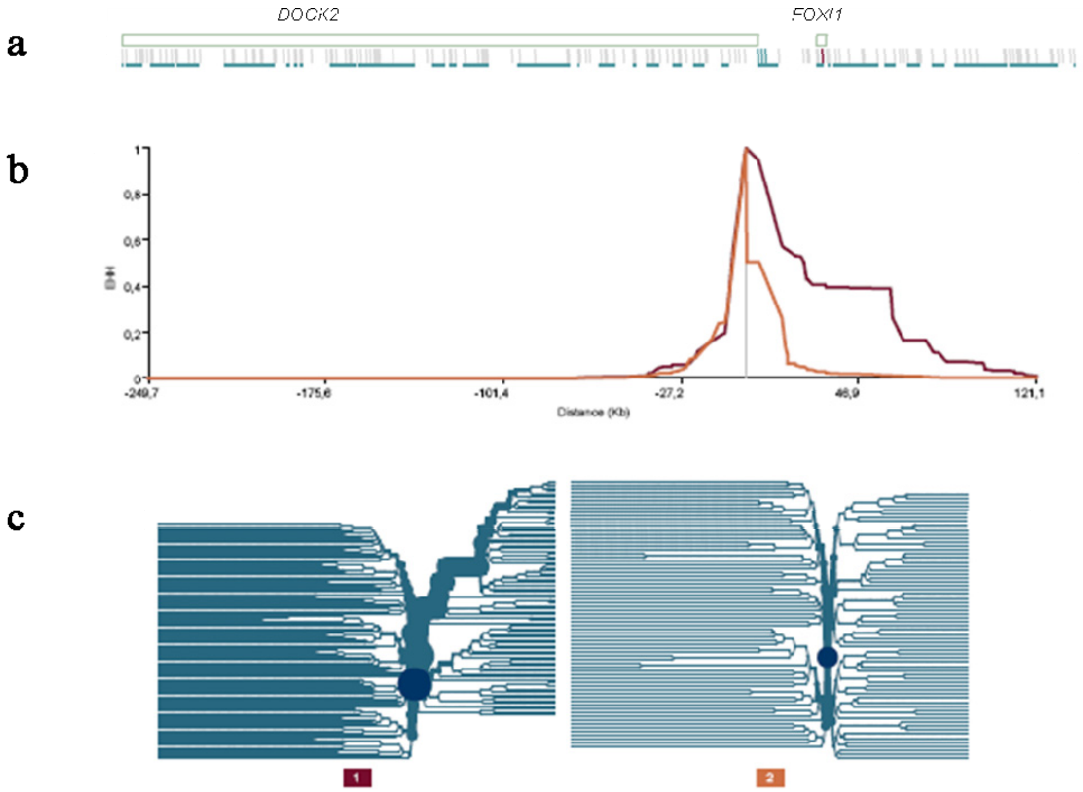
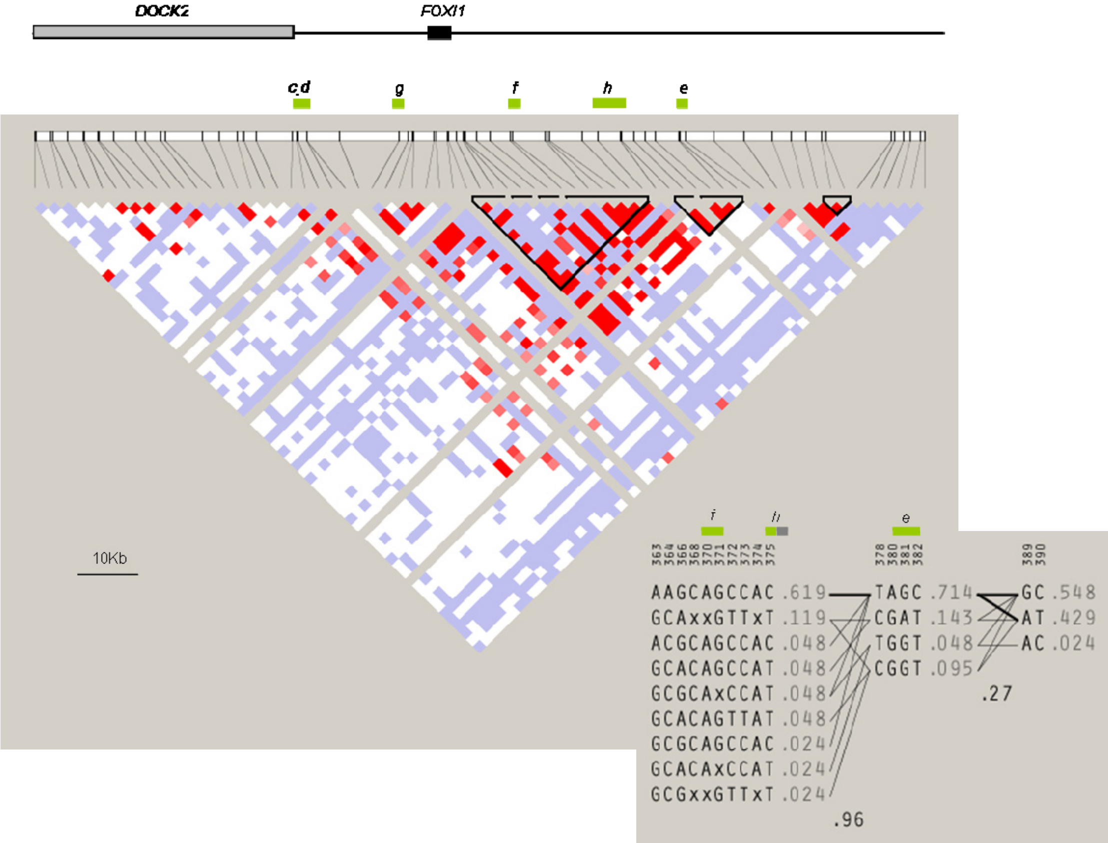

a
DOCK2
FOXI1
b
1
0,8
0,6
0,4
0,2
0
EHH
-249,7
-175,6
-101,4
-27,2
46,9
121,1
Distance (Kb)
c
1
2

Figure 4.


DOCK2
FOXI1
c,d
g
f
h
e
10Kb
f
h
e
363 364 366 368 370 371 372 373 374 375
378 380 381 382
389 390
AAGCAGCCAC .619
GCAxxGTTxT .119
ACGCAGCCAC .048
GCACAGCCAT .048
GCGCAxCCAT .048
GCACAGTTAT .048
GCGCAGCCAC .024
GCACAxCCAT .024
GCGxxGTTxT .024
.96
TAGC .714
CGAT .143
TGGT .048
CGGT .095
.27
GC .548
AT .429
AC .024

DISCUSSION

The results presented in this thesis encompass three different but related studies aimed at detecting signatures of positive natural selection in human genes. First, a genome-wide comparative study was performed to identify a subset of eleven genes with significant accelerated evolution in the human lineage for further investigation of intraspecific SNP variation. Second, two independent single candidate genes were evaluated using both SNP and resequencing data. Overall, combining publicly available and own generated data, nearly 6,000 SNP markers were analyzed in more than 1,000 human samples, and resequencing data for two candidate genes was analyzed in 99 additional individuals. Our main findings will be discussed following the initially envisaged goals of this work.

On the Identification of Positively Selected Genes in the Human Lineage

Selection on the human genome has been studied using comparative genomics and SNP variation in the lineage leading to modern humans. Comparisons of the human and chimpanzee genomes have been used to identify genes subjected to selection on the human lineage (Arbiza et al. 2006a; Bakewell et al. 2007; Bustamante et al. 2005; Clark et al. 2003; Nickel et al. 2008; Nielsen et al. 2005). Such comparative genomic approaches address genetic changes that have occurred during the 4–6 Myr since humans and chimpanzees shared a common ancestor. However, modern humans emerged in Africa less than 200 Kyr ago (Stringer and Andrews 1988) and only began to colonize other continents 50–80 Kyr ago. In connection with the African exodus, human populations have adapted to a range of different environmental conditions. To specifically target genes that have been subjected to selection in association with such more recent evolutionary events, genetic analyses based on intraspecific variation and population comparisons are needed. Following this approach, local selective sweeps have been studied at a genome-wide scale using a number of methods based on population differentiation (Akey et al. 2002; Barreiro et al. 2008; Myles et al. 2008; Weir et al. 2005), haplotype or LD structure (Sabeti et al. 2007; Tang et al. 2007; Voight et al. 2006; Wang et al. 2006), and nucleotide diversity patterns (Carlson et al. 2005; Williamson et al. 2007).

The statistical tests used in each study are thus able to recover selective events from different time periods of human history and for different stages of the selective sweep. For instance, selection may have occurred in Africa after the split from the chimpanzee lineage and prior to the separation of the major population groups (**Figure 39**, human branch), or in any continent and time point after the exodus of non-African populations (**Figure 39**, African, European, and Asian branches). The relationship between genes possibly involved in each of these opposite time periods (i.e., early versus recent evolution), has been poorly investigated in previous studies.

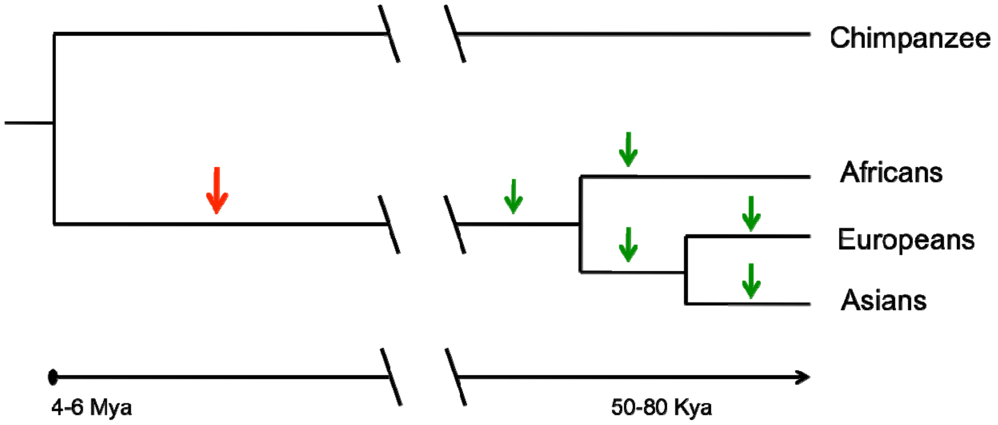


Figure 39: Different stages in which selection could have acted during human evolution. Comparative studies are able to detect selective events occurring in the human branch (red arrow) since the split from the lineage leading to Chimpanzees; whereas intraspecific studies reveal more recent selective events that occurred associated with the evolution of major human population groups (green arrows). Genome scans for selection usually focus in only one of these two opposite time periods.

In **Chapter 1** we compared a genome-wide set of Human-Chimp-Mouse-Rat-Dog orthologous genes to address selection in the human lineage over the past 4–6 Myr, and interrogated a candidate subset for variation in worldwide human populations to address selective sweeps that occurred during the last ~250 Kyr of human evolution.

To create such a subset, we have carefully chosen eleven candidate genes applying a strict quality control to the complete list of significant positively selected genes (see **Materials and Methods** for details). The full list contains 88 significant ($p < 0.05$) genes with evidence of selection in the human lineage (see **Appendix 1**), which we did not further investigate as a whole. However, given the availability of several genome-wide scans based on similar phylogenetic analyses, it is worth to compare our results with such independent studies in order to check for consistency of both the complete list of significant genes and, especially, the subset of genotyped candidate genes. In **Table 10** we show the pairwise overlap of genes between our results at these two levels and other three genome scans applying comparable methods in a varying number of species. The biggest overlap occurs between the two scans with the largest datasets, namely Nickel et al. (2008) and Bakewell et al. (2007), which share 32 genes. Notably, when comparing our study (complete set) with Arbiza et al. (2006), which is the only one based on the same species, we observed greater overlap (i.e., 30 genes), than with the other two scans (i.e., 16 and 18 genes, respectively). This means that ~35% of the genes detected in our study were also detected by an independent scan using the same phylogeny and similar branch-site models. Still, differences in the construction of multiple-species alignments and specific test parameters can account for important between-study discrepancies. When comparing the subset of eleven candidates of our study with Nickel et al. and Bakewell et al. studies (whose accuracy could be greater because they are based on closer related and/or larger number of species), we found seven overlapping genes in both cases and when doing so with Arbiza et al. the overlap stayed roughly the same (i.e., 8 genes). Moreover, these three pairwise comparisons captured a total of 10 different overlapping genes, so that all but one (*OR5G1P*) of our candidate genes are also detected either by Nickel et al. and/or Bakewell et al. and/or Arbiza et al. comparative studies. Overall, this means that we are actually studying variation in genes with confirmed evidence of positive selection on the human lineage.

Table 10: Pairwise overlap of genes with evidence of positive selection in the human branch across comparative genome scans [a]

Study	Nickel *et al.*	Bakewell *et al.*	Arbiza *et al.*	This study (all)	This study (subset)
Nickel *et al.*	242 [b]	32	17	16	7
Bakewell *et al.*		143 [c]	19	18	7
Arbiza *et al.*			107 [d]	30	8
This study (all)				86 [e]	11
This study (subset)					11 [f]

[a] The number of significant genes identified in each study are shaded across the diagonal.

[b] Number of entries with available gene symbol out of 244 significant genes ($p < 0.05$; strict branch-site test implemented in PAML, namely Model A against Model Anull).

[c] Number of entries with available gene symbol out of 154 significant genes ($p < 0.05$; strict branch-site test implemented in PAML).

[d] Number of entries with available gene symbol out of 108 significant genes (Test II of strong positive selection or strict branch-site test implemented in PAML).

[e] Number of entries with available gene symbol out of 88 significant genes ($p < 0.05$; strict branch-site test implemented in PAML).

[f] Number of genes after quality control and included for further analysis in the present study.

Our results reveal that, for eleven positively selected genes in the human lineage, there is no additional evidence of selective sweeps occurring within the time scales up to which population signatures are able to persist (i.e., up to ~250 Kya). Although we did find clear signatures of a recent selective sweep in one of the candidate gene regions (i.e., *VPS37C* gene region), a deeper characterization for functional variation of the region revealed that the downstream immediate neighbor gene (i.e., *CD5* gene) is more likely to be the target of selection.

Our empirical approximation, although limited to a small number of genes, is in agreement with the little overlap, at a genome-wide scale, between the loci reported to be under positive selection within humans and those pinpointed in between-species comparisons. **Table 11** summarizes the overlap in genes with signatures of positive selection between comparative and intraspecific studies as identified by some of the main genome-wide scans in humans with publicly available results. For example, as few as 13 out of 242 genes detected by Nickel et al. (2008) are also detected by any of the intraspecific genome scans analyzed.

The rather small (3-5%) overlap between comparative and intraspecific studies can be explained by different situations. First, there is a large difference in the time perspective of the selective events being sought. Local adaptive selection that affects one or a few human populations may be quite distinct from the selective pressure that shaped modern humans from archaic forms of *Homo*. Many adaptations, such as those resulting from spatial and temporal variation in climate, exposure to pathogens and diet, may have been restricted to particular populations and are therefore likely to remain undetected by comparative genomic studies; and *vice versa*: the accelerated protein evolution of most of the genes implicated in defining our species may have ended with the acquisition of human-specific genetic changes, well before the independent evolution of major human population groups. Second, false positive rates and low statistical power further decreases the probability of overlap. Third, most studies report only the most significant results (e.g., outliers in the 1% empirical distribution). Therefore, the results presented in **Table 11** are probably a conservative estimate of overlap between studies.

Table 11: Pairwise overlap of genes with evidence of positive selection in humans between comparative and intraspecific genome-wide scans

Signal	Study	Nickel *et al.* (n=242)	Bakewell *et al.* (n=143)	Arbiza *et al.* (n=107)	This study[a] (n=86)	This study[b] (n=11)
SFS[k]	Carlson *et al.* (n=176)[c]	2	2	2	0	0
	Williamson *et al* (n=164)[d]	5	1	1	1	1
LD	Voight *et al.* (n=14)[e]	0	0	0	0	0
	Tang *et al.* (n=290)[f]	5	4	1	1	0
	Sabeti *et al.* (n=41)[g]	0	0	0	0	0
F_{ST}	Myles *et al.* (n=34)[h]	0	0	0	0	0
	Barreiro *et al.* (n=59)[i]	1	1	0	1	0
	TOTAL[j]	**13 (5.4%)**	**8 (5.5%)**	**4 (3.7%)**	**3 (3.5%)**	**1**

[a] Referring to the complete set of significant genes identified in the present study

[b] Referring to the subset of genes included in the population analysis of the present study

[c] Number of genes within the 59 Contiguous Regions of Tajima's D Reduction (CRTRs) reported in the study

[d] Number of genes within 100 Kb of the estimate of sweep position in the 101 regions with the strongest evidence of selection according to the study (CLR test; $p < 0.00001$)

[e] Limited list of interesting candidate genes (as proposed by the authors of the study) within the 12 regions showing some of the strongest iHS signals of the genome

[f] Number of positively selected genes in Chinese and Europeans from pairwise comparisons (CE, AC, AE) taking the results of the combined dataset (HapMap + Perlegen) of the study

[g] Number of genes at or near SNPs that fulfill all three criteria of the study within the 22 strongest candidate regions for selection combining the XP-EHH, LRH, and iHS statistics

[h] Number of genes involved in highly differentiated phenotypes between populations (i.e., pigmentation, skeletal development, and carbohydrate metabolism) that overlap with the candidate regions detected in the study

[i] Number of genes containing at least one nonsynonymous SNP or 5'-UTR mutation with F_{ST} > 0.65 as reported in the study

[j] Number of pairwise matches (usually due to different genes). In some cases a gene may have a match in more than one intraspecific scan (See Table 12). In order to keep the same level of analysis across studies, *VPS37C* was counted as positively selected.

[k] Site Frequency Spectrum

On the other hand, as demonstrated in **Chapter 1** and discussed below, caution should be taken when interpreting overlapping results between comparative and intraspecific scans because, unlike the former, candidate genes proposed by statistical tests based on genome variation may not necessarily correspond to the actual target of selection. This is due to the broad area of the genome to which a significant signal of selection can be mapped to, thus making all local genes equally candidates for selection, unless further characterization of the region is performed.

Nonetheless, the small overlapping fraction of the genome between these two approaches is worthy of further consideration for detailed studies like the one presented here. **Table 12** lists the 19 overlapping genes of the genome showing evidence of selection in both comparative studies (as identified by the three aforementioned scans, plus the present study) and intraspecific studies (as identified by seven genome-wide scans capturing three different signals of recent positive selection). Most pairwise overlaps involved genes detected by Nickel et al. (2008), probably due to the larger dataset of this study (n = 242 genes), whereas, interestingly, only intraspecific genome scans with statistical power to detect complete or geographically restricted selective sweeps were involved in such overlaps (see **Table 11**). Notable also is the fact that 15 out of 19 of the overlapping genes show signatures of recent positive selection in Asian populations, whereas only three are restricted to Europeans. Finally, five genes have been identified in both populations (see **Table 12**).

As for the functional classification of these 19 overlapping genes, not surprisingly, immunity and defense, spermatogenesis, and DNA repair are some of the represented categories; but also detoxification and, notably, cation/ion transport appear on the list. The latter, which is essential for maintaining the homeostasis of the organism, is represented by three different genes, including well known cases of selection, such as the *TRPV6* gene (Hughes et al. 2008).

Further studies aimed at dissecting the molecular signature identified around these genes will help to confirm whether they have been targeted by recent selection, thus shedding light on the possible selective pressures that have persisted over long time periods during human evolution.

Table 12: Overlapping genes between comparative and intraspecific genome scans for selection

Gene	Comparative scan	Intraspecific scan	Population[a]	Biological function[b]
ABCC12	Nickel et al.	Barreiro et al.	Asians	Extracellular transport; Detoxification
KCNH7	Nickel et al.	Carlson et al.	Europeans	Cell communication
RANBP2	Nickel et al.; Bakewell et al.	Carlson et al.; Tang et al.	Asians	Nuclear transport; Protein targeting
AMPD3	Nickel et al.; Bakewell et al.; Arbiza et al.	Tang et al.	Asians	Purine metabolism
DNAH5	Nickel et al.	Tang et al.	Europeans; Asians	Spermatogenesis and cell motility
ITIH3	Nickel et al.	Tang et al.	Asians	Proteolysis
MSH4	Nickel et al.	Tang et al.	Asians	DNA repair; Meiosis
ANUBL1	Nickel et al.	Williamson et al.	Asians	Proteolysis
DTNA	Nickel et al.	Williamson et al.	Europeans	Neuromuscular synaptic transmission
IL17E	Nickel et al.	Williamson et al.	Asians	Immunity and defense
SLC9A9	Nickel et al.	Williamson et al.	Europeans; Asians	Cation transport; Homeostasis
SLC39A4	Bakewell et al.	Barreiro et al.	Europeans; Asians	Ion transport
TRPV6	Bakewell et al.	Carlson et al.; Tang et al.	Europeans; Asians	Cation transport
CLSTN2	Bakewell et al.	Tang et al.	Europeans	Cell adhesion-mediated signaling
KIAA0528	Arbiza et al.	Carlson et al.	Asians	Biological process unclassified
PHKB	Arbiza et al.	Carlson et al.	Europeans; Asians	Glycogen metabolism
LIMCH1	This study[c]	Barreiro et al.		Biological process unclassified
PPID	This study[c]	Tang et al.	Asians	Immunity and defense; Protein folding
VPS37C	This study[d]; Nickel et al.; Bakewell et al.; Arbiza et al.	Williamson et al.	Asians	Immunity and defense

[a] **Continental ancestry of the population sample in which selection has been detected by the corresponding intraspecific scan**

[b] **Functional categories based on PANTHER biological processes**

[c] **Referring to the complete set of significant genes identified in the present study**

[d] **Referring to the subset of genes included in the population analysis of the present study**

On the Localization of Signals in Candidate Regions for Positive Selection

Since comparative studies based on protein-coding sequences assess the excess of function-altering mutations between species, they focus on the beneficial alleles themselves, thus eliminating ambiguity about the target of selection. Intraspecific studies, however, seek to detect selective sweeps that affect the patterns of variation around beneficial alleles, which may involve genome regions of different sizes. The size of the region affected by a sweep depends on the strength of positive selection and, thus, the speed at which the selected allele reached high frequency. That is, rapid sweeps affect large regions. If an allele confers a selective advantage of 1% (considered moderately strong selection), the modal size of the affected genomic region has been estimated to be roughly 600 Kb (Sabeti et al. 2006). Although such a large size facilitates signal detection, it also makes the subsequent task of identifying the causal variant more difficult.

The challenge is to examine genetic variation through the candidate regions to identify the variants that are likely to have been the targets of selection. As stated above, candidate regions may span over large stretches of DNA, thus often containing multiple genes and hundreds or even thousands of common SNPs (e.g., MAF > 5%). To help highlight potential targets of selection, heuristic methods have been proposed to scrutinize such candidate regions by combining several signals of recent positive selection (Sabeti et al. 2007).

The signature of recent positive selection with the shortest time scale (i.e., <30,000 years) is given by long-range haplotypes, which may extend over considerable distances, within which many variants will share the EHH signal as they are all reciprocally linked to each other. Nonetheless other signatures may also overlap in the region not only adding support to the evidence of selection but also helping to dissect the possible target of selection. For instance, given that long-range haplotype tests are designed to identify young alleles (newly arisen), selected alleles detectable by EHH-based methods are likely to be derived. Therefore, looking for high-frequency derived alleles enhances the possibility of localizing the signal of selection (see **Figure 40**). In addition, selected alleles are likely to be highly differentiated between populations, because recent selection is probably a local environmental adaptation (Sabeti et al. 2006), thus finding high F_{ST} values for alleles that are common in only the population under selection delimitates the search too. Finally, selected alleles must have biological effects, so the subset of SNPs in a candidate region sharing the aforementioned characteristics, should be functionally characterized, resulting in an even smaller subset of possible causal variants (see **Figure 40**).

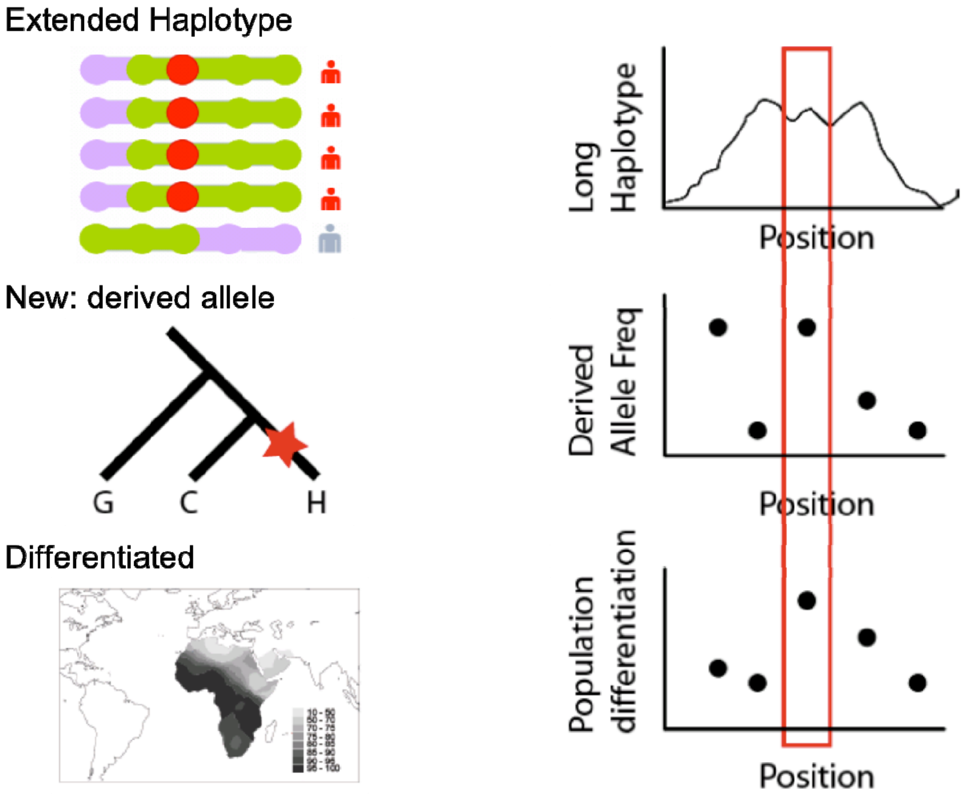


Figure 40: Localization of signals of selection in a candidate region. Five polymorphisms rising to high frequency along a positively selected (red) allele are shown within a candidate region characterized by extended haplotypes, high-frequency derived alleles and geographically restricted alleles. Only a subset SNPs in the region will share these characteristics, and an even smaller subset will be functional (modified from (Sabeti et al. 2007)).

