UNIVERSITÀ DEGLI STUDI DI TORINO

Dipartimento di Biotecnologie Molecolari e Scienze per la Salute
Department of Molecular Biotechnologies and Health Sciences

Scuola di Dottorato in Scienze della Vita e della Salute
Doctoral School in Life and Health Sciences

Dottorato di Ricerca in Scienze Biomediche ed Oncologia (XXIX ciclo)
PhD Program in Biomedical Sciences and Oncology (cycle XXIX)

Investigating the Evolution of the Human Genome at Different Scales

---

Author:
Davide MARNETTO

Supervisor:
Prof. Paolo PROVERO

SSD:
BIO11

PhD Coordinator:
Prof. Emilio HIRSCH

TESI DI DOTTORATO

20 Dicembre 2017

Abstract

Davide MARNETTO

Investigating the Evolution of the Human Genome at Different Scales

In this thesis, I will first introduce the reader to the framework of my research, briefly reviewing players, modes and times through which the human genome has been shaped, and tools to detect them, with a focus on gene regulation. Secondly, two projects at the opposite ends of the temporal and spatial spectrum previously described will be presented.

In the first we estimated the age of each region of the human genome by applying maximum parsimony to genome-wide alignments with 100 vertebrates. We then studied the age distribution of several types of functional regions, with a focus on regulatory elements, in order to assess the role of genome expansion in the evolution of gene regulation. Many transcription factors have expanded their repertoire of targets through waves of genomic expansions that can be traced to specific evolutionary times. Repeated elements contributed a major part of such expansion, with features which suggest that several binding sites were available as soon as the new sequence entered the genome, rather than being created later by accumulation of point mutations. By comparing the age of regulatory regions to the evolutionary shift in expression of nearby genes we show that rewiring through genome expansion played an important role in shaping human regulatory networks.

In the second working case we presented *Haplostrips*, a tool to visualize polymorphisms of a given region of the genome in the form of independently clustered and sorted haplotypes. This tool can reveal hidden patterns of genetic variation without losing the basic information encoded in variant sequences, and can be applied to visualize complex effects of, among others: introgression, domestication, selection, demographic events. We showed how *Haplostrips* helped in the investigation of the LCT region, a quintessential example of adaptive selection in humans, and in the discussion of regions likely subject to adaptive introgression from archaic hominins.

The study of molecular evolution at different scales in time and space involves the use of different tools and approaches, which are not trivially transferable to study the same functional features , as in the case of gene regulation analysis.

Contents

Chapter 1

Introduction

In this thesis, I will first introduce the reader to the framework of my research, briefly reviewing players, modes and times through which the human genome has been shaped, and tools to detect them, with a focus on gene regulation. Secondly, two projects at the opposite ends of the temporal and spatial spectrum previously described will be presented.

Especially for people outside evolutionary biology, it is never too much to restate that, as in a paper suggestively called "Non-Darwinian evolution" [1]:

> "The stream of spontaneous alterations in DNA, continuously fed into the genetic pool, should include far more acceptable changes that are neutral than changes that are adaptive [...]: the genome becomes virtually saturated with such changes as are not thrown off through natural selection."

Thus, this is a story about the relentless struggle between genetic drift and natural selection, to shape our genome starting from the unceasing emergence of mutations in individual organisms throughout life history.

1.1 Introduction to our focus organism: us.

Evolutionary biology is a basic science that studies the evolutionary processes that produced the diversity of life on Earth, starting from a single common ancestor. As a basic science, it has the enjoyable property of being free to study organisms other than human, as its results are not expected to give an immediate application on health and medicine. Nevertheless this thesis, and my research interests behind it, is about *Homo sapiens* because it obeys to the extreme fascination of advancing towards the solutions to our all-time unanswered questions: why are we human? where we come from, why are we like this, at all scales?

We will start this story half a billion years ago (540-510 Million Years Ago; Mya), when the first vertebrates originate in a period called "Cambrian Explosion" and known for a swift rise in organism diversity [2] . These animals had a basic vertebrate body plan: a notochord, rudimentary vertebrae, and a well-defined head and tail, but lacked jaws in the common sense, resembling modern-days lampreys. Indeed the first organism to part ways with us in the vertebrate phylogenetic tree that we will consider in the following is the *Petromyzon marinus*, or lamprey. In the first case study presented we will explore the story of our genome from here, following it as all other vertebrates split progressively from our ancestral species: first

the fishes, the amphybians, reptiles and birds together, and later, zooming on mammals, monotremes, marsupials, and so forth. Each one of these splits define a human ancestor, which is shared with smaller and smaller groups of organisms, bit by bit more similar to us. We have for example the Ur-Primate, which is the most recent common ancestor of all primates, that in turn will evolve into a plethora of species, defining several human ancestors: the Ur-Haplorrine, Ur-Simian, Ur-Catarrhine... We reach in time another human ancestor: the Ur-Homininae. This is the ancestor that we have in common with only three other living species (Gorilla, Bonobo, Chimpanzee), probably originated in Africa sometime during the early-middle Miocene and lived as a unique species until gorillas split, some 10-15 Mya [3]. This period witnesses a rapid radiation sequence where the most recent common ancestor of Chimpanzees and Humans is still in contact with the first *Gorillini*, still exchanging genetic material, but is splitting in two already: the Pan tribe, that will later develop into two species, and the future human species [3]. Through this process, sometime around 6-8 Mya, we lost contact with our closest living relatives, Chimpanzees and Bonobo; but the story to reach us, the anatomically modern humans, does not stop yet.