In **Chapter 1** we analyzed eleven 2-Mb candidate regions with five different methods (i.e., MAF and DAF threshold analyses, F_{ST} statistic, LRH test, and XP-Rsb statistic) capturing four different signatures of recent positive selection (i.e., excess of rare and high-frequency derived alleles, population differentiation, and long-range haplotypes, respectively). Our candidate regions, although centered on the gene of interest, often contained multiple genes in their vicinity. On average, they harbor 460 SNPs, most of which resulted to be common in most populations of the HGDP-CEPH Diversity Panel. However, populations from Oceania and the Americas showed a higher proportion of fixed alleles, which was due to extensive drift in the demographic history of these continents. The population richness of the HGDP-CEPH Diversity Panel allowed us to detect geographically restricted patterns of genomic variation across these candidate regions, thus leading to more detailed understandings of selective forces acting in different regions of the world.

Our results highlighted one single candidate region (i.e., the *VPS37C* region) harboring several clear signatures of recent positive selection especially in Asian populations. The different signals were mostly concentrated within ± 0.5 Mb around the *VPS37C* locus, in which we found an extended haplotype spanning ~420 Kb at high frequencies (> 60%) in one population from Central-South Asia, most East Asian populations and some from the Americas. The highest frequency (75%) was found in Han Chinese and Japanese populations, followed by South Chinese, Northeast Chinese, Pima from Mexico, Karitiana from Brazil, and Colombian (70%). Notably, this long haplotype is completely absent in Sub-Saharan populations.

Following the heuristic criteria described above we identified a C/T SNP (rs2229177) located in the last exon of the *CD5* gene (< 5 Kb apart from *VPS37C*) that: i) is embedded within the extended haplotype in the derived (T) allele state ii) is particularly frequent in Asians while not elsewhere, that is, highly differentiated (global F_{ST} = 0.27 for a tagSNP available in the HGDP-CEPH panel), and iii) it codes for a A471V amino acid substitution that is likely to alter the protein structure and function of the *CD5* receptor, which is clearly involved in the regulation of important immune responses.

Our results also demonstrate how the location of the target of selection not necessarily corresponds to the physical positions with the highest test statistics. Williamson and colleagues (2007) reported *VPS37C* as the closest gene to their physical estimate of the position of the sweep according to a composite likelihood method specifically designed to locate recent selective sweeps in the human genome. Nonetheless, they also reported the *CD5* gene as being 18 Kb apart from such estimate. The CLR test is particularly sensitive to the detection of alleles that have already reached fixation. Our putatively selected SNP (rs2229177) is indeed fixed or nearly so in Chinese and Japanese HapMap samples, and several tagSNPs for rs2229177 are fixed too in most HGDP East Asian populations. In the latter, however, we observed that the complete range of fixed alleles in the vicinity goes from rs11820879 (located ~14 Kb downstream of rs2229177) up to rs12799829 (located ~80 Kb upstream of rs2229177). For selection tests such as the CLR test, the center of this window (which coincides with the location of the *VPS37C* gene) will show the highest local test statistic, thus explaining the apparent discrepancy between their estimate of sweep position and our candidate target of selection. This also reinforces the importance that should be given to physical ranges associated with reported estimates of the location of selective sweeps in the genome. On the other hand, Williamson et al. (2007) genome scan was performed using Perlegen data, thus interrogating only three human populations (i.e., African-American, European-American and Chinese). Their results, however, are in agreement with ours in the sense that the selective sweep observed in the *VPS37C* region was found to be restricted to the Chinese sample, although additional Asian or New World populations were not studied.

Investigators from the same research group have recently applied an extension of the CLR test to analyze genotype data from the Population Reference Sample (POPRES) project (Nelson et al. 2008), which consists of 500K SNP arrays genotyped in more than 6,000 ethnically diverse samples, including a large collection of Mexican individuals from Guadalajara, Mexico. When looking at the *CD5-VPS37C* genome region, significant likelihood values comparable to those reported in Asians, were found in the Mexican sample (Wright et al. unpublished results). This could be in agreement with the fact that we also found the putatively selected long-range haplotype at high frequencies not only in Asian populations but also in some Native American populations, including a Pima population from Mexico. Nonetheless, the complexity of the genetic structure of admixed Mexican populations does not enable us to make such direct conclusions. According to a Principal Component Analysis (PCA) based on dense genome-wide SNP data from the POPRES samples, Mexican individuals show varying degrees of admixture between Europeans and a presumably (unsampled) population that likely corresponds to Native Americans (Nelson et al. 2008). In a PCA including European, East Asian, and Mexican POPRES samples, as well as HGDP Native American samples, Auton et al. (2009, In press), found that the least admixed Mexican individuals appear far away from East Asians, suggesting that there is substantial genetic differentiation between East Asian and Native American populations. This is not surprising since archeological and genetic evidence (Fagundes et al. 2008; Mulligan et al. 2004; Perego et al. 2009; Tamm et al. 2007and Sandoval et al. in review) suggests that human populations from East Asia reached the Americas between 15-20 Kya with little subsequent gene flow (until recent historical times). More importantly, such a long span of independent evolutionary history of the New World populations gives complete support to the possibility of finding local genetic adaptations imprinted in Native American genomes. Apart from very few exceptions, positive natural selection in Native American populations still remains largely unexplored. Although with the availability of next-generation sequencing and current genotyping technologies, large-scale genetic studies are beginning to pay attention on previously excluded regions of the world.

Finally, further studies on the candidate region highlighted in this work will allow confirmation of both the signal and the target of selection. The former can be confirmed by resequencing the *CD5* gene region in a number of human samples from geographically diverse populations and by applying the many available neutrality tests designed for DNA sequence data. The latter, as discussed below, is more challenging as structural and, if possible, experimental studies are needed to assess the functional impact of the candidate target of selection at the phenotypic level.

On the Analysis from Candidate to Function

Identifying candidate sites for positive selection is only one step in understanding the physiological basis of adaptation. Going beyond candidates requires a detailed analysis of molecular structure and an understanding of how changes in structure relate to changes in function. There are still relatively few examples in which molecular evolutionary changes have been linked to physiological adaptations in the organism. The identification of the sickle cell mutation in the *HBB* gene as having been the target of selection for malaria resistance represents one of the earliest successful dissections of adaptive evolution. Many years of investigation were required to unravel the association with malaria and the biochemical properties of the sickle cell mutation (Allison 1954; Ingram 1959; Pauling et al. 1949). Even nowadays, it is not completely clear the exact mechanism by which the sickle state inhibits malaria infection.

Positive selection at a specific locus can be dissected following two approaches: finding a DNA change with functional molecular consequence or finding an association to a phenotypic difference in the human population. The first approach requires good genetic annotation of the region, including both coding and regulatory regions, and can be enhanced by comparative genomic approaches. Accordingly, the functional changes might be found through comparisons between species, between populations, or between haplotypes. Such clues can be the basis of diverse means of biological experimentation. The second approach, then, usually depends on knowledge of the underlying biology of the region. If the selected variant is still polymorphic in humans, the associated phenotype might be measured in human populations (e.g., malaria resistance) or in cell lines (e.g., protein function or gene expression levels). If the selected variant has become species-specific, transgenic techniques can be used to produce somewhat "humanized" mice in which the human-specific trait could be measured. Transgenic mouse lines can also be generated to display population-specific phenotypes, such as in the case of the *EDAR* gene variant that produces the typical East Asian hair morphology (Bryk et al. 2008; Mou et al. 2008).

Following the first approach, in **Chapter 1** we presented evidence of a recent selective sweep occurring in the *CD5* gene region and proposed the A471V amino acid substitution as having been the target of selection. The A471V substitution maps to the cytoplasmic region of the CD5 receptor, which although constituting the smallest part of the protein, it seems to be the functionally essential one (Pena-Rossi et al. 1999). In consistency with this, we also found that the cytoplasmic region is highly conserved across species, as opposite to the highly variable regions corresponding to the extracellular domains of the protein (see **Figure 41**).

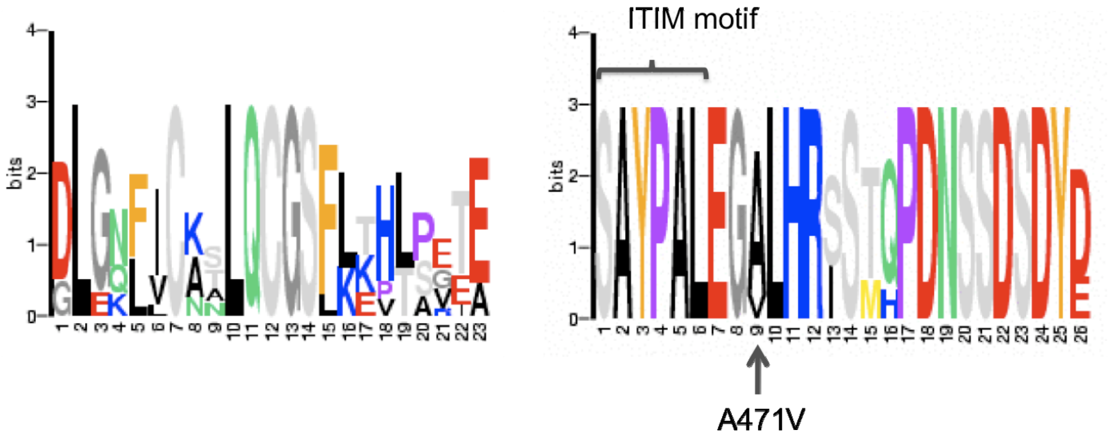


Figure 41: Degree of amino acid sequence conservation along the CD5 receptor protein. Sequence logos are proportional to their frequency and the total height of all residues at each position is proportional to the conservation (alignment based on 10 sequences from six species (i.e., human, mouse, rat, rabbit, sheep, and bovin)). Two representative motifs are shown: motif 2 (*left*) corresponding to the extracellular domain, which shows low conservation levels, and motif 10 (*right*) located in the cytoplasmic domain, which is highly conserved and contains the A471V substitution (*arrow*). The immunoreceptor tyrosine-based inhibitory motif (ITIM) next to A471V is also shown (modified from Blocks Database v14.3, entry IPR003566).

The CD5 molecule belongs to an expanding family of immune inhibitory receptors that share (in the cytoplasmic domain) a consensus amino acid sequence, the immunoreceptor tyrosine-based inhibitory motif (ITIM) (Ravetch and Lanier 2000). The conservation of ITIMs during evolution further evidences their functional relevance. As shown in **Figure 41**, the 6–amino acid ITIM sequence in the CD5 molecule is located only two amino acids apart from the human-specific A471V substitution. An amino acid replacement in such a close location from a functionally relevant motif is likely to affect the structural conformation and, thus, the biological activity of the receptor. In fact, the functional effect of this particular SNP has been predicted as “possibly damaging” according to PolyPhen (Ramensky et al. 2002), and as “pathological” according to prediction methods based on selective pressures at a codon level (Arbiza et al. 2006b). A three-dimensional model of the CD5 protein structure with and without the aforementioned amino acid substitution would help to confirm these predictions. However, such reconstruction was not possible because currently available crystal structures of molecules that would serve as templates for the CD5 protein, exclude the most distant portion of the cytoplasmic domain, exactly where the A471V substitution is located.

Given the complexity of the immune system, it is even harder to associate the A471V replacement on *CD5* to a particular phenotypic difference in human populations. We know that the derived V allele characterizes the extended haplotype segregating at high frequencies in Asia and the Americas; but it remains unknown whether this variant disrupts the inhibitory activity of the receptor or confers a new function, although the former seems more plausible.

CD5 is a T-cell marker also expressed at various developmental and activation stages on human B-cells. It is a well established negative regulator of T-cell receptor (TCR) and B-cell receptor (BCR) signaling pathways. Using animal models, it has been demonstrated that increased expression of *CD5* on either T-cells or B-cells protects against autoimmunity; as a consequence of an increase of the threshold needed for TCR or BCR-mediated activation following antigen recognition. This means that, while TCR and BCR receptors are responsible for responding against molecules that are alien to the organism, the CD5 receptor limits the extent of such response by inactivating the pathway when no longer required (see **Figure 42**). Autoimmune disorders, thus, may result from the disruption of inhibitory receptors (Dalloul 2009). In a review of the medical literature, however, we did not find any differential epidemiology for autoimmune diseases between Asians or Native Americans and populations from elsewhere.

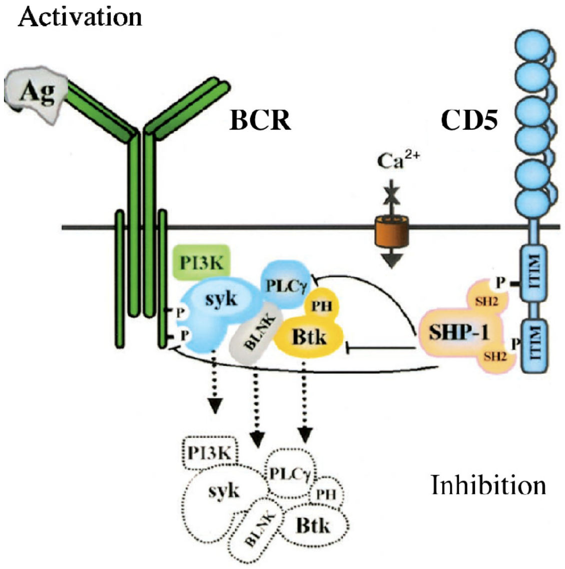


Figure 42: Inhibition of the B-cell antigen receptor (BCR) by the CD5 inhibitory receptor. CD5 inhibitory mechanism is mediated by SHP-1, which is recruited to interact with the ITIMs and ultimately downregulate the BCR signaling pathway (dotted lines) initially activated by antigen recognition (adapted from (Ravetch and Lanier 2000)).

Interestingly, although the physiological role of *CD5* might be to control the generation of aberrant immune responses, the expansion of CD5+ lymphocytes may be deleterious under certain conditions. Particularly, in different types of cancer, *CD5* expression plays a role in the fate of tumor-specific T-cells, rendering lymphocytes tolerant and unable to recognize and eliminate malignant cells. This means that the downregulation of *CD5* enhances the anti-tumoral potential of the immune system, thus providing a plausible explanation for the mechanism by which the silencing of *CD5* could have been favored by selection. If this is the case, individuals carrying the silenced version of the receptor may not be protected against autoimmune diseases, but at the same time, they may be more prepared to kill malignant cells and survive cancer. This assumption also raises the possibility that positive selection in the *CD5* locus may be a type of balancing selection.

In support to the aforementioned mechanism, it has been reported that chronic lymphocytic leukemia (CLL), a malignancy characterized by the accumulation of CD5+ B-cells (Goldin et al. 2004), is common in Europeans but uncommon in Asians and particularly rare in Chinese (Gale et al. 2000 and references therein). This different incidence is in agreement with the observed frequencies of the A471V-containing extended haplotype across different populations, which is most frequent in Asia, particularly in Chinese and Japanese. Moreover, in a genome scan of 755 genes with relevance to cancer biology in 425 patients with CLL, Sellick and colleagues (2008) identified the A471V polymorphism in *CD5* as a genetic variant determinant of survival. In particular, they reported that homozygosity for the ancestral A allele is associated with a poorer prognosis of the disease, supporting the possible protective role of the derived V allele against malignancies such as CLL.

Given the speculative nature of these assumptions, experimental confirmation is required. One possibility would be to construct cell lines, or even transgenic animals, in which one could measure any differential immunological response between populations transfected with each of the A471V alleles. It is exciting to think that this locus may have spontaneously represented for nature an example of what we would call today a therapeutic target for the development of antineoplastic drugs.

Similarly, in **Chapter 2** we presented evidence for positive selection acting on a human olfactory receptor (OR) gene, namely *OR5I1*, and proposed the F76S amino acid substitution as having been the target of selection. As in the former case, the F76S polymorphism is located in a highly conserved region of the protein, namely the second transmembrane domain (see **Figure 43**). For OR proteins, it is known that several amino acids in the various transmembrane helices form a ligand-binding pocket (Katada et al. 2005). Based on human-mouse comparisons, Man et al. (2004) predicted a number of particular sequence positions as binding site residues of OR proteins. Placing the F76S polymorphism of *OR5I1* in Man's structural homology model, we observed that it lies next to one of the predicted odorant binding sites (see **Figure 43**). Besides this observation, additional predictions of the functional effect of the F76S replacement included a large physicochemical Grantham distance (i.e., 155) (Grantham 1974), and a PolyPhen prediction of "possibly damaging".

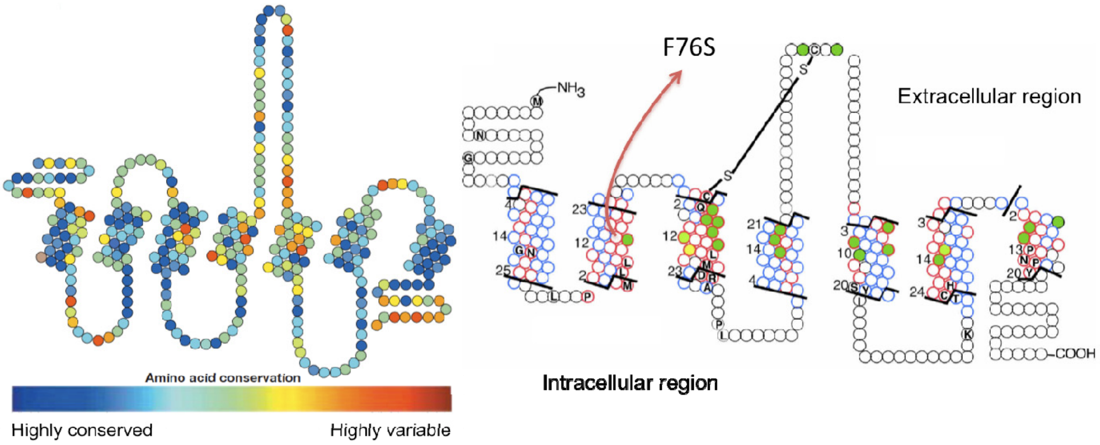


Figure 43: Schematic diagrams of the OR protein structure. *Left*: degree of amino acid conservation in the consensus sequence of ORs represented as a color spectrum, with blue being highly conserved and red highly variable (adapted from (Mombaerts 2004)) . *Right*: The predicted binding site residues (green shaded circles) proposed by Man et al. (2004) depicted on a typical OR sequence. Black lines represent the predicted location of the cellular membrane. The residue corresponding to the F76S substitution is indicated in the second transmembrane domain (red arrow).

Unlike with *CD5*, in this case we were able to further evaluate the structural impact of the different amino acid replacements defining the main human forms of the *OR5I1* protein, by means of comparative protein modeling using the bovine rhodopsin crystal structure as template. The structural analysis allowed us to rule out all other amino acid replacements as candidates for destabilizing the protein structure, whereas F76S implied a significant structural change indeed, not only in the amino acid position itself but also in neighbor residues, thus modifying the spatial configuration and the interaction among the first and the second transmembrane domains. This observation led us to propose that F76S is the most plausible target of selection.

As discussed in **Chapter 2**, we suggest the possibility that the amino acid replacement caused by the F76S polymorphism in *OR5I1* could have led to the formation of a novel human-specific binding site; although the identification of the particular odorant being recognized, and the ultimately favored phenotype, remains, so far, unknown.

The detection of chemicals in the external environment is essential for the survival of the individual and of the species. This ancient sense (known as chemosensation) enables animals to locate nutritious food and suitable mating partners, and to avoid being eaten by predators or eating toxic substances. Chemosensory systems (smell, taste) are distinguished from the other senses (vision, hearing, touch) by the qualitative heterogeneity of the stimuli: the chemical senses are responsible for detecting molecules of immense chemical variety. This task of broad chemical recognition requires a massive repertoire of receptors to match the diversity in chemical structures. Olfactory receptors represent indeed the largest gene family in mammalian genomes (Gilad et al. 2003b). More than 1,000 genes have been identified as part of the human olfactory subgenome, but few receptor–ligand interactions have been characterized. For air-breathing organisms such as humans, odorants must be sufficiently volatile to be detected by the nose. A large fraction of volatile chemicals have a discernable odor, placing the number of detectable chemicals in the range of hundreds of thousands. The number of odors that can be detected by the human nose is often quoted as 10,000 (Mombaerts 2004). Such a large number of possible ligands makes it extremely difficult to link a particular odorant with a newly arisen variant of an olfactory receptor.

Finally, in **Chapter 3** we present evidence for possible African signatures of recent positive selection in the *FOXI1* gene, and characterized ~140 Kb for functional variation. This resulted in the identification of 19 functionally relevant SNPs, out of which three are significantly differentiated between Africans and non-Africans, and were therefore proposed as the best candidates for underlying selection on this gene. The selected phenotype, however, can not be sought in a single scenario, because *FOXI1*, a transcription factor, has been implicated in the physiology of at least three different organs: inner ear, testis, and kidney (Vidarsson et al. 2009). Interestingly, *FOXI1* has been reported to be required for male fertility, as epididymal sperm maturation seems to be *FOXI1*-dependent (Blomqvist et al. 2006). Most mammalian spermatozoa are not capable to move progressively or to fertilize an oocyte at the time when they leave testis. To acquire these abilities, they require a post-testicular maturation process that to a large extent takes place in epididymis. At the molecular level, such maturation depends on active proton (H^+) secretion into the epididymal lumen, and *FOXI1* acts as an important regulator of gene expression in the major proton secretory cells of epididymal epithelia. Since this regulatory mechanism involves the vacuolar H^+ -ATPase proton pump, which is not exclusive of epididymal cells, *FOXI1* activity is also essential in other tissues such as the endolymphatic epithelium of the inner ear, and the collecting ducts of the kidney. In the former, FORE cells secretes protons into endolymph, a process important for maintaining appropriate ionic composition which in turn is vital for conversion of acoustic sound waves into neuronal action potentials, a process critical for hearing (Hulander et al. 2003). In the latter, intercalated cells of the kidney distal tubuli and collecting ducts depend on proton transport for maintaining proper systemic acid/base homeostasis (Karet et al. 1999). Not surprisingly, mutations in *FOXI1* have been associated with human syndromes involving deafness, distal renal tubular acidosis and male infertility (Blomqvist et al. 2006).

Genes that affect fertility and reproduction are expected to be subject to strong selection owing to their direct effects on fitness-related traits. This class of proteins is often reported as rapidly evolving by genome scans among and within species. Moreover, cases of differential fertility among human populations have been reported. Stefansson et al. (2005) showed that female carriers of a large inversion polymorphism, which shows evidence of selection based on population differentiation and long-range haplotype structure, have higher fertility than noncarriers. However, for the particular case of *FOXI1*, our results pointed out that a number of core haplotypes with significant REHH in Africans are significantly correlated with latitude, suggesting that climate may have driven selection on this gene. In humans, adaptation to different climates may have entailed adapting kidney function to different dehydration levels, in which *FOXI1* may have played a critical role trough its regulatory mechanism affecting water homeostasis. Given its multiple physiological implications, further evolutionary and functional studies on *FOXI1* variation are needed to assess whether the observed signatures of selection in African populations are indeed related to climate adaptation.

Overall, the results presented in this thesis illustrate three particular cases of positive selection in the human genome, which involve different individual genes that, not casually belong to some of the most represented functional categories under positive selection in humans, namely immunity (*CD5*), olfaction (*OR5I1*) and homeostasis or reproduction (*FOXI1*). This is notable since the present work was not intended to focus in any previously defined biological function of the human body.

Concluding Remarks

From a genome-wide perspective, our results suggest that most of the genes implicated in selective events during early human evolution differ from those involved in recent human adaptations. Although probably affecting similar functional categories, the particular targeted phenotypes, and thus the genes underlying them, may have been different along the varying stages of human evolutionary history as a result of new encountered environments, pathogens and food resources as humans colonized the world. Such progressive differentiation of the human branch (as species) into thinner branches (populations) reinforces the idea of uniqueness at both levels, in which we, individuals, represent the tips of the tree of life.

From a candidate gene perspective, our results highlighted at least three different genomic regions in which, although no relation with earlier selective events in the human lineage was found, there is strong evidence for a selective sweep occurring within the past ~80,000 years of human evolution. These consist of three previously undissected cases of particular genes undergoing positive selection in the human genome. In one of them (i.e., *OR5I1*) the signal is shared among all human populations, thus likely occurring right before the dispersal of modern humans; whereas in the other two (i.e., *CD5* and *FOXI1*) the signal is geographically restricted to East Asian and African populations, respectively. This implies that positive selection acting on these regions must have occurred after the separation of the major population human groups. For each genomic region, detailed characterization for functional variation was provided, and when appropriate, specific targets of selection were proposed.

As demonstrated by the in-depth analysis of these three cases, the relevance of studying natural selection goes beyond providing knowledge for understanding human nature. Finding variants that have been subject to selection can provide insights about those genes influencing human phenotypic variability, including differences in health and disease with profound epidemiological implications. This is true for at least one of the strongest candidates for selection, namely the *CD5* gene, which is clearly implicated in immune-related disorders.

The most challenging issue is to provide convincing evidence that particular genetic changes proposed to be subjected to natural selection, are actually involved in defining a population-specific or human-specific trait, that is, that they actually contribute to adaptive phenotypes, and ultimately to make us humans.

Despite we believe to be applying the most comprehensive genetic analyses to dissect human-specific adaptive changes, it would be too simplistic to think that humanness is limited to genetics. Nonetheless it certainly represents a milestone in understanding human evolution.

BIBLIOGRAPHY

References

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: 1805-14

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P (2007) Molecular Biology of the Cell, Fifth edn

Allison AC (1954) Protection afforded by sickle-cell trait against subtertian malareal infection. Br Med J 1: 290-4

Andres AM, Soldevila M, Navarro A, Kidd KK, Oliva B, Bertranpetit J (2004) Positive selection in MAOA gene is human exclusive: determination of the putative amino acid change selected in the human lineage. Hum Genet 115: 377-86

Arbiza L, Dopazo J, Dopazo H (2006a) Positive selection, relaxation, and acceleration in the evolution of the human and chimp genome. PLoS Comput Biol 2: e38

Arbiza L, Duchi S, Montaner D, Burguet J, Pantoja-Uceda D, Pineda-Lucena A, Dopazo J, Dopazo H (2006b) Selective pressures at a codon-level predict deleterious mutations in human disease genes. J Mol Biol 358: 1390-404

Bakewell MA, Shi P, Zhang J (2007) More genes underwent positive selection in chimpanzee evolution than in human evolution. Proc Natl Acad Sci U S A 104: 7489-94

Bamshad MJ, Mummidi S, Gonzalez E, Ahuja SS, Dunn DM, Watkins WS, Wooding S, Stone AC, Jorde LB, Weiss RB, Ahuja SK (2002) A strong signature of balancing selection in the 5' cis-regulatory region of CCR5. Proc Natl Acad Sci U S A 99: 10539-44

Barbujani G, Magagni A, Minch E, Cavalli-Sforza LL (1997) An apportionment of human DNA diversity. Proc Natl Acad Sci U S A 94: 4516-9

Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L (2008) Natural selection has driven population differentiation in modern humans. Nat Genet 40: 340-5

Barsh GS (2003) What controls variation in human skin color? PLoS Biol 1: E27

Barton NH, Briggs DEG, Eisen JA, Goldstein DB, Patel NH (2007) EVOLUTION, New York, USA

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IM, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DM, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Chiara ECM, Chang S, Neil Cooley R, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin

SS, Ewan MR, Fedurco M, Fraser LJ, Fuentes Fajardo KV, Scott Furey W, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT, Horgan AM, et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456: 53-9

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN (2004) Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet 74: 1111-20

Biswas S, Akey JM (2006) Genomic insights into positive selection. Trends Genet 22: 437-46

Blomqvist SR, Vidarsson H, Soder O, Enerback S (2006) Epididymal expression of the forkhead transcription factor Foxi1 is required for male fertility. Embo J 25: 4131-41

Bond J, Roberts E, Mochida GH, Hampshire DJ, Scott S, Askham JM, Springell K, Mahadevan M, Crow YJ, Markham AF, Walsh CA, Woods CG (2002) ASPM is a major determinant of cerebral cortical size. Nat Genet 32: 316-20

Botchkarev VA, Fessing MY (2005) Edar signaling in the control of hair follicle development. J Investig Dermatol Symp Proc 10: 247-51

Boyd R, Silk JB (2004) How Humans Evolved, New York, USA

Bryk J, Hardouin E, Pugach I, Hughes D, Strotmann R, Stoneking M, Myles S (2008) Positive selection in East Asians for an EDAR allele that enhances NF-kappaB activation. PLoS ONE 3: e2209

Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD, Civello D, Adams MD, Cargill M, Clark AG (2005) Natural selection on protein-coding genes in the human genome. Nature 437: 1153-7

Calafell F, Shuster A, Speed WC, Kidd JR, Kidd KK (1998) Short tandem repeat polymorphism evolution in humans. Eur J Hum Genet 6: 38-49

Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-Tamir B, Cambon-Thomsen A, Chen Z, Chu J, Carcassi C, Contu L, Du R, Excoffier L, Ferrara GB, Friedlaender JS, Groot H, Gurwitz D, Jenkins T, Herrera RJ, Huang X, Kidd J, Kidd KK, Langaney A, Lin AA, Mehdi SQ, Parham P, Piazza A, Pistillo MP, Qian Y, Shu Q, Xu J, Zhu S, Weber JL, Greely HT, Feldman MW, Thomas G, Dausset J, Cavalli-Sforza LL (2002) A human genome diversity cell line panel. Science 296: 261-2

Cann RL, Stoneking M, Wilson AC (1987) Mitochondrial DNA and human evolution. Nature 325: 31-6

Carlson CS, Thomas DJ, Eberle MA, Swanson JE, Livingston RJ, Rieder MJ, Nickerson DA (2005) Genomic regions exhibiting positive selection identified from dense genotype data. Genome Res 15: 1553-65

Carroll SB (2003) Genetics and the making of Homo sapiens. Nature 422: 849-57

Cavalli-Sforza LL, Piazza A, Menozzi P (1994) History and Geography of Human Genes, Princeton

Clark AG (2005) Hot spots unglued. Nat Genet 37: 563-4

Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, Todd MA, Tanenbaum DM, Civello D, Lu F, Murphy B, Ferriera S, Wang G, Zheng X, White TJ, Sninsky JJ, Adams MD, Cargill M (2003) Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science 302: 1960-3

Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R (2005) Ascertainment bias in studies of human genome-wide polymorphism. Genome Res 15: 1496-502

Conde L, Vaquerizas JM, Dopazo H, Arbiza L, Reumers J, Rousseau F, Schymkowitz J, Dopazo J (2006) PupaSuite: finding functional single nucleotide polymorphisms for large-scale genotyping purposes. Nucleic Acids Res 34: W621-5

Consortium CSaA (2005a) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437: 69-87

Consortium IHGS (2004) Finishing the euchromatic sequence of the human genome. Nature 431: 931-45

Consortium TIH (2005b) A haplotype map of the human genome. Nature 437: 1299-320

Currat M, Excoffier L, Maddison W, Otto SP, Ray N, Whitlock MC, Yeaman S (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: 172; author reply 172

Currat M, Trabuchet G, Rees D, Perrin P, Harding RM, Clegg JB, Langaney A, Excoffier L (2002) Molecular analysis of the beta-globin gene cluster in the Niokholo Mandenka population reveals a recent origin of the beta(S) Senegal mutation. Am J Hum Genet 70: 207-23

Chamary JV, Parmley JL, Hurst LD (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat Rev Genet 7: 98-108

Chen FC, Vallender EJ, Wang H, Tzeng CS, Li WH (2001) Genomic divergence between human and chimpanzee estimated from large-scale alignments of genomic sequences. J Hered 92: 481-9

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001) High-resolution haplotype structure in the human genome. Nat Genet 29: 229-32

Dalloul A (2009) CD5: a safeguard against autoimmunity and a shield for cancer cells. Autoimmun Rev 8: 349-53

De la Vega FM, Lazaruk KD, Rhodes MD, Wenz MH (2005) Assessment of two flexible and compatible SNP genotyping platforms: TaqMan SNP Genotyping Assays and the SNPlex Genotyping System. Mutat Res 573: 111-35

Enard W, Paabo S (2004) Comparative primate genomics. Annu Rev Genomics Hum Genet 5: 351-78

Enard W, Przeworski M, Fisher SE, Lai CS, Wiebe V, Kitano T, Monaco AP, Paabo S (2002) Molecular evolution of FOXP2, a gene involved in speech and language. Nature 418: 869-72

Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Jarvela I (2002) Identification of a variant associated with adult-type hypolactasia. Nat Genet 30: 233-7

Evans PD, Anderson JR, Vallender EJ, Gilbert SL, Malcom CM, Dorus S, Lahn BT (2004) Adaptive evolution of ASPM, a major determinant of cerebral cortical size in humans. Hum Mol Genet 13: 489-94

Evans PD, Gilbert SL, Mekel-Bobrov N, Vallender EJ, Anderson JR, Vaez-Azizi LM, Tishkoff SA, Hudson RR, Lahn BT (2005) Microcephalin, a gene regulating brain size, continues to evolve adaptively in humans. Science 309: 1717-20

Fagundes NJ, Kanitz R, Eckert R, Valls AC, Bogo MR, Salzano FM, Smith DG, Silva WA, Jr., Zago MA, Ribeiro-dos-Santos AK, Santos SE, Petzl-Erler ML, Bonatto SL (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: 583-92

Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. Genetics 155: 1405-13

Feuk L, Marshall CR, Wintle RF, Scherer SW (2006) Structural variants: changing the landscape of chromosomes and design of disease studies. Hum Mol Genet 15 Spec No 1: R57-66

Fisher SE, Vargha-Khadem F, Watkins KE, Monaco AP, Pembrey ME (1998) Localisation of a gene implicated in a severe speech and language disorder. Nat Genet 18: 168-70

Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Shen Y, Sun W, Wang H, Wang Y, Wang Y, Xiong X, Xu L, Waye MM, Tsui SK, Xue H, Wong JT, Galver LM, Fan JB, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier JF, Phillips MS, Roumy S, Sallee C, Verner A, Hudson TJ, Kwok PY, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui LC, Mak W, Song YQ, Tam PK, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851-61

Fu YX (1996) New statistical tests of neutrality for DNA samples from a population. Genetics 143: 557-70

Fu YX (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics 147: 915-25

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

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D (2002) The structure of haplotype blocks in the human genome. Science 296: 2225-9

Gale RP, Cozen W, Goodman MT, Wang FF, Bernstein L (2000) Decreased chronic lymphocytic leukemia incidence in Asians in Los Angeles County. Leuk Res 24: 665-9

Gilad Y, Bustamante CD, Lancet D, Paabo S (2003a) Natural selection on the olfactory receptor gene family in humans and chimpanzees. Am J Hum Genet 73: 489-501

Gilad Y, Man O, Paabo S, Lancet D (2003b) Human specific loss of olfactory receptor genes. Proc Natl Acad Sci U S A 100: 3324-7

Gilad Y, Rosenberg S, Przeworski M, Lancet D, Skorecki K (2002) Evidence for positive selection and population structure at the human MAO-A gene. Proc Natl Acad Sci U S A 99: 862-7

Glazko GV, Nei M (2003) Estimation of divergence times for major lineages of primate species. Mol Biol Evol 20: 424-34

Goldin LR, Pfeiffer RM, Li X, Hemminki K (2004) Familial risk of lymphoproliferative tumors in families of patients with chronic lymphocytic leukemia: results from the Swedish Family-Cancer Database. Blood 104: 1850-4

Gould JS (1981) The Mismeasure of Man, New York, London

Graf J, Hodgson R, van Daal A (2005) Single nucleotide polymorphisms in the MATP gene are associated with normal human pigmentation variation. Hum Mutat 25: 278-84

Grantham R (1974) Amino acid difference formula to help explain protein evolution. Science 185: 862-4

Graur D, Li W-H (2000) Fundamentals of Molecular Evolution, Second edn

Green RE, Krause J, Ptak SE, Briggs AW, Ronan MT, Simons JF, Du L, Egholm M, Rothberg JM, Paunovic M, Paabo S (2006) Analysis of one million base pairs of Neanderthal DNA. Nature 444: 330-6

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: 1669-79

Hamblin MT, Thompson EE, Di Rienzo A (2002) Complex signatures of natural selection at the Duffy blood group locus. Am J Hum Genet 70: 369-83

Harding RM, Healy E, Ray AJ, Ellis NS, Flanagan N, Todd C, Dixon C, Sajantila A, Jackson IJ, Birch-Machin MA, Rees JL (2000) Evidence for variable selective pressures at MC1R. Am J Hum Genet 66: 1351-61

Harpending HC, Batzer MA, Gurven M, Jorde LB, Rogers AR, Sherry ST (1998) Genetic traces of ancient demography. Proc Natl Acad Sci U S A 95: 1961-7

Hartl DL, Clark AG (2007) Principles of Population Genetics, 4th edn, Sunderland, Massachusetts

Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR (2005) Whole-genome patterns of common DNA variation in three human populations. Science 307: 1072-9

Hollox EJ, Poulter M, Zvarik M, Ferak V, Krause A, Jenkins T, Saha N, Kozlov AI, Swallow DM (2001) Lactase haplotype diversity in the Old World. Am J Hum Genet 68: 160-172

Hudson RR, Kreitman M, Aguade M (1987) A test of neutral molecular evolution based on nucleotide data. Genetics 116: 153-9

Hughes AL, Yeager M (1998) Natural selection at major histocompatibility complex loci of vertebrates. Annu Rev Genet 32: 415-35

Hughes DA, Tang K, Strotmann R, Schoneberg T, Prenen J, Nilius B, Stoneking M (2008) Parallel selection on TRPV6 in human populations. PLoS ONE 3: e1686

Hulander M, Kiernan AE, Blomqvist SR, Carlsson P, Samuelsson EJ, Johansson BR, Steel KP, Enerback S (2003) Lack of pendrin expression leads to deafness and expansion of the endolymphatic compartment in inner ears of Foxi1 null mutant mice. Development 130: 2013-25

Huxley TH (1863) Evidence as to Man's Place in Nature. In: Norgate Wa (ed), London

Ingram VM (1959) Abnormal human haemoglobins. III. The chemical difference between normal and sickle cell haemoglobins. Biochim Biophys Acta 36: 402-11

Izagirre N, Garcia I, Junquera C, de la Rua C, Alonso S (2006) A scan for signatures of positive selection in candidate loci for skin pigmentation in humans. Mol Biol Evol 23: 1697-706

Jackson AP, Eastwood H, Bell SM, Adu J, Toomes C, Carr IM, Roberts E, Hampshire DJ, Crow YJ, Mighell AJ, Karbani G, Jafri H, Rashid Y, Mueller RF, Markham AF, Woods CG (2002) Identification of microcephalin, a protein implicated in determining the size of the human brain. Am J Hum Genet 71: 136-42

Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC, Szpiech ZA, Degnan JH, Wang K, Guerreiro R, Bras JM, Schymick JC, Hernandez DG, Traynor BJ, Simon-Sanchez J, Matarin M, Britton A, van de Leemput J, Rafferty I, Bucan M, Cann HM, Hardy JA, Rosenberg NA, Singleton AB (2008) Genotype, haplotype and copy-number variation in worldwide human populations. Nature 451: 998-1003

Jobling MA, Hurles ME, Tyler-Smith C (2004) Human Evolutionary Genetics: origins, peoples and disease., London/New York

Jorde LB, Watkins WS, Bamshad MJ, Dixon ME, Ricker CE, Seielstad MT, Batzer MA (2000) The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. Am J Hum Genet 66: 979-88

Kaessmann H, Heissig F, von Haeseler A, Paabo S (1999) DNA sequence variation in a non-coding region of low recombination on the human X chromosome. Nat Genet 22: 78-81

Kaessmann H, Paabo S (2002) The genetical history of humans and the great apes. J Intern Med 251: 1-18

Kaessmann H, Wiebe V, Weiss G, Paabo S (2001) Great ape DNA sequences reveal a reduced diversity and an expansion in humans. Nat Genet 27: 155-6

Kaiser J (2008) DNA sequencing. A plan to capture human diversity in 1000 genomes. Science 319: 395

Karet FE, Finberg KE, Nelson RD, Nayir A, Mocan H, Sanjad SA, Rodriguez-Soriano J, Santos F, Cremers CW, Di Pietro A, Hoffbrand BI, Winiarski J, Bakkaloglu A, Ozen S, Dusunsel R, Goodyer P, Hulton SA, Wu DK, Skvorak AB, Morton CC, Cunningham MJ, Jha V, Lifton RP (1999) Mutations in the gene encoding B1 subunit of H+-ATPase cause renal tubular acidosis with sensorineural deafness. Nat Genet 21: 84-90

Katada S, Hirokawa T, Oka Y, Suwa M, Touhara K (2005) Structural basis for a broad but selective ligand spectrum of a mouse olfactory receptor: mapping the odorant-binding site. J Neurosci 25: 1806-15

Kayser M, Brauer S, Stoneking M (2003) A genome scan to detect candidate regions influenced by local natural selection in human populations. Mol Biol Evol 20: 893-900

Kelley JL, Swanson WJ (2008) Positive selection in the human genome: from genome scans to biological significance. Annu Rev Genomics Hum Genet 9: 143-60

Kidd JR, Pakstis AJ, Zhao H, Lu RB, Okonofua FE, Odunsi A, Grigorenko E, Tamir BB, Friedlaender J, Schulz LO, Parnas J, Kidd KK (2000) Haplotypes and linkage disequilibrium at the phenylalanine hydroxylase locus, PAH, in a global representation of populations. Am J Hum Genet 66: 1882-99

Kidd KK, Pakstis AJ, Speed WC, Kidd JR (2004) Understanding human DNA sequence variation. J Hered 95: 406-20

Kimura M (1968) Evolutionary rate at the molecular level. Nature 217: 624-6

Kimura M (1983) The neutral theory of molecular evolution. , Cambridge

Kimura R, Fujimoto A, Tokunaga K, Ohashi J (2007) A practical genome scan for population-specific strong selective sweeps that have reached fixation. PLoS ONE 2: e286

King JL, Jukes TH (1969) Non-Darwinian evolution. Science 164: 788-98

Kouprina N, Pavlicek A, Mochida GH, Solomon G, Gersch W, Yoon YH, Collura R, Ruvolo M, Barrett JC, Woods CG, Walsh CA, Jurka J, Larionov V (2004) Accelerated evolution of the ASPM gene controlling brain size begins prior to human brain expansion. PLoS Biol 2: E126

Krause J, Lalueza-Fox C, Orlando L, Enard W, Green RE, Burbano HA, Hublin JJ, Hanni C, Fortea J, de la Rasilla M, Bertranpetit J, Rosas A, Paabo S (2007) The derived FOXP2 variant of modern humans was shared with Neandertals. Curr Biol 17: 1908-12

Kuehn BM (2008) 1000 Genomes Project promises closer look at variation in human genome. Jama 300: 2715

Kunz S, Rojek JM, Kanagawa M, Spiropoulou CF, Barresi R, Campbell KP, Oldstone MB (2005a) Posttranslational modification of alpha-dystroglycan, the cellular receptor for arenaviruses, by the glycosyltransferase LARGE is critical for virus binding. J Virol 79: 14282-96

Kunz S, Rojek JM, Perez M, Spiropoulou CF, Oldstone MB (2005b) Characterization of the interaction of lassa fever virus with its cellular receptor alpha-dystroglycan. J Virol 79: 5979-87

Lahr MM, Foley RA (1998) Towards a theory of modern human origins: geography, demography, and diversity in recent human evolution. Am J Phys Anthropol Suppl 27: 137-76

Lai CS, Fisher SE, Hurst JA, Vargha-Khadem F, Monaco AP (2001) A forkhead-domain gene is mutated in a severe speech and language disorder. Nature 413: 519-23

Lamason RL, Mohideen MA, Mest JR, Wong AC, Norton HL, Aros MC, Jurynec MJ, Mao X, Humphreville VR, Humbert JE, Sinha S, Moore JL, Jagadeeswaran P, Zhao W, Ning G, Makalowska I, McKeigue PM, O'Donnell D, Kittles R, Parra EJ, Mangini NJ, Grunwald DJ, Shriver MD, Canfield VA, Cheng KC (2005) SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. Science 310: 1782-6

Landegren U, Kaiser R, Sanders J, Hood L (1988) A ligase-mediated gene detection technique. Science 241: 1077-80

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921

Lao O, de Gruijter JM, van Duijn K, Navarro A, Kayser M (2007) Signatures of positive selection in genes associated with human skin pigmentation as revealed from analyses of single nucleotide polymorphisms. Ann Hum Genet 71: 354-69

Lao O, Lu TT, Nothnagel M, Junge O, Freitag-Wolf S, Caliebe A, Balascakova M, Bertranpetit J, Bindoff LA, Comas D, Holmlund G, Kouvatsi A, Macek M, Mollet I, Parson W, Palo J, Ploski R, Sajantila A, Tagliabracci A, Gether U, Werge T, Rivadeneira F, Hofman A, Uitterlinden AG, Gieger C, Wichmann HE, Ruther A, Schreiber S, Becker C, Nurnberg P, Nelson MR, Krawczak M, Kayser M (2008) Correlation between genetic and geographic structure in Europe. Curr Biol 18: 1241-8

Lewontin RC (1964) The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. Genetics 49: 49-67

Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics 74: 175-95

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, Myers RM (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319: 1100-4

Livingstone FB (1984) The Duffy blood groups, vivax malaria, and malaria selection in human populations: a review. Hum Biol 56: 413-25

MacDermot KD, Bonora E, Sykes N, Coupe AM, Lai CS, Vernes SC, Vargha-Khadem F, McKenzie F, Smith RL, Monaco AP, Fisher SE (2005) Identification of FOXP2 truncation as a novel cause of developmental speech and language deficits. Am J Hum Genet 76: 1074-80

Man O, Gilad Y, Lancet D (2004) Prediction of the odorant binding site of olfactory receptor proteins by human-mouse comparisons. Protein Sci 13: 240-54

McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in Drosophila. Nature 351: 652-4

Mekel-Bobrov N, Gilbert SL, Evans PD, Vallender EJ, Anderson JR, Hudson RR, Tishkoff SA, Lahn BT (2005) Ongoing adaptive evolution of ASPM, a brain size determinant in Homo sapiens. Science 309: 1720-2

Mombaerts P (2004) Genes and ligands for odorant, vomeronasal and taste receptors. Nat Rev Neurosci 5: 263-78

Mou C, Thomason HA, Willan PM, Clowes C, Harris WE, Drew CF, Dixon J, Dixon MJ, Headon DJ (2008) Enhanced ectodysplasin-A receptor (EDAR) signaling alters multiple fiber characteristics to produce the East Asian hair form. Hum Mutat 29: 1405-11

Mulligan CJ, Hunley K, Cole S, Long JC (2004) Population genetics, history, and health patterns in native americans. Annu Rev Genomics Hum Genet 5: 295-315

Myers S, Bottolo L, Freeman C, McVean G, Donnelly P (2005) A fine-scale map of recombination rates and hotspots across the human genome. Science 310: 321-4

Myles S, Bouzekri N, Haverfield E, Cherkaoui M, Dugoujon JM, Ward R (2005) Genetic evidence in support of a shared Eurasian-North African dairying origin. Hum Genet 117: 34-42

Myles S, Somel M, Tang K, Kelso J, Stoneking M (2007) Identifying genes underlying skin pigmentation differences among human populations. Hum Genet 120: 613-21

Myles S, Tang K, Somel M, Green RE, Kelso J, Stoneking M (2008) Identification and analysis of genomic regions with large between-population differentiation in humans. Ann Hum Genet 72: 99-110

Nelson MR, Bryc K, King KS, Indap A, Boyko AR, Novembre J, Briley LP, Maruyama Y, Waterworth DM, Waeber G, Vollenweider P, Oksenberg JR, Hauser SL, Stirnadel HA, Kooner JS, Chambers JC, Jones B, Mooser V, Bustamante CD, Roses AD, Burns DK, Ehm MG, Lai EH (2008) The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. Am J Hum Genet 83: 347-58

Nickel GC, Tefft D, Adams MD (2008) Human PAML browser: a database of positive selection on human genes using phylogenetic methods. Nucleic Acids Res 36: D800-8

Nielsen R (2005) Molecular signatures of natural selection. Annu Rev Genet 39: 197-218

Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, Fledel-Alon A, Tanenbaum DM, Civello D, White TJ, J JS, Adams MD, Cargill M (2005) A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol 3: e170

Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG (2007) Recent and ongoing selection in the human genome. Nat Rev Genet 8: 857-68

Noonan JP, Coop G, Kudaravalli S, Smith D, Krause J, Alessi J, Chen F, Platt D, Paabo S, Pritchard JK, Rubin EM (2006) Sequencing and analysis of Neanderthal genomic DNA. Science 314: 1113-8

Norton HL, Kittles RA, Parra E, McKeigue P, Mao X, Cheng K, Canfield VA, Bradley DG, McEvoy B, Shriver MD (2007) Genetic evidence for the convergent evolution of light skin in Europeans and East Asians. Mol Biol Evol 24: 710-22

Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR, Stephens M, Bustamante CD (2008) Genes mirror geography within Europe. Nature 456: 98-101

Ohashi J, Naka I, Patarapotikul J, Hananantachai H, Brittenham G, Looareesuwan S, Clark AG, Tokunaga K (2004) Extended linkage disequilibrium surrounding the hemoglobin E variant due to malarial selection. Am J Hum Genet 74: 1198-208

Olson MV, Varki A (2003) Sequencing the chimpanzee genome: insights into human evolution and disease. Nat Rev Genet 4: 20-8

Oota H, Pakstis AJ, Bonne-Tamir B, Goldman D, Grigorenko E, Kajuna SL, Karoma NJ, Kungulilo S, Lu RB, Odunsi K, Okonofua F, Zhukova OV, Kidd JR, Kidd KK (2004) The evolution and population genetics of the ALDH2 locus: random genetic drift, selection, and low levels of recombination. Ann Hum Genet 68: 93-109

Osier MV, Pakstis AJ, Soodyall H, Comas D, Goldman D, Odunsi A, Okonofua F, Parnas J, Schulz LO, Bertranpetit J, Bonne-Tamir B, Lu RB, Kidd JR, Kidd KK (2002) A global perspective on genetic variation at the ADH genes reveals unusual patterns of linkage disequilibrium and diversity. Am J Hum Genet 71: 84-99

Paabo S (2003) The mosaic that is our genome. Nature 421: 409-12

Pauling L, Itano HA, et al. (1949) Sickle cell anemia, a molecular disease. Science 109: 443

Pena-Rossi C, Zuckerman LA, Strong J, Kwan J, Ferris W, Chan S, Tarakhovsky A, Beyers AD, Killeen N (1999) Negative regulation of CD4 lineage development and responses by CD5. J Immunol 163: 6494-501

Pennisi E (2007) Breakthrough of the year. Human genetic variation. Science 318: 1842-3

Perego UA, Achilli A, Angerhofer N, Accetturo M, Pala M, Olivieri A, Kashani BH, Ritchie KH, Scozzari R, Kong QP, Myres NM, Salas A, Semino O, Bandelt HJ, Woodward SR, Torroni A (2009) Distinctive Paleo-Indian migration routes from Beringia marked by two rare mtDNA haplogroups. Curr Biol 19: 1-8

Przeworski M, Hudson RR, Di Rienzo A (2000) Adjusting the focus on human variation. Trends Genet 16: 296-302

Quintana-Murci L, Semino O, Bandelt HJ, Passarino G, McElreavey K, Santachiara-Benerecetti AS (1999) Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. Nat Genet 23: 437-41

Ramensky V, Bork P, Sunyaev S (2002) Human non-synonymous SNPs: server and survey. Nucleic Acids Res 30: 3894-900

Ravetch JV, Lanier LL (2000) Immune inhibitory receptors. Science 290: 84-9

Rees JL (2003) Genetics of hair and skin color. Annu Rev Genet 37: 67-90

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, Lander ES (2001) Linkage disequilibrium in the human genome. Nature 411: 199-204

Reich DE, Schaffner SF, Daly MJ, McVean G, Mullikin JC, Higgins JM, Richter DJ, Lander ES, Altshuler D (2002) Human genome sequence variation and the influence of gene history, mutation and recombination. Nat Genet 32: 135-42

Rooney AP, Zhang J (1999) Rapid evolution of a primate sperm protein: relaxation of functional constraint or positive Darwinian selection? Mol Biol Evol 16: 706-10

Rosenberg NA (2006) Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. Ann Hum Genet 70: 841-7

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW (2002) Genetic structure of human populations. Science 298: 2381-5

Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, Ackerman HC, Campbell SJ, Altshuler D, Cooper R, Kwiatkowski D, Ward R, Lander ES (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832-7

Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES (2006) Positive natural selection in the human lineage. Science 312: 1614-20

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, Schaffner SF, Lander ES, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Shen Y, Sun W, Wang H, Wang Y, Wang Y, Xiong X, Xu L, Waye MM, Tsui SK, Xue H, Wong JT, Galver LM, Fan JB, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier JF, Phillips MS, Roumy S, Sallee C, Verner A, Hudson TJ, Kwok PY, Cai D, Koboldt DC, Miller RD, Pawlikowska L, et al. (2007) Genome-wide detection and characterization of positive selection in human populations. Nature 449: 913-8

Sabeti PC, Walsh E, Schaffner SF, Varilly P, Fry B, Hutcheson HB, Cullen M, Mikkelsen TS, Roy J, Patterson N, Cooper R, Reich D, Altshuler D, O'Brien S, Lander ES (2005) The case for selection at CCR5-Delta32. PLoS Biol 3: e378

Saunders MA, Hammer MF, Nachman MW (2002) Nucleotide variability at G6pd and the signature of malarial selection in humans. Genetics 162: 1849-61

Sellick GS, Wade R, Richards S, Oscier DG, Catovsky D, Houlston RS (2008) Scan of 977 nonsynonymous SNPs in CLL4 trial patients for the identification of genetic variants influencing prognosis. Blood 111: 1625-33

Serre D, Paabo S (2004) Evidence for gradients of human genetic diversity within and among continents. Genome Res 14: 1679-85

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

Soranzo N, Bufe B, Sabeti PC, Wilson JF, Weale ME, Marguerie R, Meyerhof W, Goldstein DB (2005) Positive selection on a high-sensitivity allele of the human bitter-taste receptor TAS2R16. Curr Biol 15: 1257-65

Stefansson H, Helgason A, Thorleifsson G, Steinthorsdottir V, Masson G, Barnard J, Baker A, Jonasdottir A, Ingason A, Gudnadottir VG, Desnica N, Hicks A, Gylfason A, Gudbjartsson DF, Jonsdottir GM, Sainz J, Agnarsson K, Birgisdottir B, Ghosh S, Olafsdottir A, Cazier JB, Kristjansson K, Frigge ML, Thorgeirsson TE, Gulcher JR, Kong A, Stefansson K (2005) A common inversion under selection in Europeans. Nat Genet 37: 129-37

Stringer C, Andrews P (2005) The Complete World of Human Evolution, London

Stringer CB, Andrews P (1988) Genetic and fossil evidence for the origin of modern humans. Science 239: 1263-8

Sulem P, Gudbjartsson DF, Stacey SN, Helgason A, Rafnar T, Magnusson KP, Manolescu A, Karason A, Palsson A, Thorleifsson G, Jakobsdottir M, Steinberg S, Palsson S, Jonasson F, Sigurgeirsson B, Thorisdottir K, Ragnarsson R, Benediktsdottir KR, Aben KK, Kiemeney LA, Olafsson JH, Gulcher J, Kong A, Thorsteinsdottir U, Stefansson K (2007) Genetic determinants of hair, eye and skin pigmentation in Europeans. Nat Genet 39: 1443-52

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

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585-95

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: e829

Tang K, Thornton KR, Stoneking M (2007) A New Approach for Using Genome Scans to Detect Recent Positive Selection in the Human Genome. PLoS Biol 5: e171

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673-80

Tishkoff SA, Dietzsch E, Speed W, Pakstis AJ, Kidd JR, Cheung K, Bonne-Tamir B, Santachiara-Benerecetti AS, Moral P, Krings M (1996) Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science 271: 1380-7

Tishkoff SA, Kidd KK (2004) Implications of biogeography of human populations for 'race' and medicine. Nat Genet 36: S21-7

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

Tishkoff SA, Varkonyi R, Cahinhinan N, Abbes S, Argyropoulos G, Destro-Bisol G, Drousiotou A, Dangerfield B, Lefranc G, Loiselet J, Piro A, Stoneking M, Tagarelli A, Tagarelli G, Touma EH, Williams SM, Clark AG (2001) Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. Science 293: 455-62

Tishkoff SA, Verrelli BC (2003) Patterns of human genetic diversity: implications for human evolutionary history and disease. Annu Rev Genomics Hum Genet 4: 293-340

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, et al. (2001) The sequence of the human genome. Science 291: 1304-51

Verrelli BC, McDonald JH, Argyropoulos G, Destro-Bisol G, Froment A, Drousiotou A, Lefranc G, Helal AN, Loiselet J, Tishkoff SA (2002) Evidence for balancing selection from nucleotide sequence analyses of human G6PD. Am J Hum Genet 71: 1112-28

Verrelli BC, Tishkoff SA (2004) Signatures of selection and gene conversion associated with human color vision variation. Am J Hum Genet 75: 363-75

Vidarsson H, Westergren R, Heglind M, Blomqvist SR, Breton S, Enerback S (2009) The forkhead transcription factor Foxi1 is a master regulator of vacuolar H-ATPase proton pump subunits in the inner ear, kidney and epididymis. PLoS ONE 4: e4471

Vigilant L, Stoneking M, Harpending H, Hawkes K, Wilson AC (1991) African populations and the evolution of human mitochondrial DNA. Science 253: 1503-7

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

Wang ET, Kodama G, Baldi P, Moyzis RK (2006) Global landscape of recent inferred Darwinian selection for Homo sapiens. Proc Natl Acad Sci U S A 103: 135-40

Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Zhang J, Guo Y, Feng B, Li H, Lu Y, Fang X, Liang H, Du Z, Li D, Zhao Y, Hu Y, Yang Z, Zheng H, Hellmann I, Inouye M, Pool J, Yi X, Zhao J, Duan J, Zhou Y, Qin J, Ma L, Li G, Yang Z, Zhang G, Yang B, Yu C, Liang F, Li W, Li S, Li D, Ni P, Ruan J, Li Q, Zhu H, Liu D, Lu Z, Li N, Guo G, Zhang J, Ye J, Fang L, Hao Q, Chen Q, Liang Y, Su Y, San A, Ping C, Yang S, Chen F, Li L, Zhou K, Zheng H, Ren Y, Yang L, Gao Y, Yang G, Li Z, Feng X, Kristiansen K, Wong GK, Nielsen R, Durbin R, Bolund L, Zhang X, Li S, Yang H, Wang J (2008) The diploid genome sequence of an Asian individual. Nature 456: 60-5

Wang YQ, Su B (2004) Molecular evolution of microcephalin, a gene determining human brain size. Hum Mol Genet 13: 1131-7

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyras E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA, Fulton RS, Furey TS, Gage D, Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigo R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A, Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones TA, Joy A, Kamal M, Karlsson EK, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520-62

Weir BS, Cardon LR, Anderson AD, Nielsen DM, Hill WG (2005) Measures of human population structure show heterogeneity among genomic regions. Genome Res 15: 1468-76

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song XZ, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM (2008) The complete genome of an individual by massively parallel DNA sequencing. Nature 452: 872-6

Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, Nielsen R (2007) Localizing recent adaptive evolution in the human genome. PLoS Genet 3: e90

Wolpoff MH (1996) Interpretations of multiregional evolution. Science 274: 704-7

Wong WS, Nielsen R (2004) Detecting selection in noncoding regions of nucleotide sequences. Genetics 167: 949-58

Wooding SP, Watkins WS, Bamshad MJ, Dunn DM, Weiss RB, Jorde LB (2002) DNA sequence variation in a 3.7-kb noncoding sequence 5' of the CYP1A2 gene: implications for human population history and natural selection. Am J Hum Genet 71: 528-42

Wright S (1969) The Theory of Gene Frequencies. In: Press CUoC (ed) Evolution and the Genetics of Populations, vol 2

Wyckoff GJ, Wang W, Wu CI (2000) Rapid evolution of male reproductive genes in the descent of man. Nature 403: 304-9

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13: 555-6

Yang Z, Nielsen R (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol 19: 908-17

Yu F, Hill RS, Schaffner SF, Sabeti PC, Wang ET, Mignault AA, Ferland RJ, Moyzis RK, Walsh CA, Reich D (2007) Comment on "Ongoing adaptive evolution of ASPM, a brain size determinant in Homo sapiens". Science 316: 370

Zhang, Nielsen R, Yang Z (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol 22: 2472-9

Zhang C, Bailey DK, Awad T, Liu G, Xing G, Cao M, Valmeekam V, Retief J, Matsuzaki H, Taub M, Seielstad M, Kennedy GC (2006) A whole genome long-range haplotype (WGLRH) test for detecting imprints of positive selection in human populations. Bioinformatics 22: 2122-8

Zhang J (2004) Frequent false detection of positive selection by the likelihood method with branch-site models. Mol Biol Evol 21: 1332-9

APPENDIX

Appendix 1: List of significant (p < 0.05) positively selected genes in the human lineage

Ensembl Gene ID	Gene symbol	L null	L alternative	X2 difference	P-value Chisq	Codons (posterior probability ≥ 95%)
ENSG00000127603	MACF1	23307,21152	22613,75535	1386,912348	1,4669E-303	Not displayed (>1,000 codons)
ENSG00000124201	ZNFX1	6271,848791	6120,775017	302,147548	1,12176E-67	
ENSG00000162897	FCAMR	4830,535144	4749,074425	162,921438	2,60246E-37	357 0.966 // 359 0.986 // 362 0.971 // 363 0.961 // 366 0.986 // 368 0.950 // 369 0.990 // 376 0.997 // 378 0.988 // 383 0.985 // 384 0.985
ENSG00000197705	KLHL14	3382,630646	3309,919795	145,421702	1,73687E-33	215 0.964 // 216 1.000 // 217 0.968 // 219 0.998 // 220 0.965 // 221 0.999 // 222 0.990 // 223 0.998 // 224 0.998 // 226 0.962 // 227 0.997 // 228 0.974 // 229 0.998 // 230 0.999 // 232 0.970 // 234 1.000 // 235 0.997 // 236 0.965 // 237 0.968 // 238 0.998 // 239 0.968 // 240 0.998 // 241 0.999 // 242 0.998 // 243 1.000 // 245 0.967 // 246 0.998 // 247 0.971 // 363 0.998
ENSG00000196914	ARHGEF12	5423,230396	5354,958158	136,544476	1,51678E-31	318 0.999 // 319 0.997 // 320 0.997 // 321 0.968 // 323 0.998 // 324 1.000 // 325 1.000 // 326 0.981 // 327 0.996
ENSG00000137343	C6orf134	1578,180653	1524,49823	107,364846	3,70317E-25	1 1.000 // 2 0.999 // 3 0.952 // 4 0.989 // 5 1.000 // 6 0.992 // 7 0.990 // 8 1.000 // 9 0.998 // 11 1.000 // 12 1.000 // 13 1.000 // 16 0.999 // 17 1.000 // 19 1.000 // 20 1.000 // 21 1.000 // 22 0.989 // 23 0.999 // 24 0.988
ENSG00000158717	RNF166	1000,875331	948,004426	105,74181	8,39963E-25	
ENSG00000077150	NFKB2	4333,278319	4290,320689	85,91526	1,87814E-20	448 0.999 // 449 0.998 // 454 1.000 // 455 0.998 // 456 1.000 // 457 0.999

ENSG00000067191	CACNB1	1802,381695	1776,106396	52,550598	4,19305E-13	156 1.000 // 157 1.000 // 158 0.998 // 160 0.985
ENSG00000124205	EDN3	1590,219581	1571,760073	36,919016	1,23139E-09	150 0.954 // 151 0.972 // 154 0.996
ENSG00000185585	OLFML2A	1315,707634	1297,403788	36,607692	1,44461E-09	1 0.981 // 2 0.976 // 3 0.989 // 4 0.981 // 5 0.998
ENSG00000136754	SSH3BP	2194,403382	2176,401644	36,003476	1,96966E-09	156 0.997
ENSG00000144535	DIS3L2	2511,62402	2496,805255	29,63753	5,20866E-08	
ENSG00000088367	EPB41L1	8576,194614	8562,097804	28,19362	1,09766E-07	
ENSG00000064042	LIMCH1	286,330074	273,798134	25,06388	5,54621E-07	
ENSG00000183914	DNAH2	7295,128296	7282,745477	24,765638	6,47414E-07	190 0.975 // 357 0.970 // 560 0.969 // 827 0.973 // 926 0.970 // 997 0.998 // 998 0.999
ENSG00000167566	KIAA1602	6315,666351	6304,409883	22,512936	2,08733E-06	
ENSG00000100208		1321,049479	1311,335605	19,427748	1,04478E-05	
ENSG00000138030	KHK	2519,552759	2510,063681	18,978156	1,32224E-05	98 0.950
ENSG00000148634	HERC4	4391,587521	4382,718624	17,737794	2,53539E-05	
ENSG00000023902	PLEKHO1	2615,651705	2606,911622	17,480166	2,90321E-05	
ENSG00000101464	CDC91L1	200,286462	192,001285	16,570354	4,68781E-05	
ENSG00000174948	GPR149	2555,973595	2548,258567	15,430056	8,56157E-05	250 0.996
ENSG00000172886	KRTAP5-7	487,155318	479,668954	14,972728	0,000109076	29 0.994 // 32 0.997 // 33 0.994 // 34 1.000 // 36 0.971 // 39 0.999 // 41 0.995 // 44 1.000 // 45 0.981 // 46 0.977 // 48 0.998 // 49 0.987

ENSG00000004139	SARM1	3996,442822	3990,004137	12,87737	0,000332579	174 0.998
ENSG00000147044	CASK	1148,804337	1142,886292	11,83609	0,000580937	1 0.993
ENSG00000154099	LRRC50	1945,19968	1939,520701	11,357958	0,000751254	77 0.977 // 173 0.979 // 254 0.997
ENSG00000188683	AC008749.6	3037,329918	3031,811712	11,036412	0,000893396	
ENSG00000146013	GFRA3	2973,517159	2968,230637	10,573044	0,001147486	304 0.959
ENSG00000171522	PTGER4	3321,125581	3316,029375	10,192412	0,001410197	235 0.995
ENSG00000119431	HDHD3	2111,664953	2106,586982	10,155942	0,001438369	28 0.964
ENSG00000146232	NFKBIE	1124,393182	1119,352354	10,081656	0,001497527	79 0.993
ENSG00000107902	LHPP	2006,564867	2001,651941	9,825852	0,00172076	78 0.993 // 94 0.984
ENSG00000132677	RHBG	3660,577023	3655,759278	9,63549	0,001908533	416 0.977
ENSG00000168878	SFTPB	1407,022324	1402,245817	9,553014	0,001996214	118 0.994 // 128 0.991
ENSG00000131187	F12	5368,557853	5363,836693	9,44232	0,002120351	
ENSG00000162227	TAF6L	3995,358244	3990,770253	9,175982	0,002452118	
ENSG00000122862	SRGN	1561,976308	1557,474795	9,003026	0,00269533	
ENSG00000118298	CA14	2584,598798	2580,23877	8,720056	0,003147284	153 0.954
ENSG00000033867	SLC4A7	5863,037996	5858,858119	8,359754	0,003836225	412 0.951
ENSG00000100889	PCK2	4338,63575	4334,590254	8,090992	0,004448581	
ENSG00000141577	AZI1	7169,354942	7165,383065	7,943754	0,004825363	237 0.978

ENSG00000001631	KRIT1	4331,021118	4327,129819	7,782598	0,005275188	
ENSG00000002726	ABP1	6270,433472	6266,779468	7,308008	0,006864799	
ENSG00000197122	SRC	1401,38894	1397,814167	7,149546	0,007498297	113 0.977
ENSG00000187872	OR2A14	2652,831462	2649,283845	7,095234	0,00772892	4 0.976 // 14 0.971 // 75 0.978 // 132 0.997 // 157 0.978 // 163 0.980 // 256 0.983
ENSG00000196228	SULT1C3	2736,04862	2732,506008	7,085224	0,007772211	72 0.975
ENSG00000116521	SCAMP3	2279,963309	2276,540344	6,84593	0,008884344	
ENSG00000165060	FXN	1287,490254	1284,067823	6,844862	0,008889657	
ENSG00000144283	PKP4	4006,403973	4002,984869	6,838208	0,008922831	48 0.977
ENSG00000167987	VPS37C	2746,991902	2743,588837	6,80613	0,009084545	70 0.981 // 197 0.981 // 248 0.980 // 275 0.992
ENSG00000189306	RRP7A	1267,687852	1264,366626	6,642452	0,009957679	98 0.980
ENSG00000148805	AL161645.14-201	1135,742833	1132,482552	6,520562	0,010663431	17 0.983
ENSG00000183066	WBP2NL	2783,550739	2780,322201	6,457076	0,011051136	
ENSG00000011028	MRC2	6977,925536	6974,697788	6,455496	0,011060967	158 0.974
ENSG00000135083	CCNJL	2014,644347	2011,470925	6,346844	0,011759032	
ENSG00000186174	BCL9L	6560,539923	6557,397849	6,284148	0,012182264	
ENSG00000183798	EMILIN3	5509,095943	5505,988513	6,21486	0,012668238	413 0.987 // 488 0.974
ENSG00000196118	AC106886.3-	2832,411688	2829,405222	6,012932	0,014201413	

	202					
ENSG00000180509	KCNE1	1091,628248	1088,649901	5,956694	0,014661509	37 0.958
ENSG00000163071	SPATA18	4481,855134	4478,949919	5,81043	0,015931399	
ENSG00000036672	USP2	3982,777027	3979,880628	5,792798	0,016091958	140 0.968
ENSG00000170891	CYTL1	1256,922587	1254,034221	5,776732	0,016239709	41 0.988
ENSG00000181296	OR5G1P	1731,830644	1728,969799	5,72169	0,016756595	4 0.989 // 43 0.996 // 48 0.995 // 54 0.979 // 108 0.996 // 183 0.987 // 193 0.987
ENSG00000181195	PENK	1425,100851	1422,277204	5,647294	0,017482331	
ENSG00000101057	MYBL2	4437,278029	4434,455463	5,645132	0,0175039	
ENSG00000115561	VPS24	1246,369878	1243,550614	5,638528	0,017569954	
ENSG00000141458	NPC1	9634,908128	9632,157716	5,500824	0,019007517	1007 0.990
ENSG00000072958	AP1M1	2359,599883	2356,906615	5,386536	0,020292718	
ENSG00000144381	HSPD1	3319,518969	3316,826036	5,385866	0,020300512	
ENSG00000100941	PNN	2812,431267	2809,79995	5,262634	0,021788023	
ENSG00000187475	HIST1H1T	1993,434458	1990,857325	5,154266	0,023189338	
ENSG00000173546	CSPG4	12816,99275	12814,41965	5,14619	0,023297441	
ENSG00000155657	TTN	6121,265456	6118,731828	5,067256	0,024381954	
ENSG00000182551	ADI1	850,644059	848,153635	4,980848	0,025629419	62 0.983

ENSG00000182348	ZNF804B	11358,57522	11356,1077	4,93505	0,026317303	501 0.974
ENSG00000086288	TXNDC3	4876,238348	4873,805919	4,864858	0,027409056	
ENSG00000159884	CCDC107	2176,13584	2173,724548	4,822584	0,028089208	
ENSG00000197591	OR11L1	2770,574576	2768,203924	4,741304	0,029446746	40 0.971 // 52 0.975 // 154 0.975 // 171 0.989
ENSG00000001617	SEMA3F	2731,259179	2728,889797	4,738764	0,029490254	
ENSG00000184640	SEPT9	183,001326	180,632876	4,7369	0,029522225	
ENSG00000174959		1091,390213	1089,206097	4,368232	0,03661497	
ENSG00000171497	PPID	2344,258017	2342,137882	4,24027	0,039475892	
ENSG00000173230	GOLGB1	8370,742702	8368,641746	4,201912	0,040378432	
ENSG00000166411	IDH3A	2233,127267	2231,094296	4,065942	0,043756279	
ENSG00000186714	CCDC73	5050,866454	5048,886783	3,959342	0,046611928	6 0.982 // 9 0.958 // 12 0.951 // 34 0.960 // 47 0.952 // 89 0.968 // 113 0.955 // 154 0.967 // 155 0.972 // 160 0.953 // 179 0.971 // 186 0.962 // 211 0.975 // 305 0.954 // 314 0.960 // 332 0.976 // 333 0.964 // 350 0.953
ENSG00000100376	FAM118A	2408,391211	2406,442281	3,89786	0,048347664	155 0.998
ENSG00000135521	LTV1	3633,327166	3631,381833	3,890666	0,048555174	

Appendix 2: Supplementary data to Chapter 1

Legends to Supplementary Figures

Figures S1 – S7: Distribution of low-frequency minor alleles. The proportion of SNPs with minor allele frequency (MAF) less than 0.10 within 100 kb sliding windows is plotted for each genomic region and population. The vertical gray rectangles delimit the physical coordinates of the gene of interest in each region. Solid colored circles represent the top 5% values of the distribution within each population across all regions. Open circles are values below the top 5% and hence are considered non-significant for that population. Gaps are the consequence of sliding windows having less than 5 SNPs, which has been set as the minimum for computing allele frequency proportions.

Figures S8 – S14: Distribution of high-frequency derived alleles. The proportion of SNPs with derived allele frequency (DAF) greater than 0.80 within 100 kb sliding windows is plotted for each genomic region and population. The vertical gray rectangles delimit the physical coordinates of the gene of interest in each region. Solid colored circles represent the top 5% values of the distribution within each population across all regions. Open circles are values below the top 5% and hence are considered non-significant for that population. Gaps are the consequence of sliding windows having less than 5 SNPs, which has been set as the minimum for computing allele frequency proportions.

Figure S15: Global F_{ST} for 39 populations across the eleven genomic regions analyzed. The vertical gray rectangles delimit the physical coordinates of the gene of interest in each region. Solid blue circles represent SNPs with F_{ST} values above the 95th percentile of the distribution across all regions. Open blue circles are SNPs below the 95^{th} percentile and hence are considered non-significant.

Figure S16: Distribution of REHH against frequency for populations within Middle East-North Africa. Core haplotypes within ±100 kb of the candidate genes (black dots) are plotted over the background distribution of cores from the eleven full 2 Mb regions (gray dots) analyzed. REHH is shown at a distance of about 0.42 cM for all populations included. Dashed lines indicate 0.95, 0.99 and 0.999 percentiles of REHH considering all cores. Cores that remained significant after multiple test correction ($q < 0.05$) are indicated with a black open diamond.

Figure S17: Distribution of –log p-values of XP-Rsb for each major geographic area across the eleven genomic regions analyzed. The genomewide significance of XP-Rsb at every SNP site for each population is plotted against physical distance over the candidate regions. Gray rectangles indicate the location of the gene of interest in each region. Dotted and dashed lines show 0.05 and 0.01 significance levels, respectively. Values above the latter are additionally represented with solid color circles, while open circles indicate values below the 0.01 significance level.

Figures S18 – S28: Distribution of populational –log p-values of XP-Rsb grouped across main geographical regions. The genome-wide significance of XP-Rsb at every SNP site and population is plotted against distance over each candidate region. Gray rectangles indicate the location of the gene of interest in each region. Dotted and dashed lines show 0.05 and 0.01 significance levels respectively. Values above the latter are additionally represented with solid color circles while open circles indicate values below the 0.01 significance level.

Figure S1:

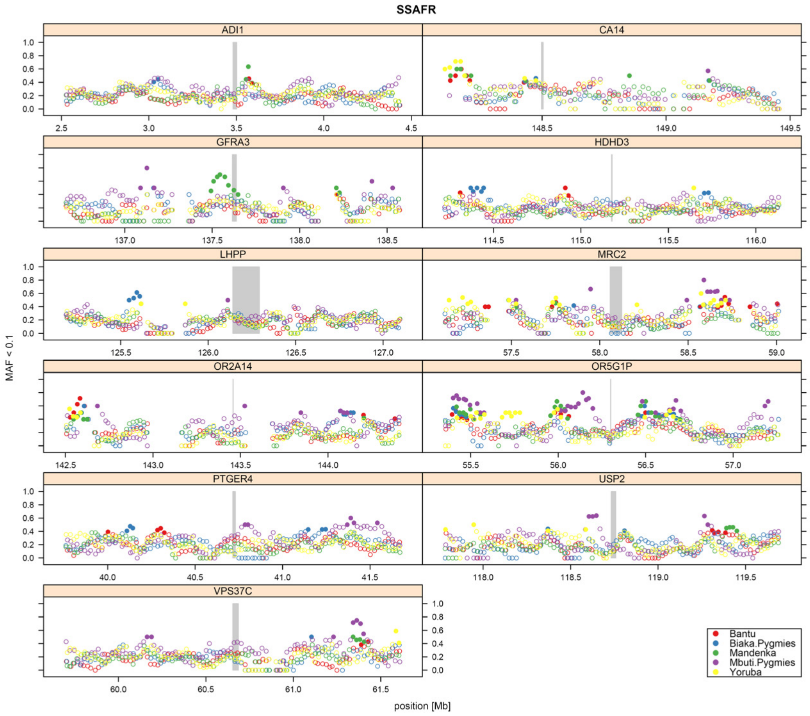


Figure S2:

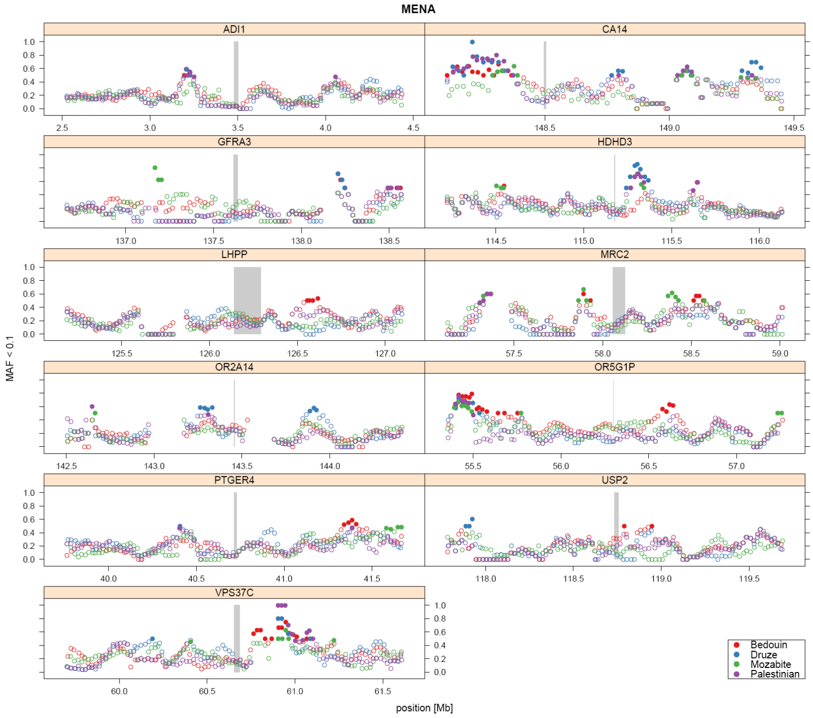


Figure S3:

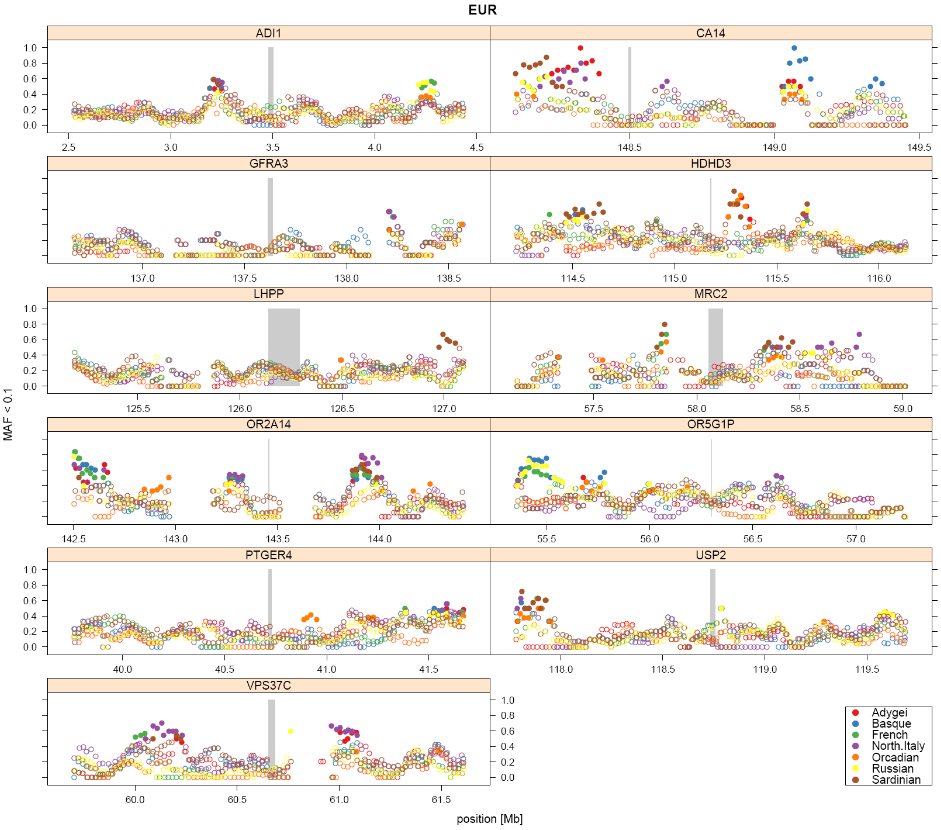


Figure S4:

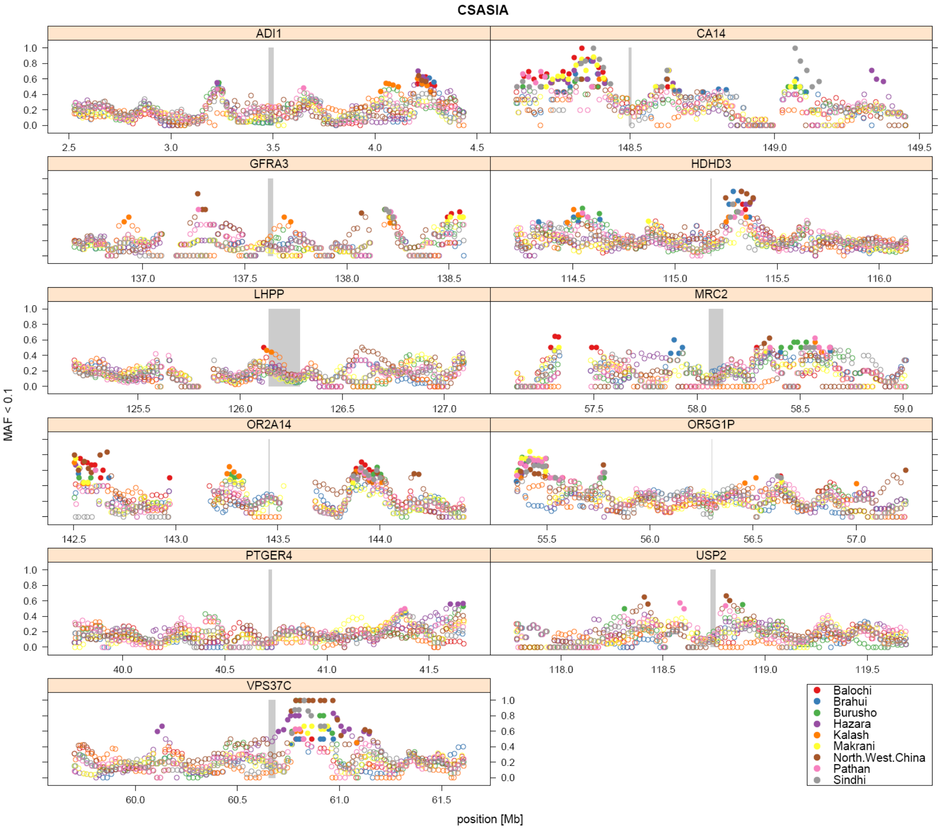


Figure S5:

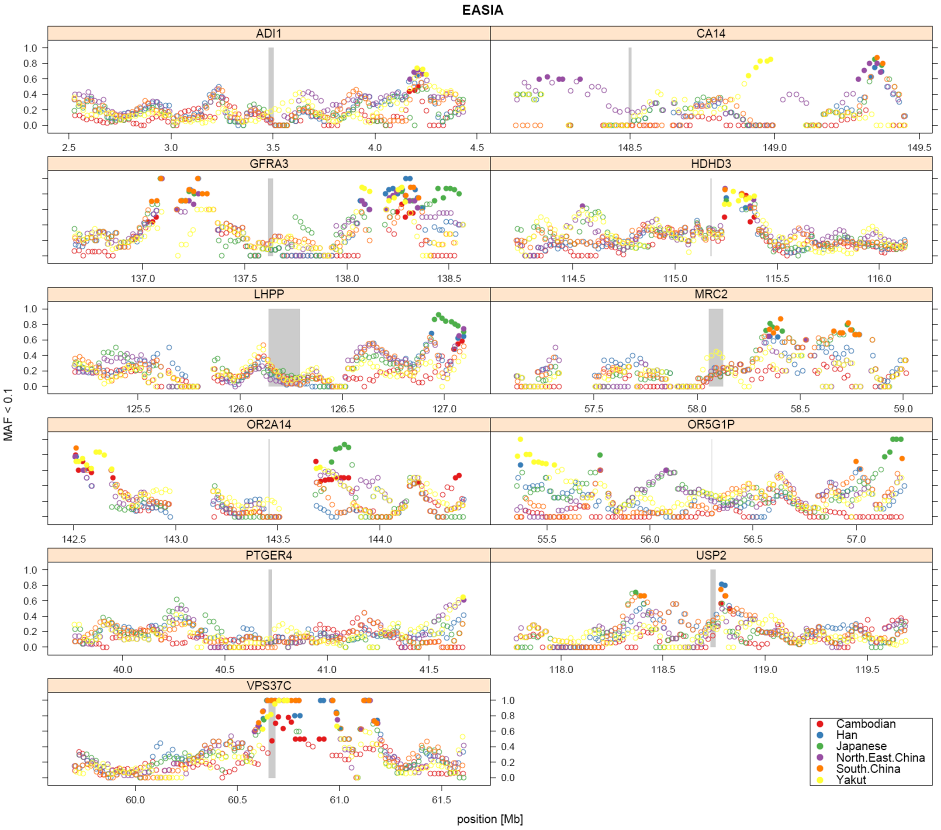


Figure S6:

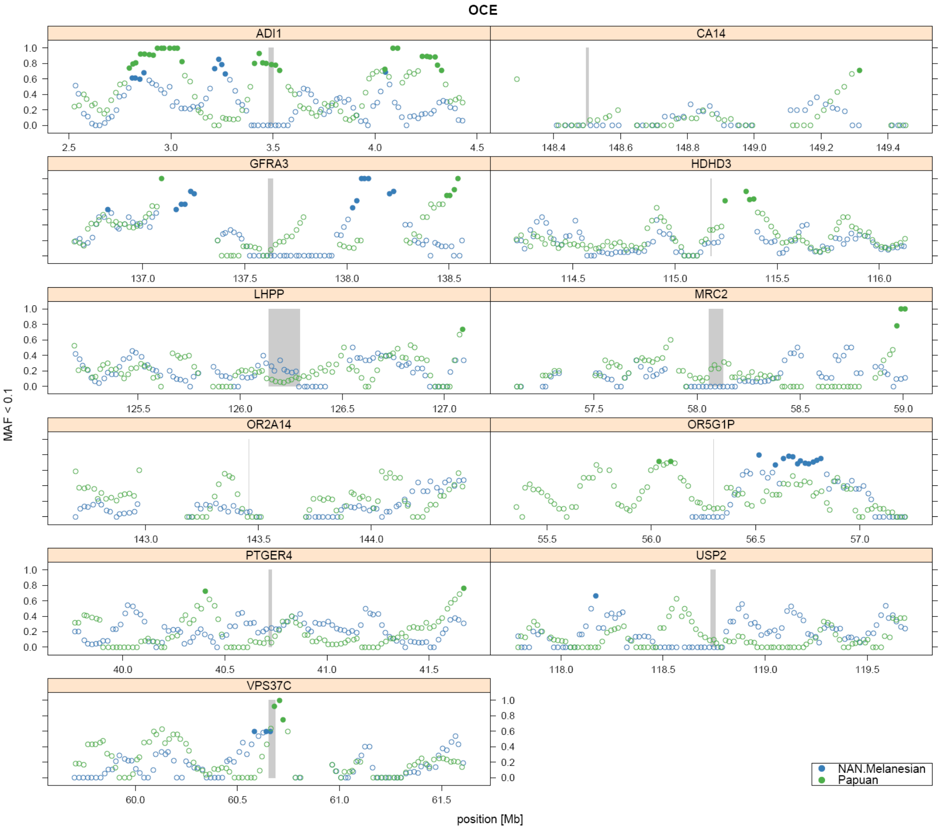


Figure S7:

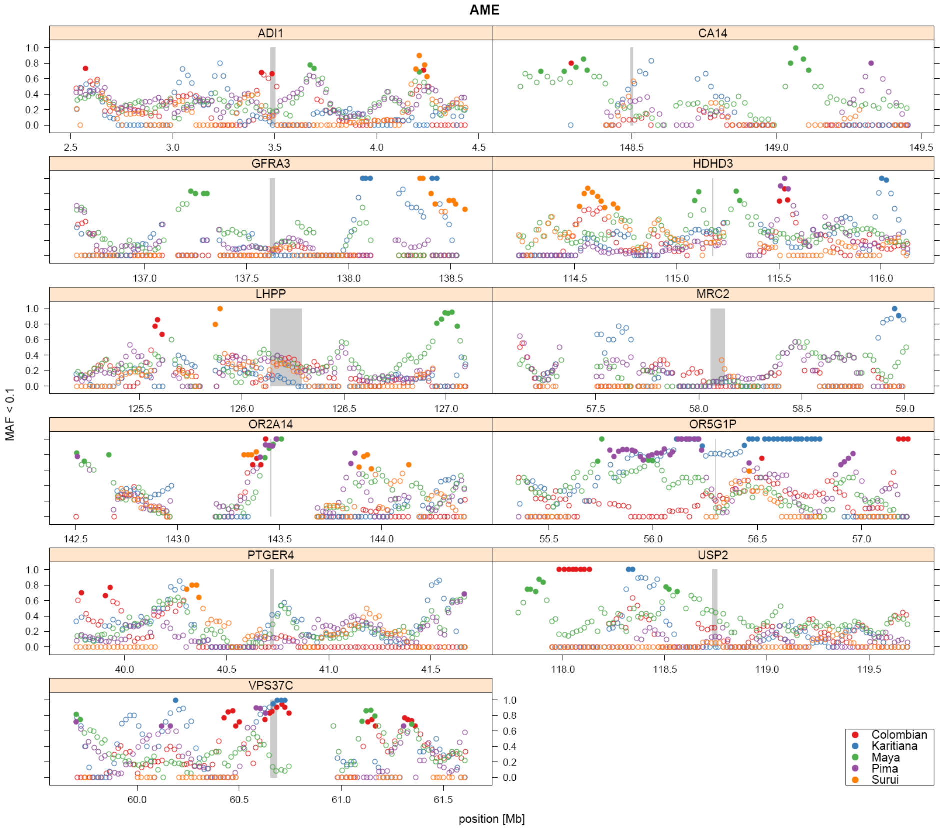


Figure S8:

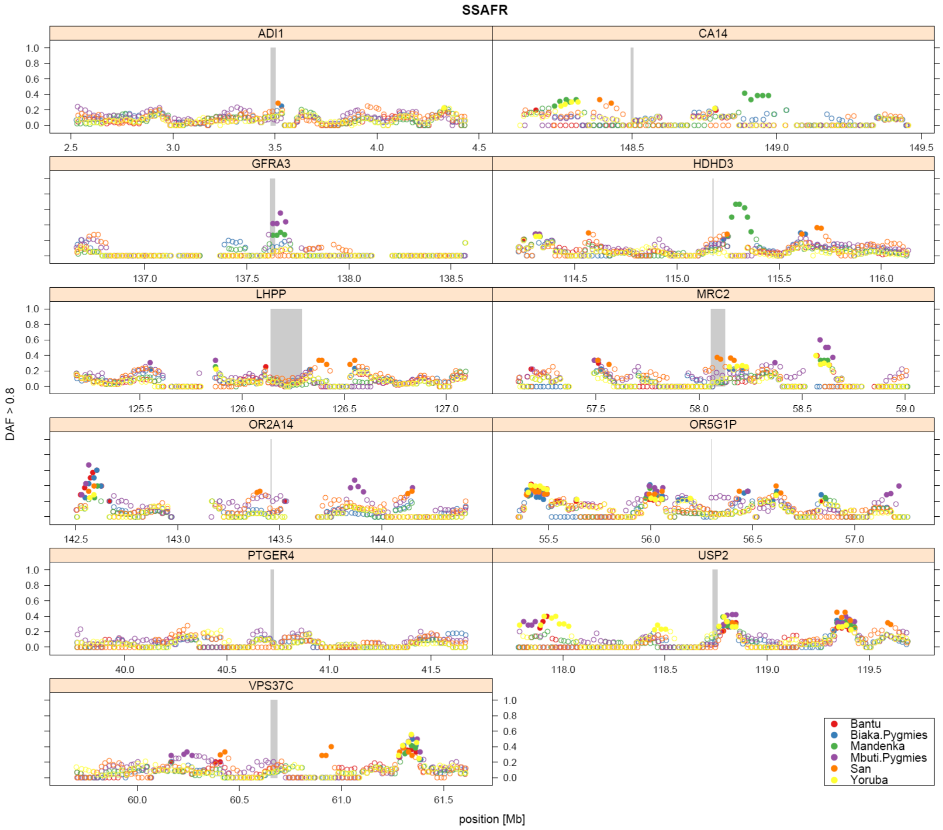


Figure S9:

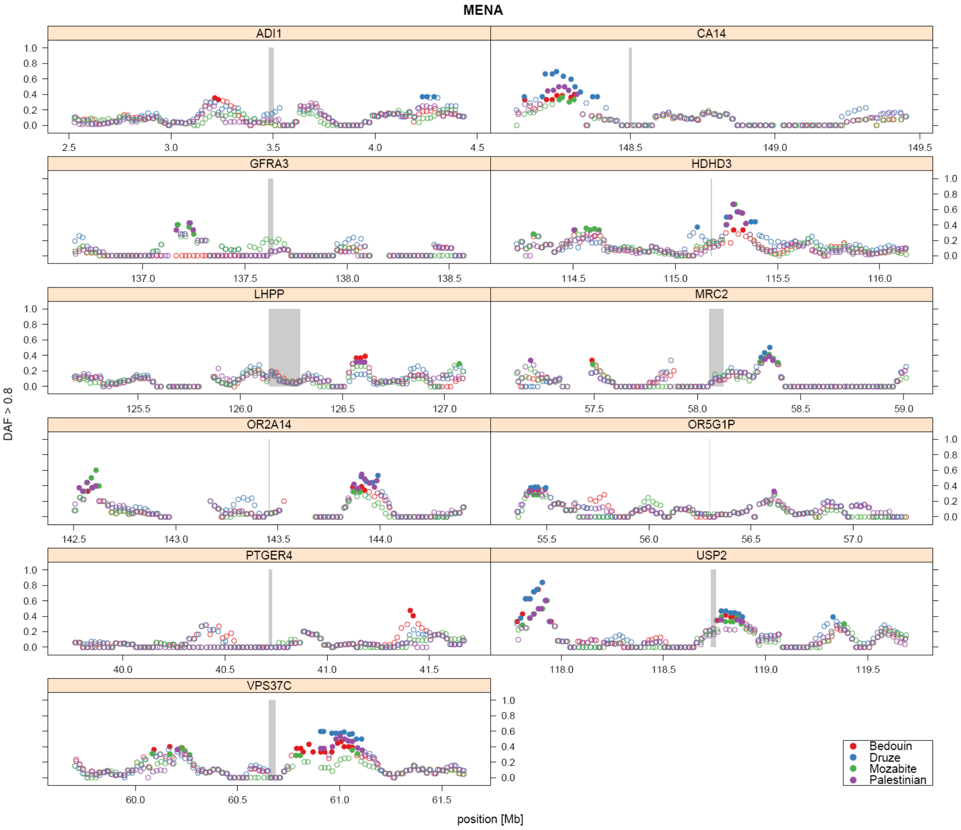


Figure S10:

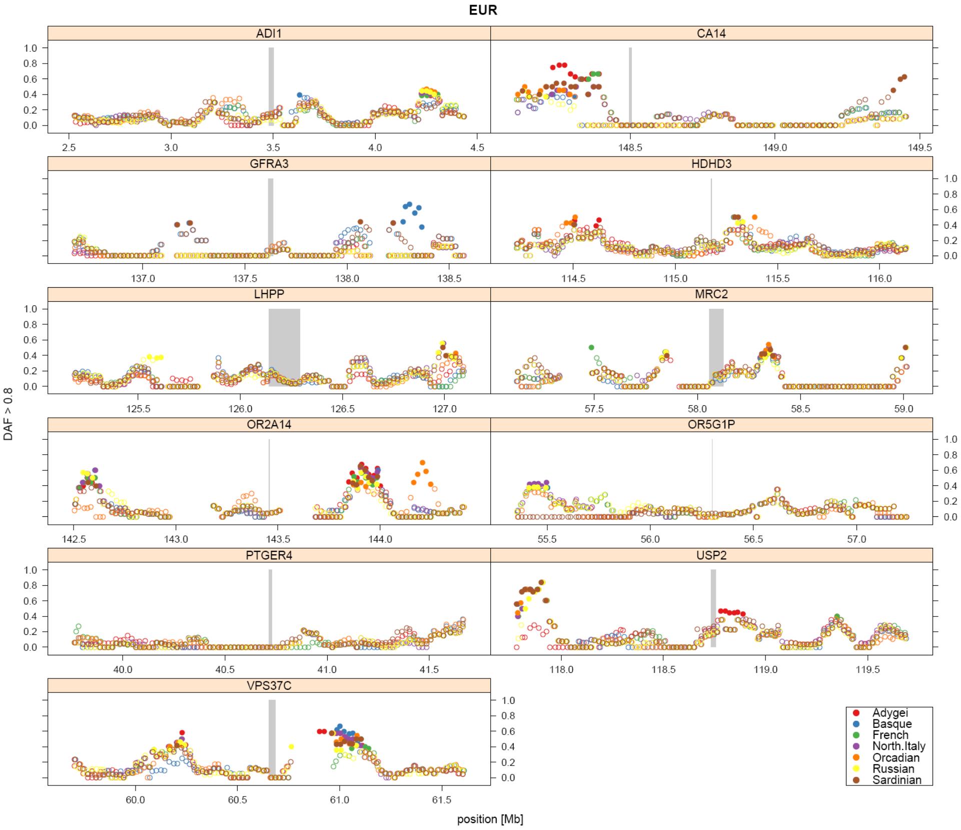


Figure S11:

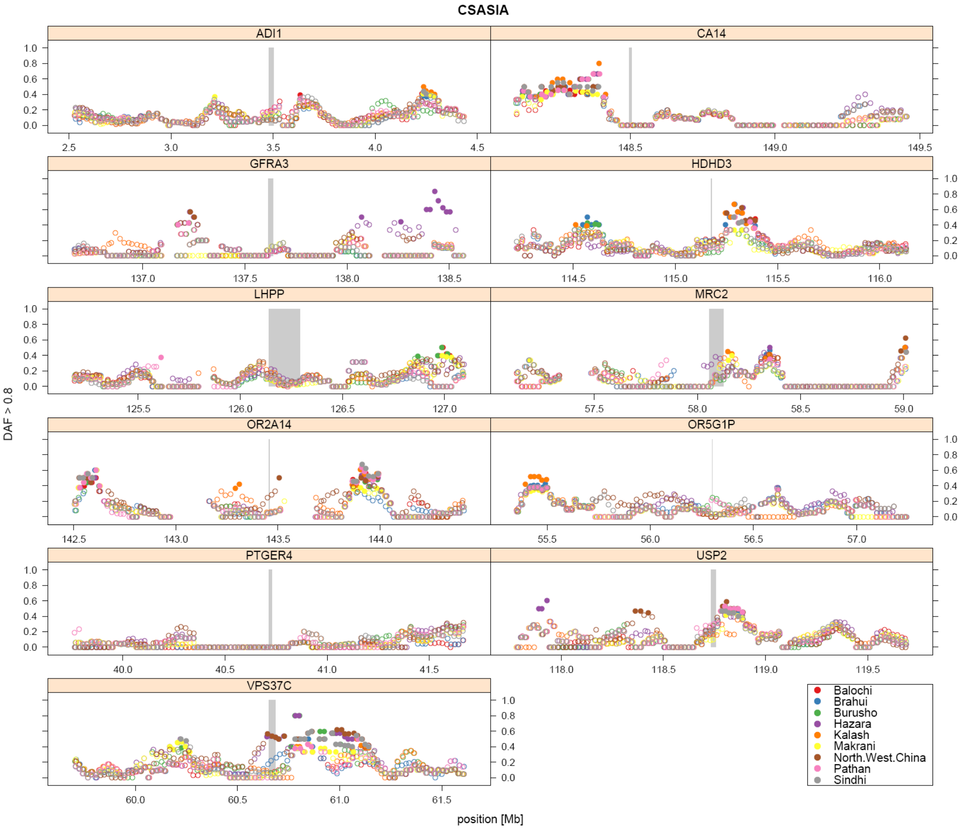


Figure S12:

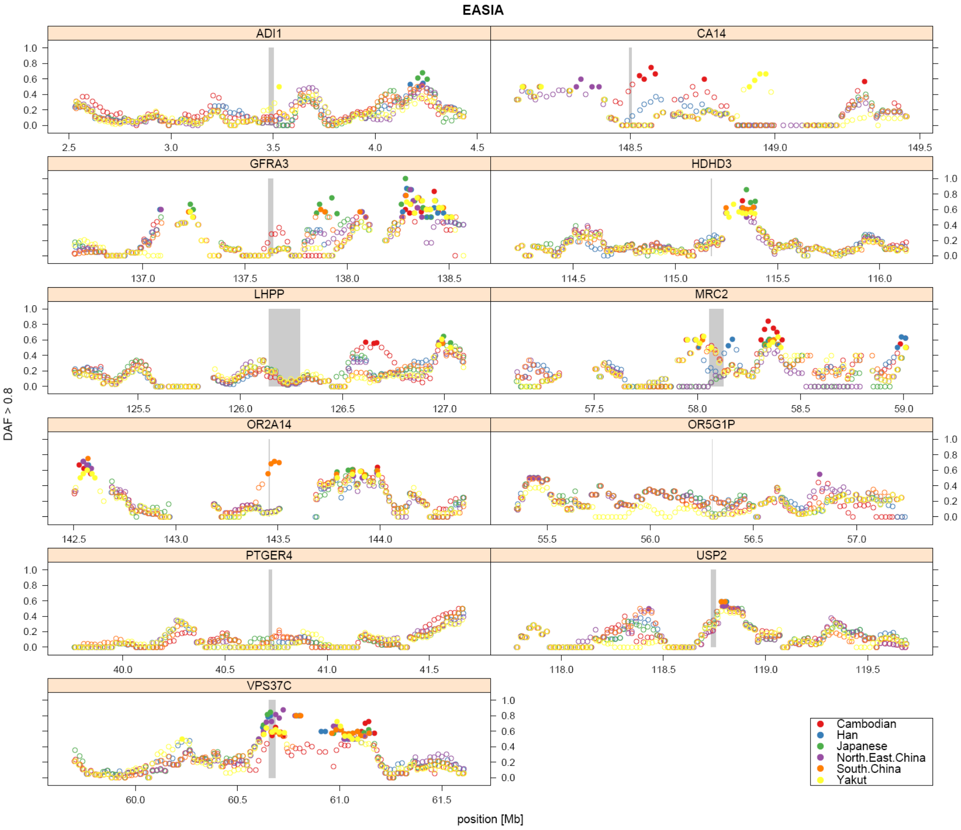


Figure S13:

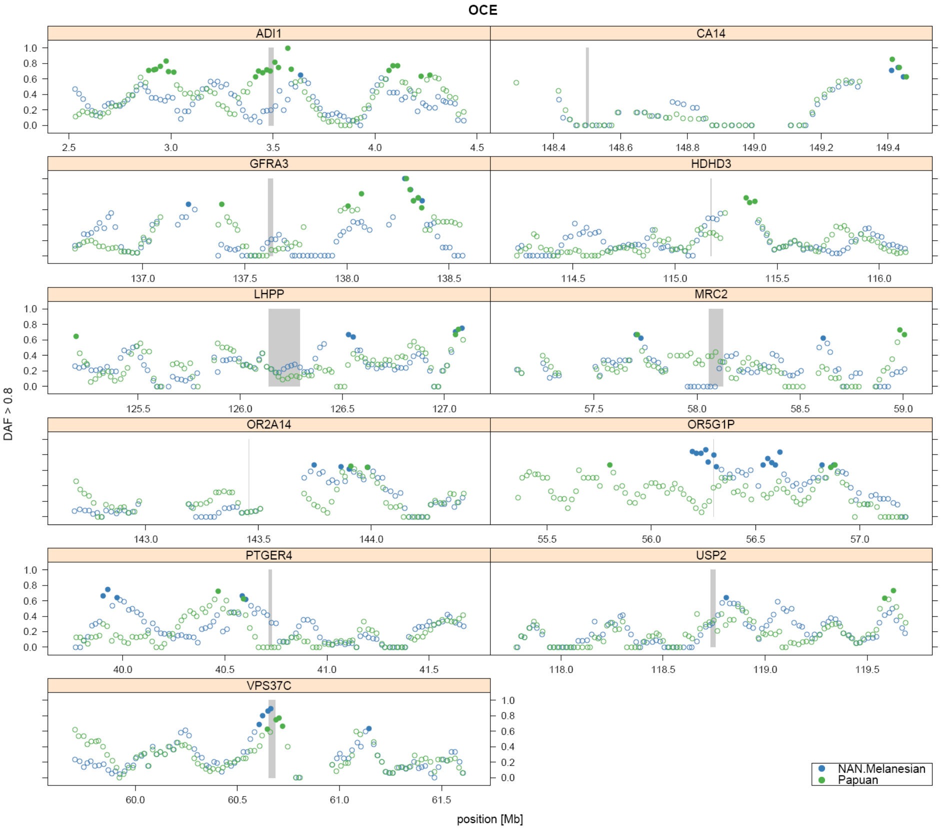


Figure S14:

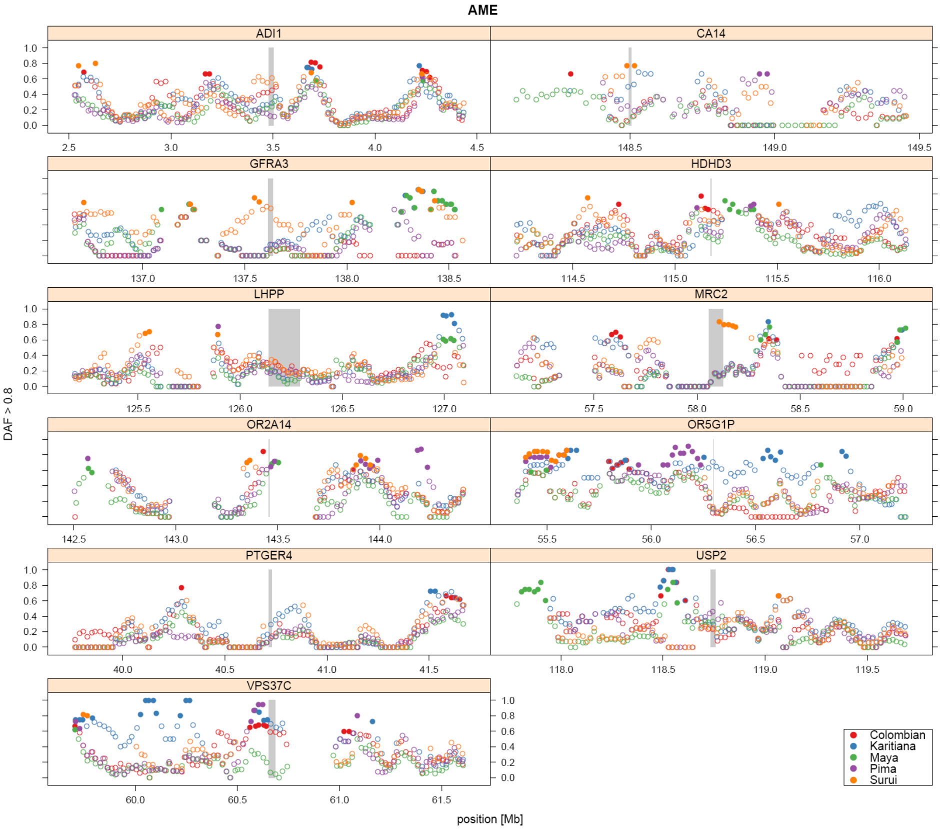


Figure S15:

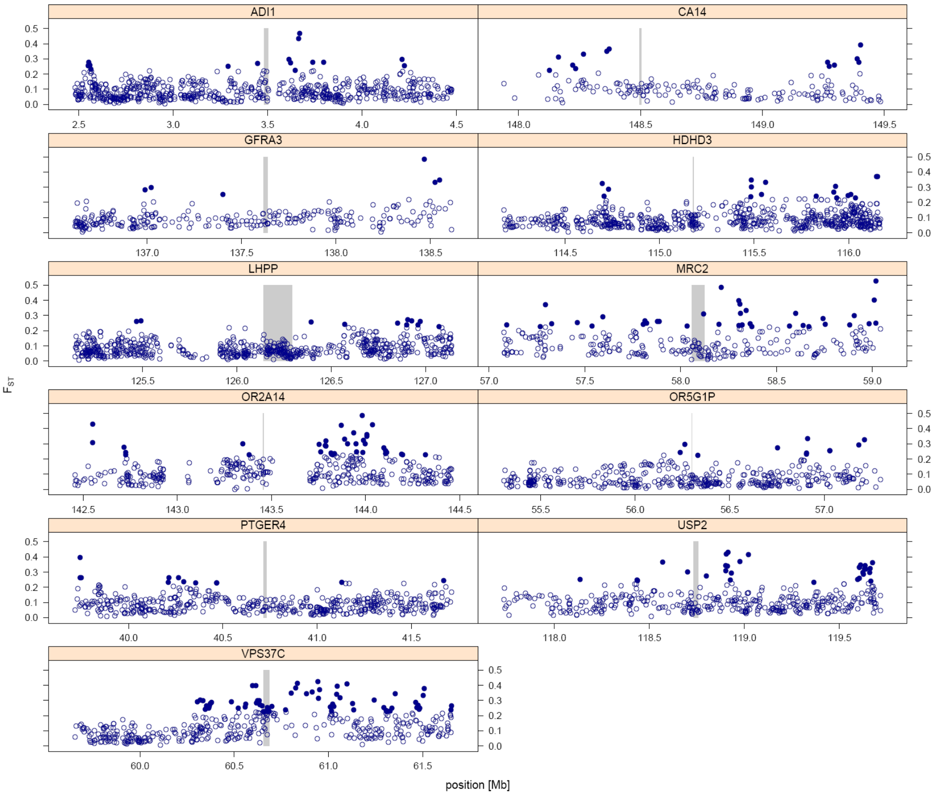


Figure S16:

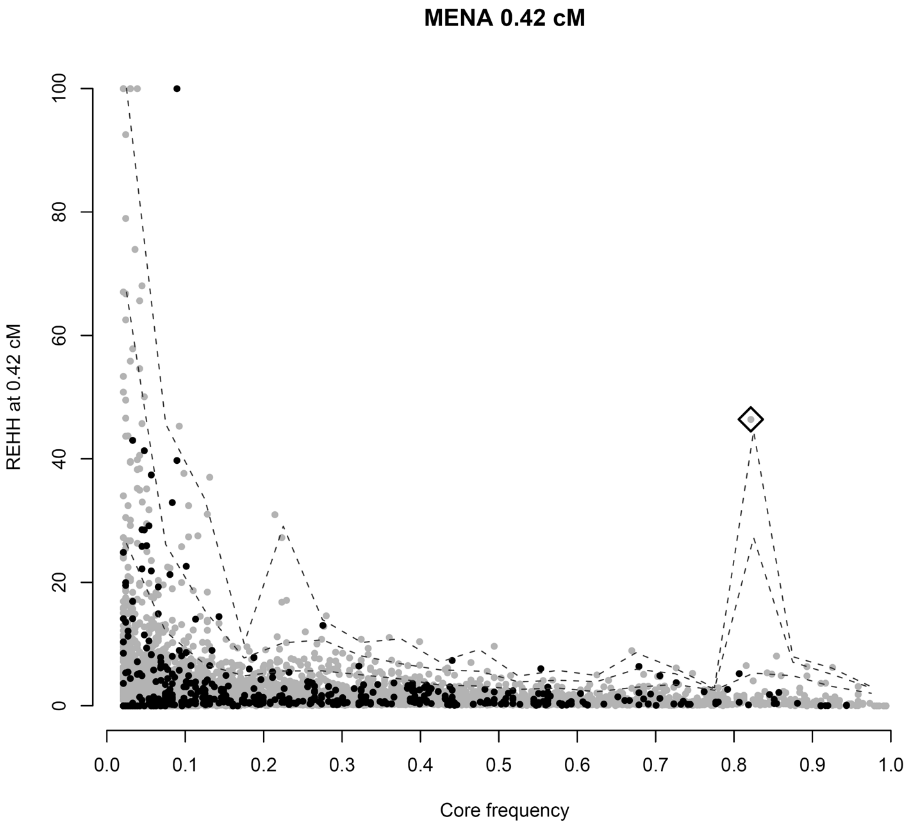


Figure S17:

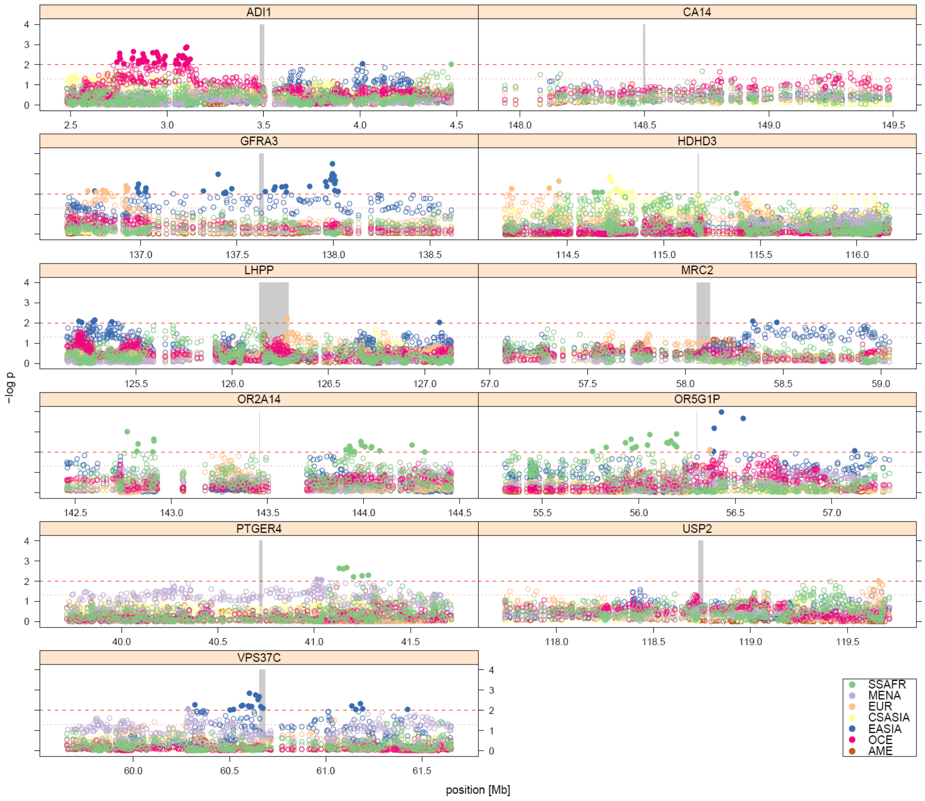


Figure S18:

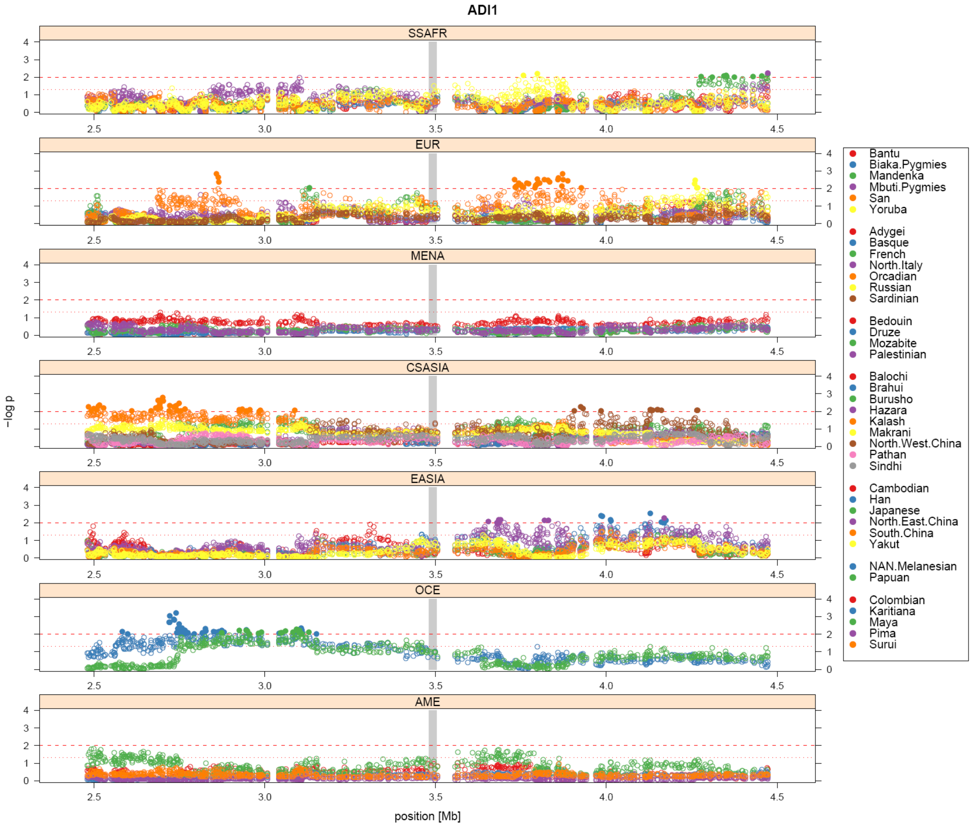


Figure S19:

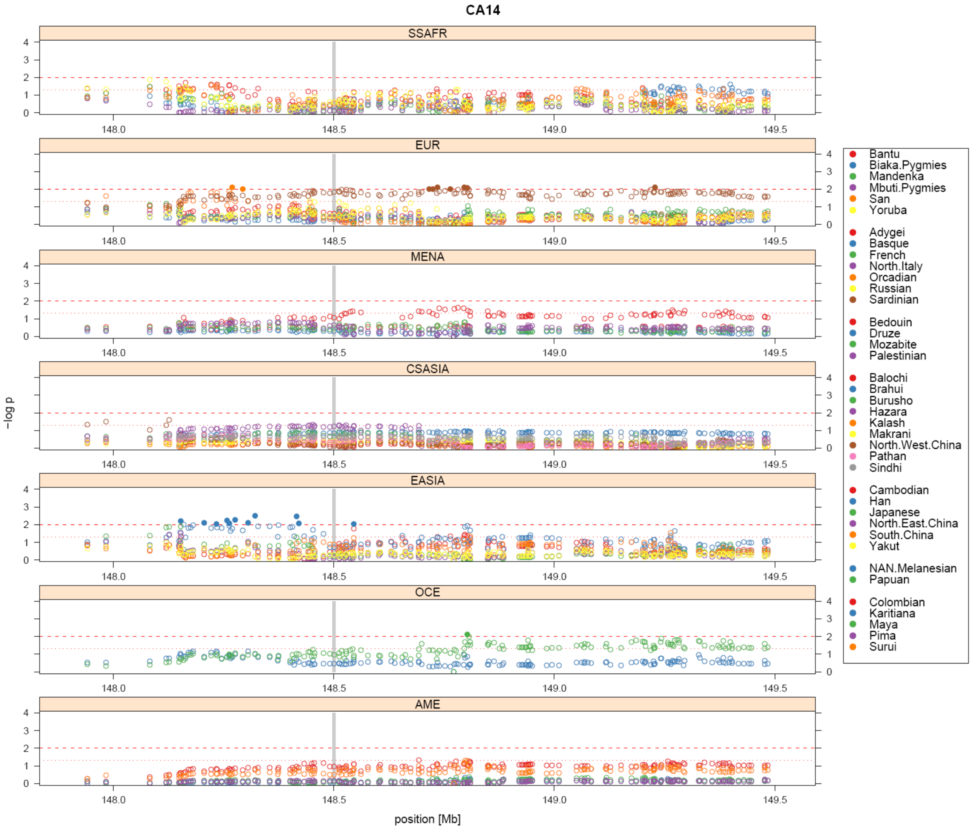


Figure S20:

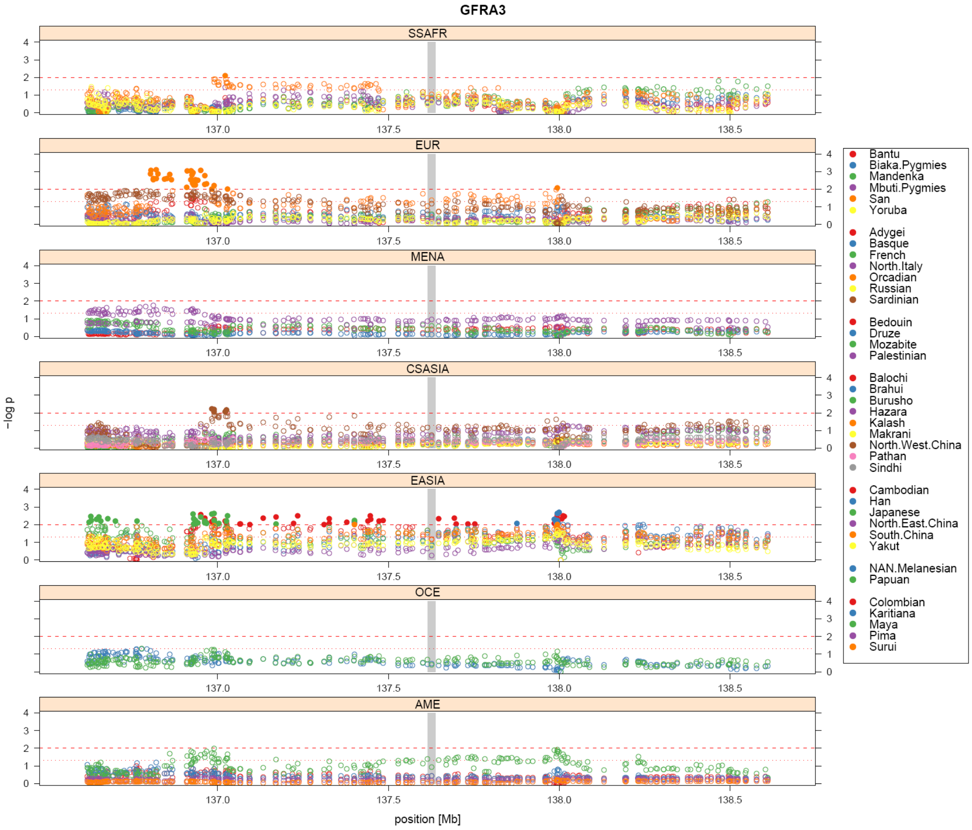


Figure S21:

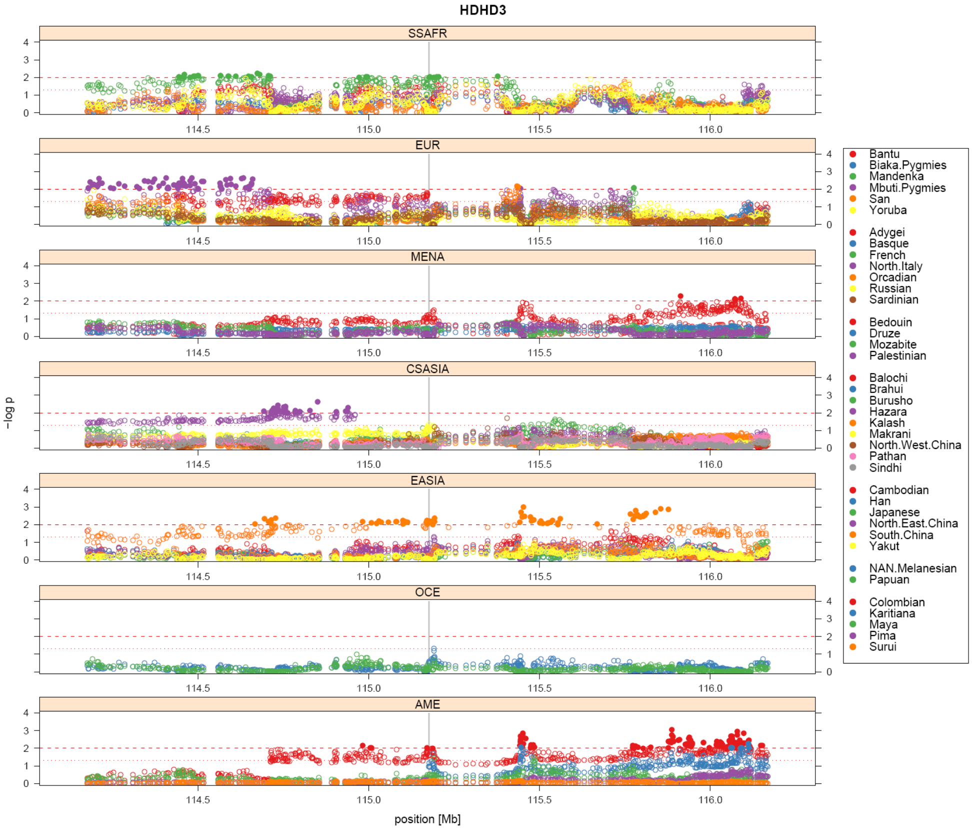


Figure S22:

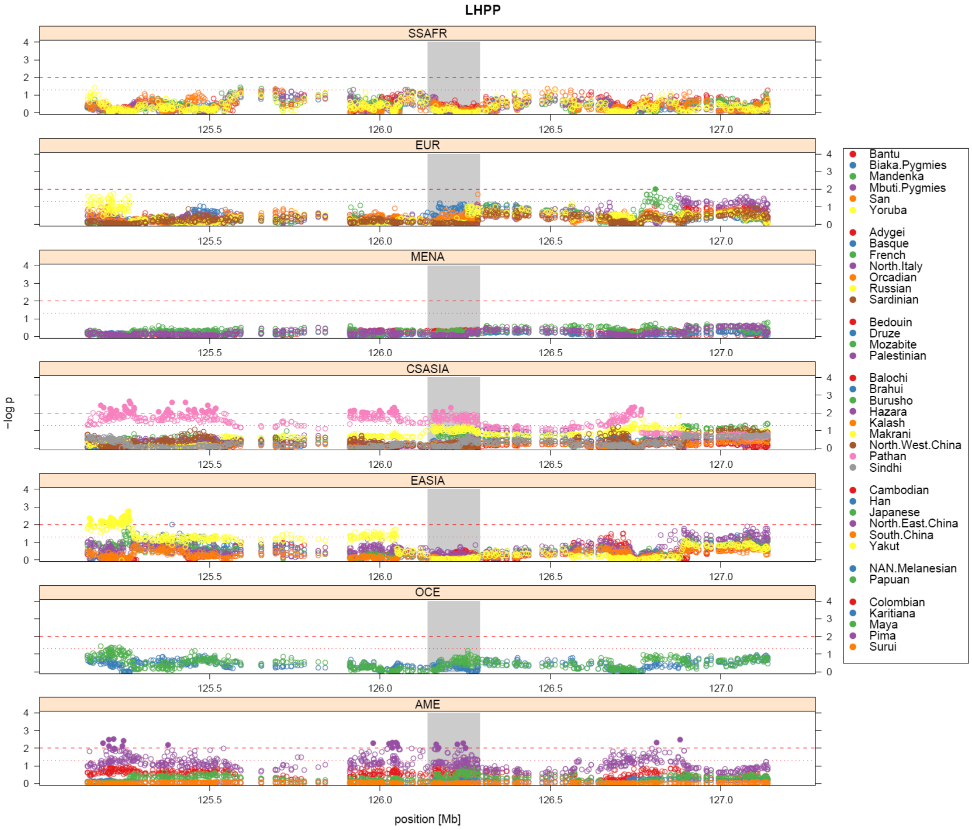


Figure S23:

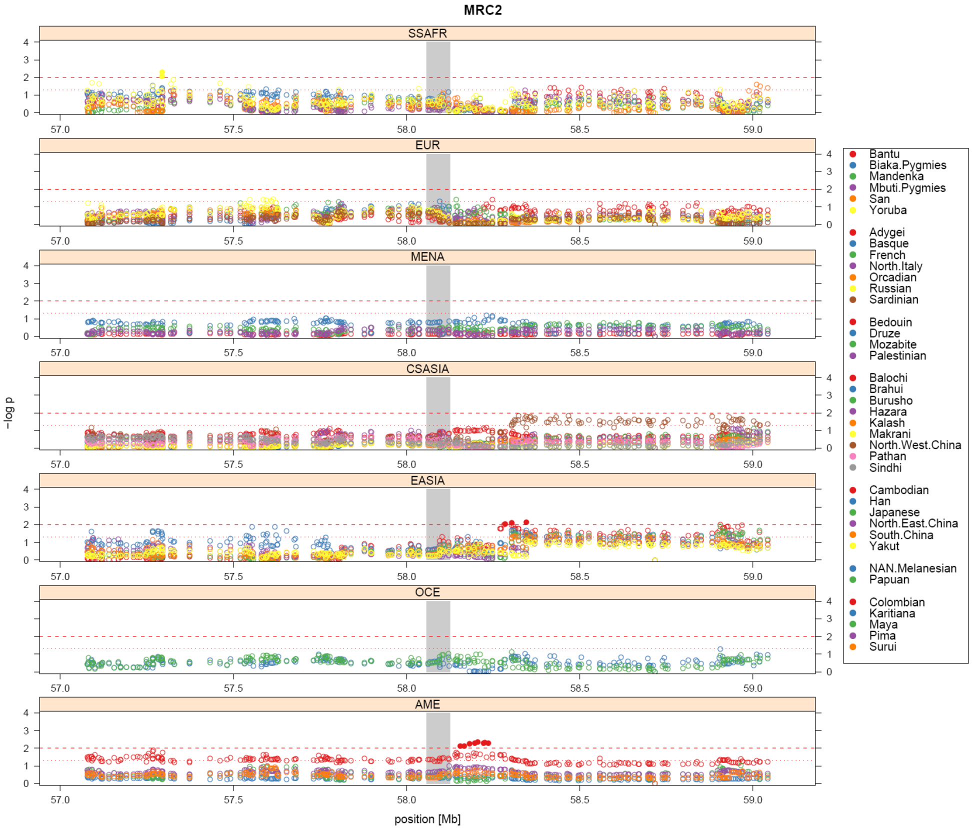


Figure S24:

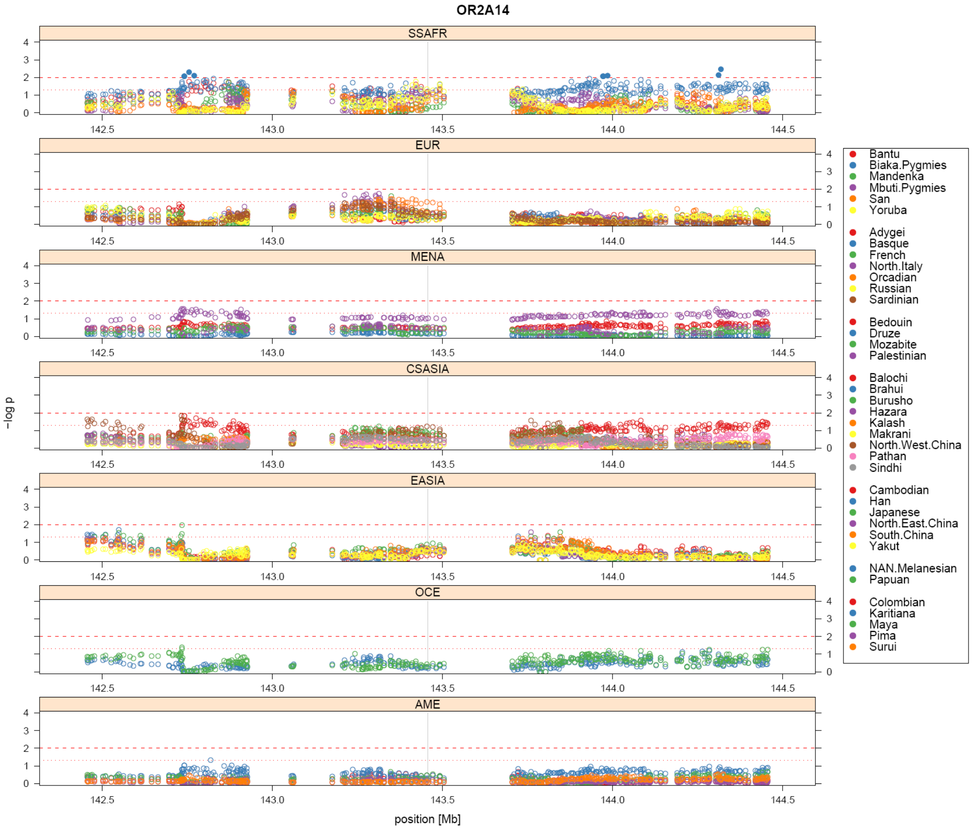


Figure S25:

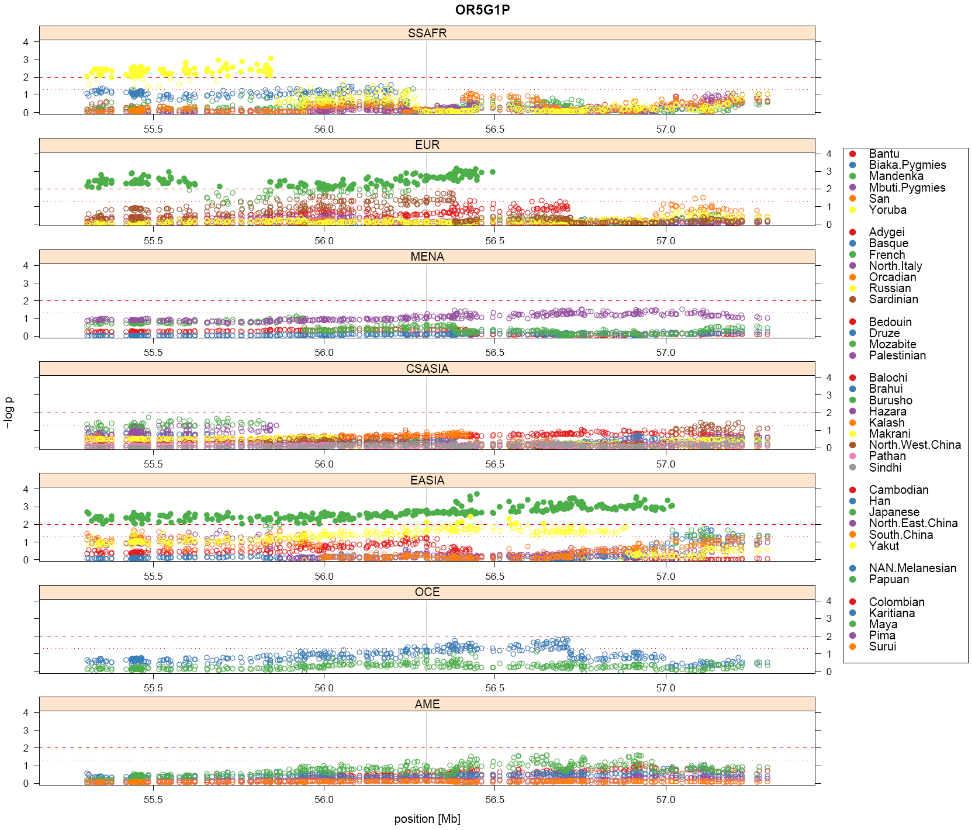


Figure S26:

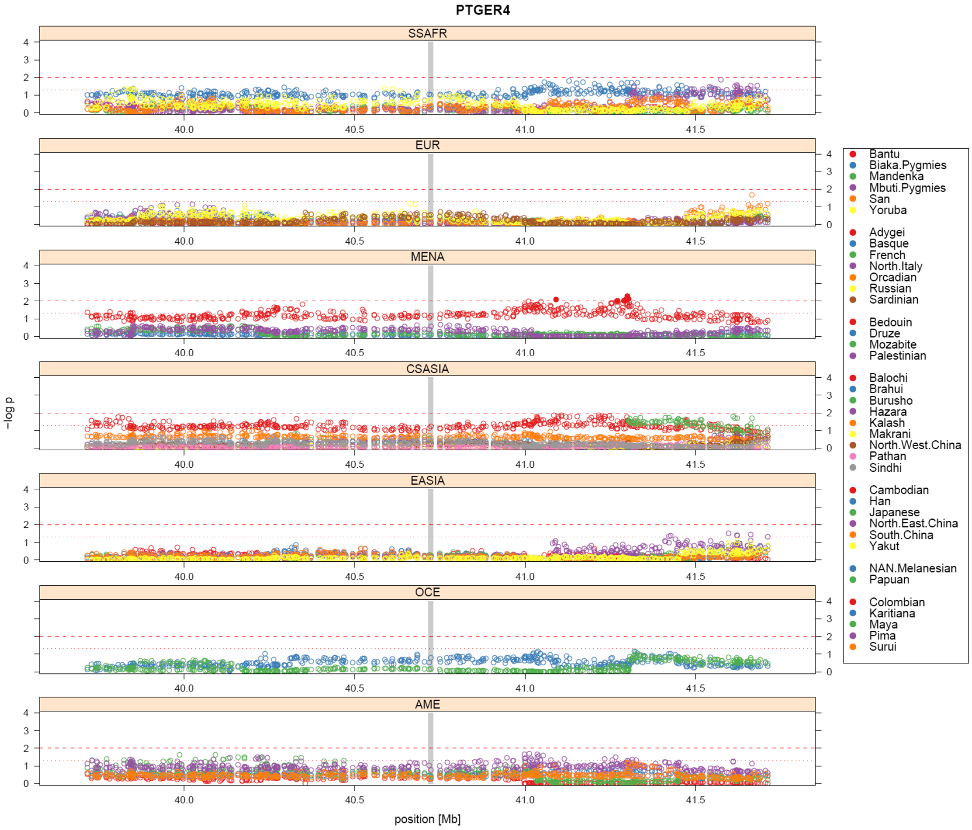


Figure S27:

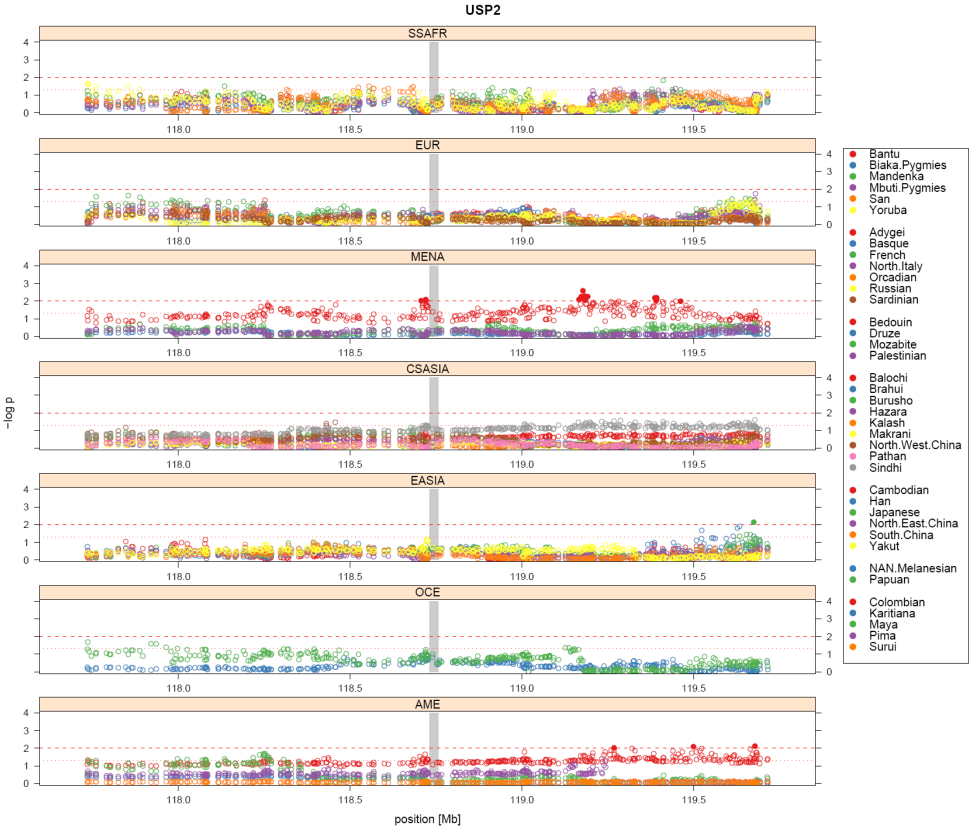
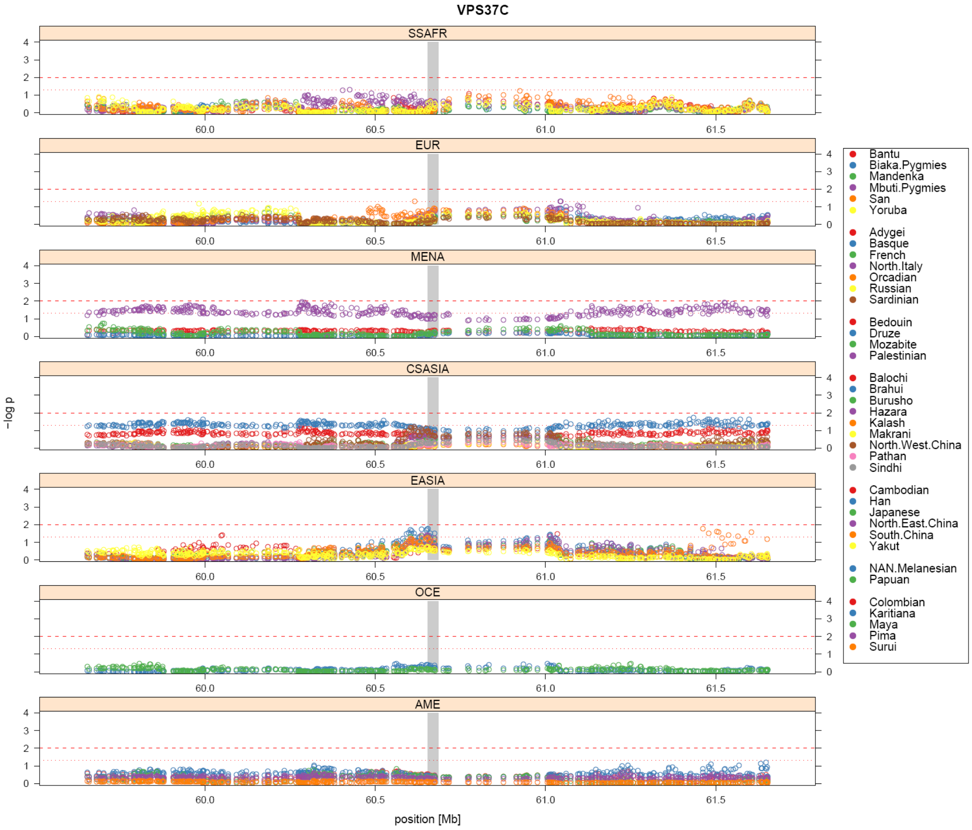


Figure S28



Appendix 3: Supplementary data to Chapter 2

Supplementary note S1

Signatures of selection

Non-polymorphic SNPs in the overall 39 populations were not considered in the threshold analyses of minor and derived allele frequencies. However, SNPs fixed in a population but polymorphic elsewhere were counted as having MAF < 0.10. The definition of minor allele was specific for each population rather than global and uniform across populations. As the ancestral allele of rs3923162 was ambiguous, we did not include this SNP in the DAF threshold analysis.

The following *OR5I1* ortholog sequences were also retrieved from Ensembl: ENSMMUG00000016153 (macaque), ENSMUSG00000068816 (mouse), ENSRNOG00000005793 (rat), ENSMODG00000020000 (opossum). Their multiple alignment with the chimpanzee and three different *OR5I1* human versions (ancestral-like, majoritary or human consensus sequence and majoritary derived) was done with RevTrans (Wernersson and Pedersen 2003). Phylogenetic analysis was conducted using MEGA3 (Kumar, Tamura, and Nei 2004) and a branch site likelihood method was used to test for positive selection (Zhang, Nielsen, and Yang 2005). In particular, we compared for the chimpanzee branch and for each human branch of the obtained phylogeny a null hypothesis with omega fixed to one *versus* an alternative hypothesis of omega being estimated at the branch of interest. Null and alternative hypotheses and the corresponding Bayesian empirical inference of amino acid sites under selection (Yang, Wong, and Nielsen 2005) were performed using the codeml program in the PAML package (Yang 1997).

Departures from neutrality were tested in resequencing data from Nielsen et al. (2005) by means of the Tajima's D, Fu and Li's F, F*, D and D* and Fay and Wu's H tests. In order to obtain realistic distributions for the statistics and thus evaluate evidence for natural selection, we performed 10,000 coalescent simulations using Cosi version 1.1 (Schaffner et al. 2005). As some demographic effects (such as population expansions) and positive selection have similar effects over genealogies (Charlesworth, Morgan, and Charlesworth 1993) those simulations include the ad-hoc human demographic calibration described in Schaffner et al. (2005) and provided with the Cosi source code. We have assumed an infinite-sites model, we have fixed S to 11, the number of segregating sites in the gene found by Nielsen et al. (2005), and the length of simulated sequence has been set to 945 bp (the *OR5I1* length). As for recombination, we have used the recombination rates found in the region (Myers et al. 2005). The critical value for each statistic has been obtained from the empirical distribution of the corresponding neutral model with a significance level of 0.05. We used DnaSP 4.0 (Rozas et al. 2003) to produce coalescent neutral simulations with a constant population size.

Yoruba, European and Chinese+Japanese phased haplotype data for those SNPs typed in all three populations and polymorphic in at least one of them was downloaded from the HapMap web page (Release 21 July 06). The extended haplotype homozygosity (EHH) and the relative extended haplotype homozygosity (REHH) for core haplotypes involving the *OR5I1* gene region were explored at 0.20, 0.25 and 0.30 cM for each HapMap population separately using the Sweep^{TM} software (http://www.broad.mit.edu/mpg/sweep/index.html) and considering the default core definition parameters.

Literature Cited

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

Kumar S, Tamura K, Nei M. 2004. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief Bioinform. 5:150-163.

Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. 2005. A fine-scale map of recombination rates and hotspots across the human genome. Science. 310:321-324.

Nielsen R, Bustamante C, Clark AG, et al. (12 co-authors). 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol. 3:e170.

Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics. 19:2496-2497.

Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. 2005. Calibrating a coalescent simulation of human genome sequence variation. Genome Res. 15:1576-1583.

Wernersson R, Pedersen AG. 2003. RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences. Nucleic Acids Res. 31:3537-3539.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 13:555-556.

Yang Z, Wong WS, Nielsen R. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. Mol Biol Evol. 22:1107-1118.

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol. 22:2472-2479.

Supplementary note S2

Recombination detection

As shown in figure 3, two main haplotype patterns differing at seven SNPs were found. This allowed inferring which haplotypes may be recombinants by noting changes in pattern along the sequence. A pattern shift had to be maintained for at least two consecutive SNPs for it to be considered the outcome of recombination rather than of a double recurrent mutation. Average recombination rate in the *OR5I1* genotyped region is extremely low (0.08 cM/Mb; Myers et al. (2005)) with a probability of recombination for the whole region of 5×10^{-5}. Also according to Myers at al. (2005) the *OR5I1* genotyped region contains no recombination hotspots. Moreover, three of the genotyped SNPs (namely rs7115131, rs4367963 and rs3923162) correspond to C/T polymorphisms at hypermutable CpG dinucleotides. We did observe recurrent substitutions at rs7115131 and rs3923162 but also in four non CpG dinucleotides positions.

Literature Cited

Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. 2005. A fine-scale map of recombination rates and hotspots across the human genome. Science. 310:321-324.

Supplementary Table S1. Relative frequencies of the *OR5I1* haplotypes across 39 worldwide populations. Haplotype definition as in figure 3

Population	2N[a]	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12	H13	H14	H15	H16
Sub-Saharan Africa																	
Bantu	36										0.417		0.333		0.083	0.028	0.139
Biaka Pyg	54	0.019						0.185			0.352		0.185		0.111	0.019	0.130
Mbuti Pyg	26	0.039						0.039			0.346		0.039		0.039	0.115	0.385
Mandenka	46							0.022			0.261		0.348		0.130	0.196	0.044
San	12				0.083						0.167	0.250	0.083			0.167	0.250
Yoruba	44							0.023			0.409		0.250		0.136		0.182
Middle East-North Africa																	
Mozabite	58		0.052				0.017				0.603		0.155		0.035		0.138
Palestinian	100		0.050					0.010			0.590	0.050	0.170		0.010		0.120
Bedouin	94		0.032			0.021					0.692		0.192		0.021		0.043
Druze	88		0.080								0.716		0.102	0.011			0.091
Europe																	
French	56		0.018								0.714		0.143				0.125
Basque	48		0.042								0.833		0.083				0.042
Orcadian	28		0.179								0.607		0.143				0.071
Sardinian	56		0.214								0.607		0.089	0.018			0.071
N Italian	42		0.119								0.833		0.048				
Adygei	34		0.118						0.029		0.735		0.088				0.029
Russian	50		0.020								0.820		0.100				0.060

Continued (Supplementary Table S1)

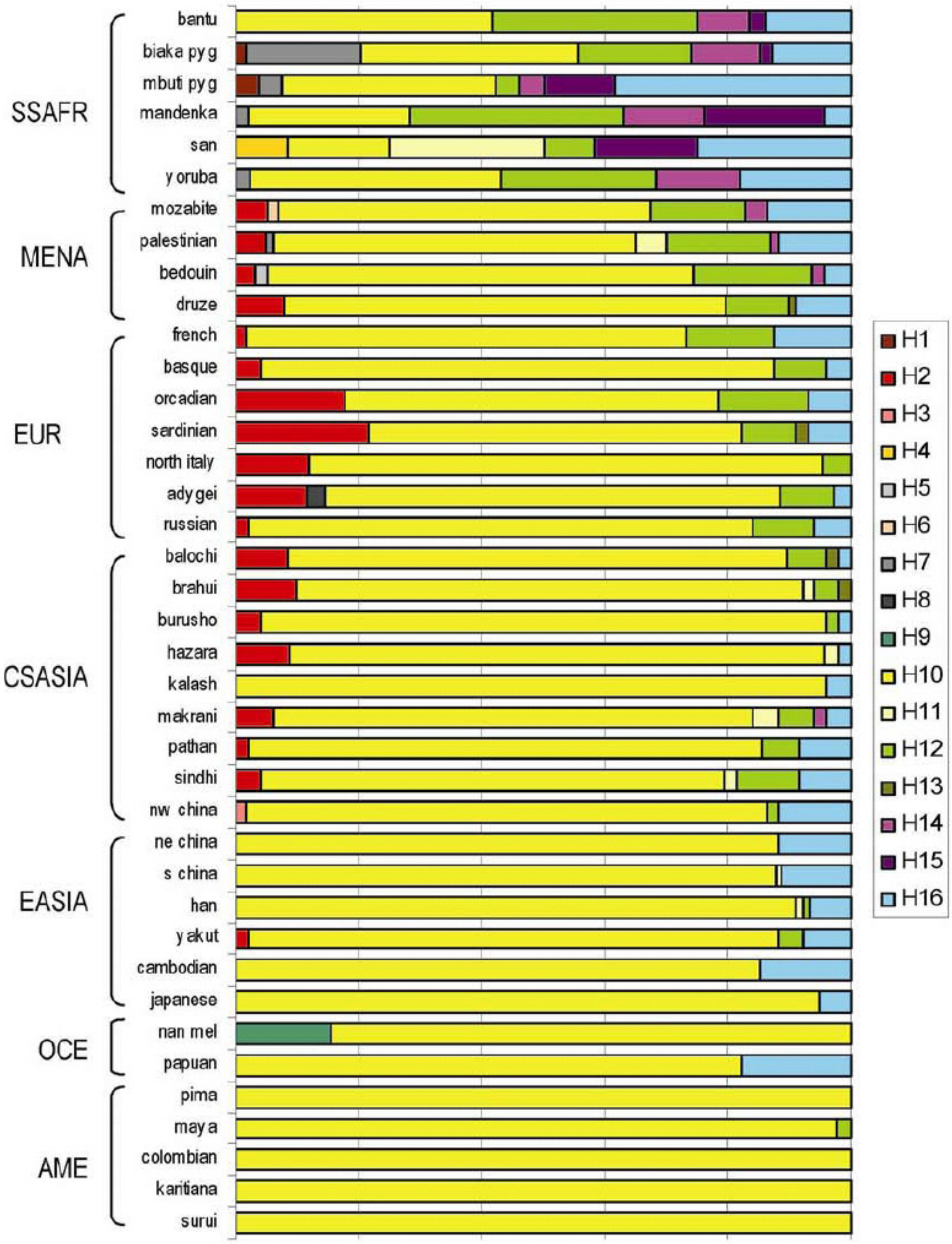
Population	2N [a]	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12	H13	H14	H15	H16
Central-South Asia																	
Balochi	48		0.083								0.813		0.063	0.021			0.021
Brahui	50		0.100								0.820	0.020	0.040	0.020			
Burusho	50		0.040								0.920		0.020				0.020
Hazara	46		0.087								0.870	0.022					0.022
Kalash	48										0.958						0.042
Makrani	50		0.060								0.780	0.040	0.060		0.020		0.040
Pathan	48		0.021								0.833		0.063				0.083
Sindhi	48		0.042								0.750	0.021	0.104				0.083
NW China	58			0.017							0.845		0.017				0.121
East Asia																	
NE China	76										0.882						0.118
S China	132										0.879	0.008					0.114
Han	88										0.909	0.011	0.011				0.068
Yakut	50		0.020								0.860		0.040				0.080
Cambodian	20										0.850						0.150
Japanese	58										0.948						0.052
Oceania																	
Nasioi	26									0.154	0.846						
Papuan	34										0.824						0.177
America																	
Pima	28										1.000						
Maya	44										0.977		0.023				
Colombian	14										1.000						
Karitiana	28										1.000						
Surui	18										1.000						

[a] Number of chromosomes

Supplementary Table S2. Nonsynonymous variation at the *OR5I1* gene.

SNP ID	Codon position	Ancestral codon	Derived codon	Grantham distance	PolyPhen prediction
rs17597625	6	Gly (G)	Arg (R)	56	Benign
rs4367963	50	Leu (L)	Ser (S)	145	Probably damaging
rs966086	76	Ser (S)	Phe (F)	155	Possibly damaging
rs9665861	306	Val (V)	Ile (I)	29	Benign

Supplementary Figure S1



Appendix 4: Supplementary data to Chapter 3

Supplementary Table 1. Significance of the likelihood ratio tests of positive selection performed on the human lineage for the *FOXI1* gene

Compared organisms	Branch site test	Strict branch + site test
Hsa, Ppa, Ptr, Ggo, Ppy, Hkl, Mmu, Pha, Sla, Rno and Mms	0.18714	0.80650
Hsa, Ppa, Ptr, Mmu, Pha, Sla, Rno and Mms	0.59670	0.62421
Hsa, Ppa, Ptr, Ggo, Ppy, Hkl, Mmu, Pha, Rno and Mms	0.18714	0.08065
Hsa, Ppa, Ptr, Ggo, Ppy, Hkl, Mmu, Pha, Sla	0.15939	1.00000
Hsa, Ppa, Ptr, Ggo, Ppy, Hkl, Sla, Rno, Mms	0.18463	0.8065
Hsa, Ptr, Ppy, Mmu, Rno, Mms, Ocu, Cfa, Bta, Dno, Laf, Mdo	0.59670	0.84148

In all cases p = Chi dist (2X(lnL1-lnL2), (np1-np2)). The three letter species codes are (% coverage): Hsa (human, 100%), Ppa (bonobo, 81.31%), Ptr (chimp, 100%), Ggo (gorilla, 78.57%), Ppy (orangutan, 100%), Hkl (gibbon, 79.63%), Mmu (macaque, 100%), Pha (babbon, 15.34%), Sla (Tamarin, 75.13%), Rno (rat, 100%), Mms (mouse, 100%), Ocu (rabbit, 100%), Cfa (dog, 100%), Bta (cow, 100%), Dno (armadillo, 100%), Laf (elephant, 100%) and Mdo (opossum, 100%)

Supplementary Table 2. Functional characterization and allele frequencies for functional relevant SNPS within ~ 140 Kb containing *FOXI1*

SNP ID	Position[a]	Alleles[b]	Africans	Europeans	Asians	Sources	Functional Effect[d]
rs6555882	169401666	G/**C**	0.683	1	1	HapMap-YRI, CEU, CHB, JPT	ESE (*DOCK2*)
rs17647491	169416284	**C**/T	1	0.966	1	HapMap-YRI, CEU, CHB, JPT	ESE (*DOCK2*)
rs13179480	169428747	**A**/C	1	1	0.988	HapMap-YRI, CEU, JPT	CNS, ESE (*DOCK2*)
rs13179490	169428777	**A**/C	1	1	1	HapMap-YRI, CEU, CHB, JPT	CNS (*DOCK2*)
rs2270900	169435654	**T**/C	1	1	0.989	HapMap-YRI, CEU, CHB, JPT	ESE (*DOCK2*)
rs1045168	169437321	**T**/C	0.619	0.758	0.895	HGDP-Yoruba, HapMap- CEU, JPT	ESE (*DOCK2*)
rs2270898	169441471	**T**/A	1	1	0.989	HapMap-YRI, CEU, CHB, JPT	CNS, ESE (*DOCK2*)
rs1045176	169442598	G/**T**	0.595	0.121	0.478	HapMap-YRI, CEU, CHB	ESE (*DOCK2*)
rs9307	169442601	A/**G**	NA[c]	NA[c]	0.460	CEPH individuals (dbSNP)	ESE (*DOCK2*)
rs17072089	169461251	**C**/G	0.978	0.979	1	Afr-Am, European, Asian (Perlegen)	New TFBS (*FOXI1*)
rs7704953	169461667	C/**T**	0.929	NA[c]	0.818	HGDP-Yoruba, HapMap- JPT	New TFBS (*FOXI1*)
rs34218925	169465509	G/T	NA[c]	NA[c]	NA[c]		ESE (*FOXI1*)
rs2277944	169465818	A/**G**	0.714	0.188	0.479	HGDP-Yor, European, Asian (Perlegen)	ESE (*FOXI1*)
rs35678180	169467782	G/A	0.974	NA[c]	NA[c]	Afr-Am (Applera)	ESE (*FOXI1*)
rs10063424	169468100	T/**C**	0.024	0.092	0.044	HGDP-Yoruba, HapMap- CEU, CHB	ESE (*FOXI1*)
rs3828625	169468141	**T**/C	1	1	0.988	HGDP-Yoruba, HapMap- CEU, JPT	CNS, ESE (*FOXI1*)
rs6873124	169468312	A/**C**	0.575	0.164	0.489	HapMap-YRI, CEU, CHB	ESE (*FOXI1*)
rs6555887	169468633	**A**/G	0.847	0.908	0.978	HapMap-YRI, CEU, JPT	ESS, ESE, miRNA (*FOXI1*)
rs6555888	169468728	G/**A**	0.842	0.808	0.944	HapMap-YRI, CEU, JPT	miRNA (*FOXI1*)

[a] SNP positions are based on NCBI build 36.3, [b] Ancestral allele in bold, [c] Not available, [d] Functional effect as predicted by PupaSuite (Conde et al. 2006; Reumers et al. 2008): ESE, exonic splicing enhancer;CNS, coding non-synonymous SNPs; TFBS, transcription factor binding sites ; ESS, exonic splicing silencer; miRNA, microRNAs and their targets.

Legends to Supplementary Figures

Figure S1. Median Joining Network of human *FOXI* haplotypes. Nodes in the median joining network are proportional to frequencies and branch lengths to the number of polymorphic base substitutions.

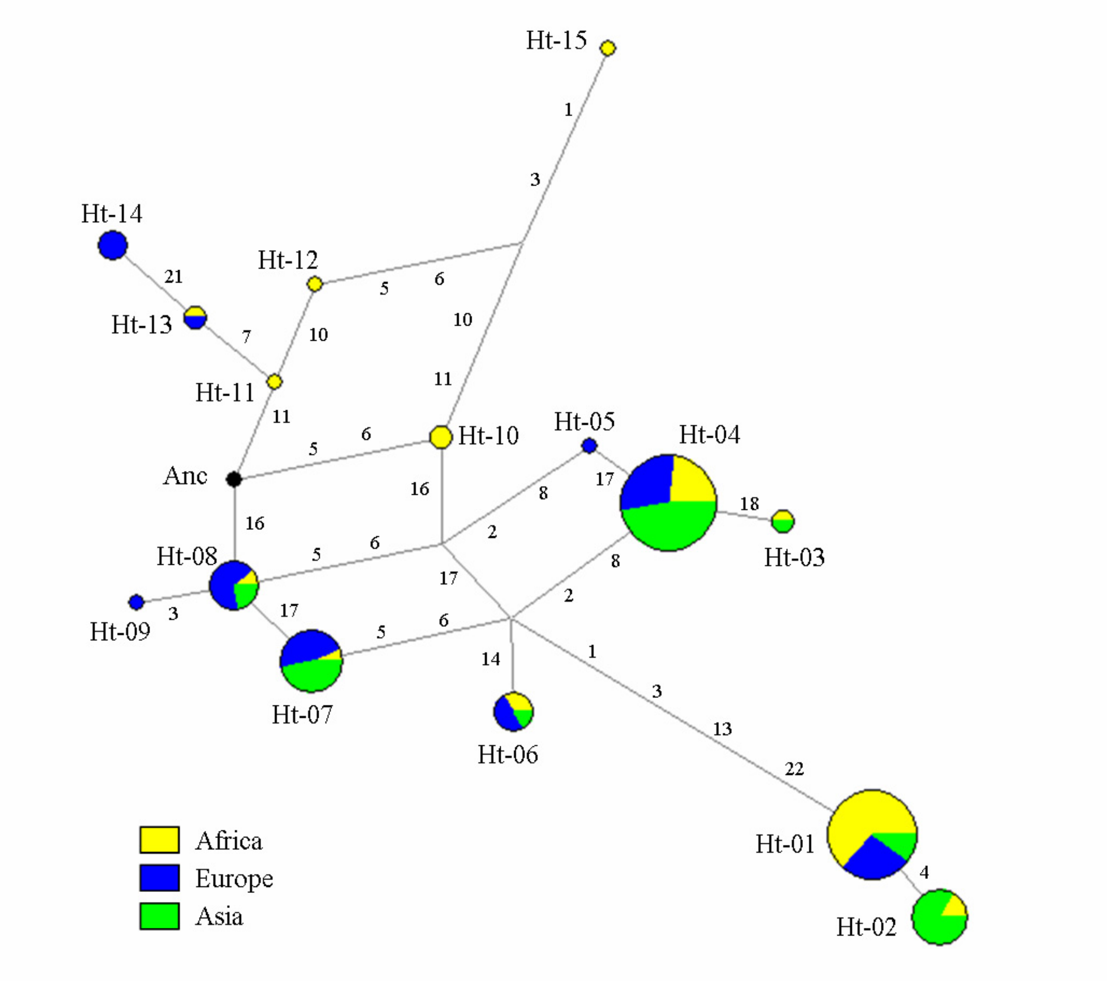
Figure S2. Distribution of low-frequency minor alleles. The proportion of SNPs with minor allele frequency (MAF) less than 0.10 within 100 Kb sliding windows is plotted for each population.

Figure S3. Distribution of high-frequency derived alleles. The proportion of SNPs with derived allele frequency (DAF) greater than 0.80 within 100 Kb sliding windows is plotted for each genomic region and population.

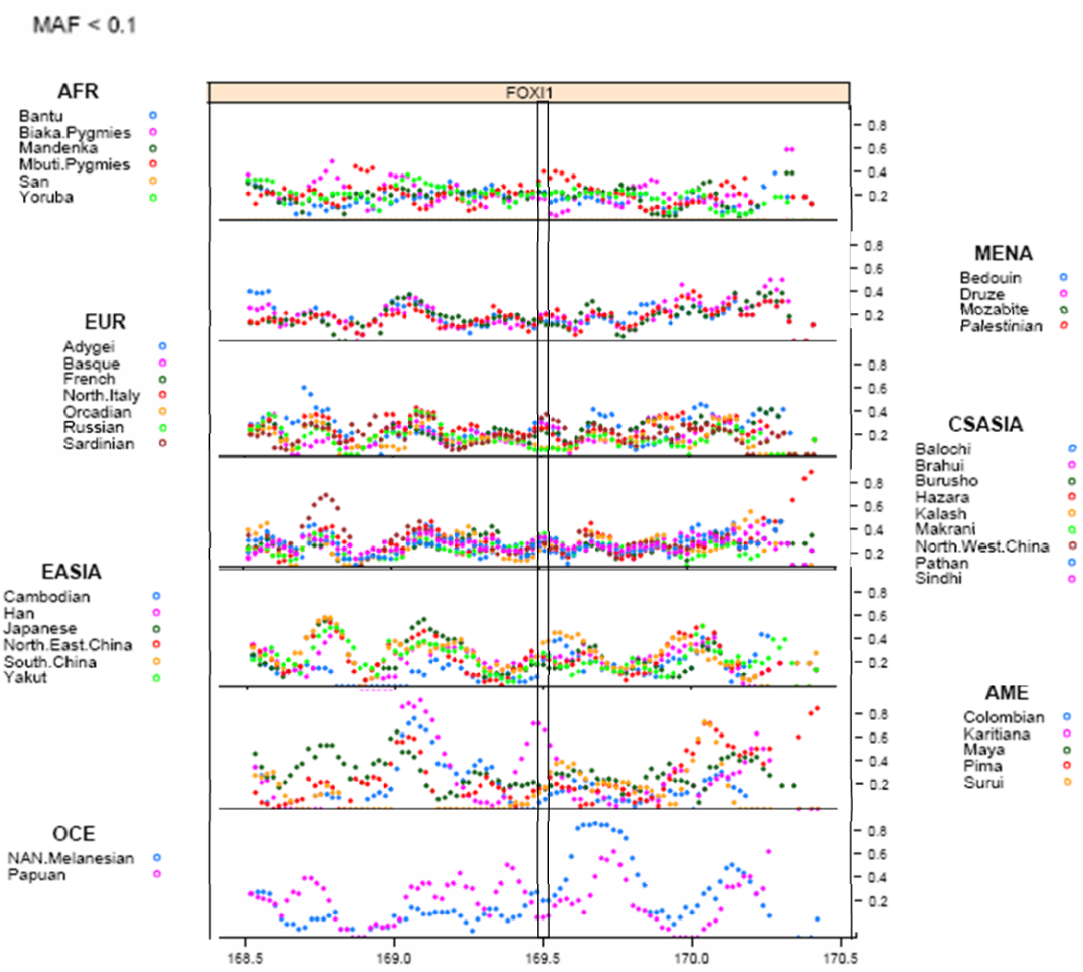
Figure S4. Decay of linkage disequilibrium around *FOXI1* in Africa. (A). Genes and SNPs on the region. Boxes represent genes, vertical gray lines are SNPs, the vertical blue line denote that constituting the core and vertical red lines indicate non-synonymous SNPs. Underlined SNPs represent other cores within the region. (B). Breakdown of EHH over genetic distance. (C). Decay of REHH over physical distance. (D) Haplotype Bifurcation Plots.

Figure S5. Allele frequencies at rs7736379 as a function of absolute latitude in 39 HGDP-CEPH populations. The line represents linear regression.

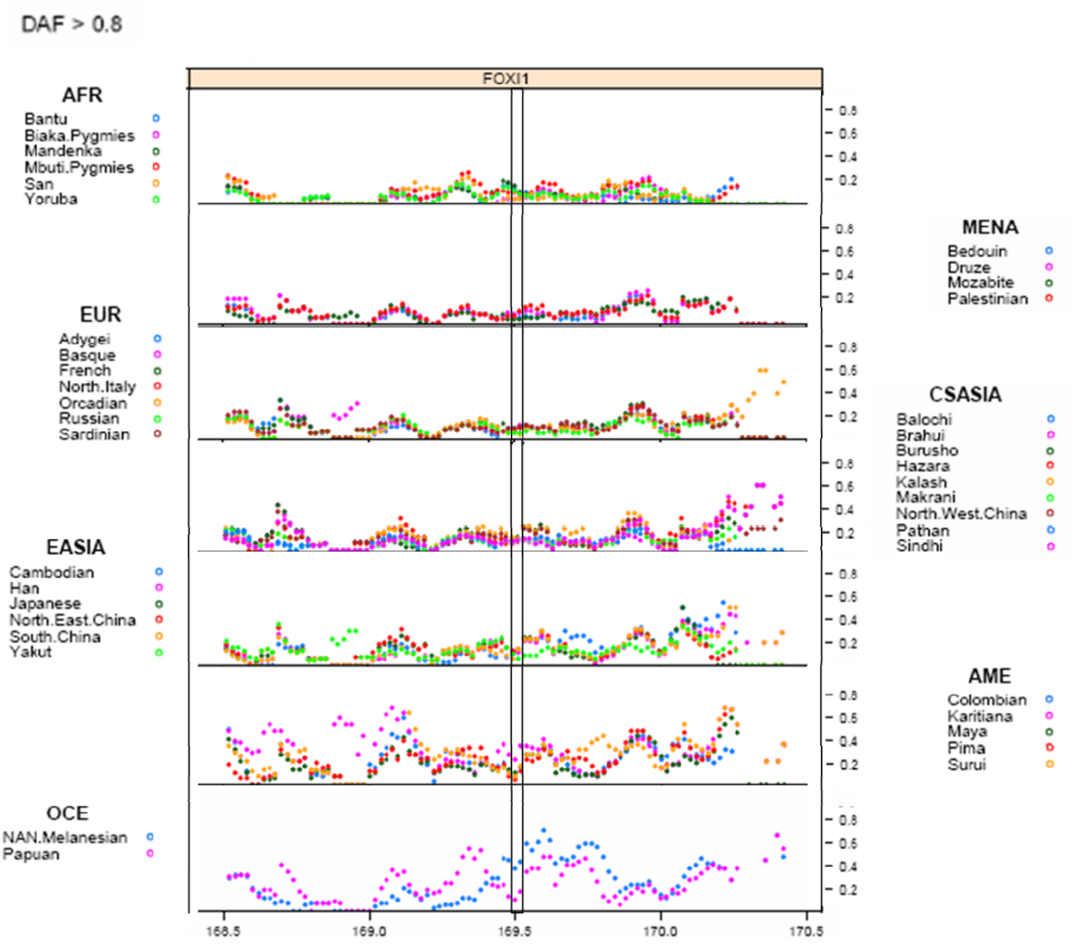
Supplementary Figure S1.



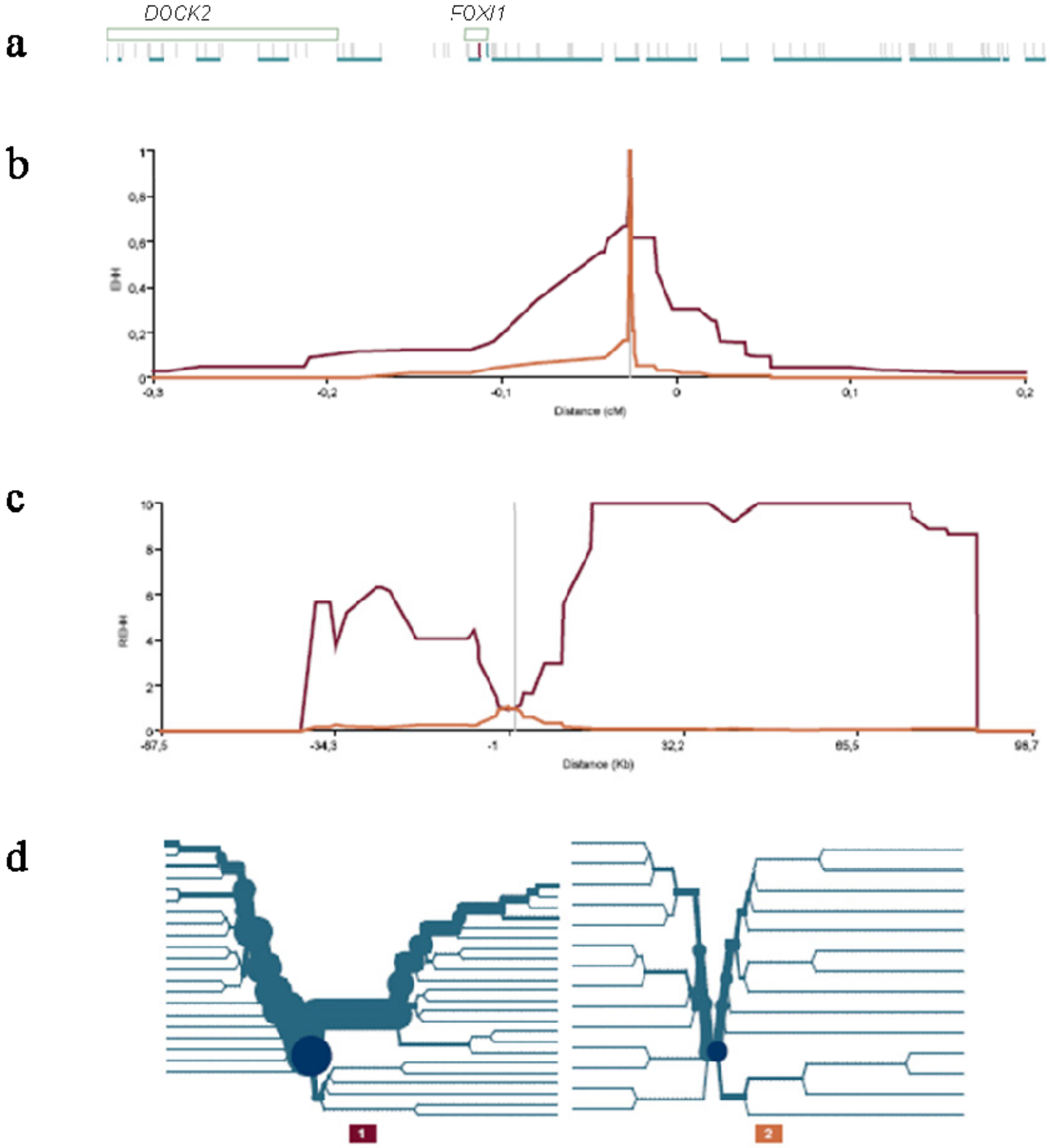
Supplementary Figure S2.



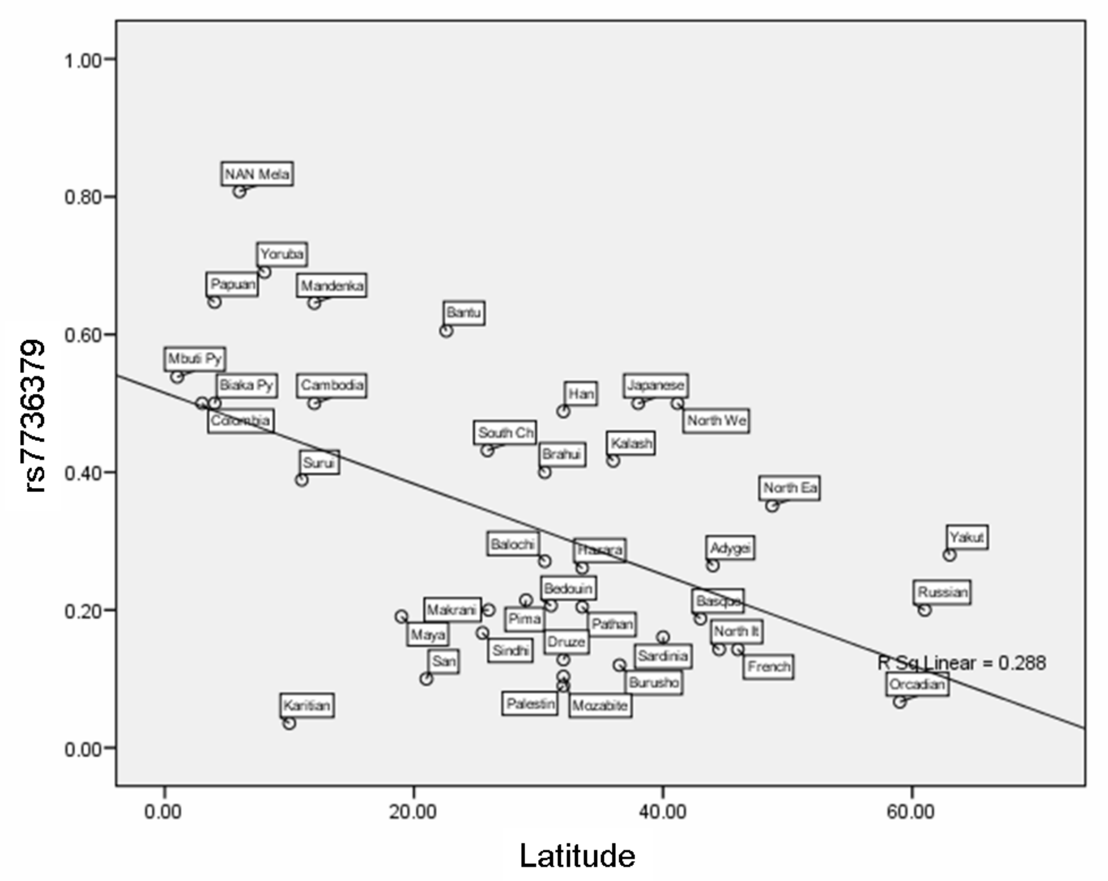
Supplementary Figure S3.



Supplementary Figure S4.



Supplementary Figure S5.


rs7736379
Latitude
NAN Mela
Yoruba
Papuan
Mandenka
Bantu
Mbuti Py
Biaka Py
Cambodia
Colombia
Han
Japanese
North We
South Ch
Brahui
Kalash
Surui
North Ea
Balochi
Hazara
Adygei
Yakut
Bedouin
Russian
Makrani
Pima
Basque
Pathan
Maya
Sindhi
Druze
North It
Sardinia
San
French
R Sq Linear = 0.288
Burusho
Orcadian
Karitian
Palestin
Mozabite