The earliest fossils candidate to reside in the exclusively human lineage belong to *Sahelanthropus*, *Orrorin*, *Ardipithecus* genera. The subsequent Australopithecus genus evolved in eastern Africa around 4 Mya before spreading throughout the continent and eventually becoming extinct 2 million years ago [4]. Another step towards the anatomically modern humans is represented by the onset of the genus Homo some 2.8 Mya in nowadays Ethiopia [5]: *Homo habilis*, *ergaster* and *erectus* were the first to use stone tools and to develop increased brain dimensions, likely following the duplication of the SRGAP gene [6]. *Homo erectus* is also thought to be the first to leave Africa and colonize the rest of the world as rests were found in western and eastern Asia dating back to 1.8 Mya [7] and 780 thousand year ago (kya) [8] respectively.

With a jump to 200 thousand years ago in Eurasia we might encounter some sporadic group of archaic humans that look and behave very closely to modern ones. They are not *Homo erectus* descendants, that probably disappeared everywhere with the possible exception of Indonesia [9]. These sturdy archaic humans are the well known Neanderthals, likely the result of a second migration out of Africa given that their split time with anatomically modern humans is estimated to be around 600 kya [10]. They are not alone, as individuals belonging to a sister group called Denisova have been found in the Altai mountains. Meanwhile in Africa the last remaining lineage of the *Homo* genus already started to move its first steps: the deepest difference among anatomically modern humans, from recent estimates, happened no later than 260 kya [11]. Africa is therefore the cradle of our species but we find modern human samples in Siberia already 45 kya [12]. Is under discussion if all modern day non-african populations are the result of a single emigration out of Africa, or the result of multiple events [13, 14]. What is known is that certain modern human populations admixed with archaic groups of humans after expanding out of Africa. In particular, non-African populations have 1-2% Neanderthal ancestry [10, 15], and Melanesians and East Asians have 3% and 0.2% ancestry, respectively, from Denisovans [10, 16, 17]. Some of these contributions were adaptive [18, 19] though, a larger proportion of introgressed genetic material was likely maladaptive to modern humans, and therefore selected against [20]. The simultaneous decline of the archaic humans has been an enigma for long, though these and other data suggest that they were instead integrated in our ancestry [21]. The subsequent peopling of the whole world, first Europe, Asia and Oceania, then the Americas through Bering

strait, concludes for now this story, as the unceasing migrations and admixtures of our populations defines what we are now and what we will be in the next future.

1.2 Evolution at Different Spatial and Temporal Scales

1.2.1 "Chromosomal" Evolution: Gain and Loss of Genomic Regions

The largest spatial scale in terms molecular evolution of the genome, is represented by chromosomal rearrangements. Such changes are usually caused by a breakage in the DNA double helix at two distant locations, followed by a disordered rejoining of the broken ends. This can produce the classical deletion, duplication, inversion or translocation events, but also a relatively smaller scale events generically described as segmental rearrangements [22]. Such events can either result in lethal or dangerous effects for the organism, chromosomal rearrangements are a hallmark of tumoral cell lines [23], or turn out to be neutral and even advantageous, giving rise to a lineage of organisms that incorporate these changes.

Another group of players in shaping the genome at this scale is represented by the transposable elements (TEs), which are repeated and mobile DNA sequences, with the ability to invade genomes: they generally represent a substantial fraction of the genome, but vary depending on the species [24, 25]. Class I TEs, or retrotransposons, use reverse transcriptase to copy an RNA genome into the host DNA, they are divided into Long Terminal Repeat (LTR) and non-LTR elements; among LTR-elements, the human endogenous retroviruses (HERVs) resemble retroviruses, among the non-LTRs, some lost their mobile autonomy, becoming shorter. Class II elements, or DNA transposons, use the DNA element itself as the template for transposition; see figure 1.1, extracted from *Kazazian et al.* [25] for a depiction of elements that we can find in our genome.

Segmental duplications and TEs thus generate elements which are resembling to each other and scattered around the genome, collectively named repeated elements. They can give place to gene paralogs, gain other functionalities, remain neutral: in any case they are recognisable as insertions when comparing genomes of species affected or unaffected by one of these events; not infrequently they can as well get deleted in time. Any genome is hence the result of unceasing insertions and deletions, which define the genome size in a perpetual arms-race [26]: e.g. in Amniotes the (often large) amount of DNA gained via lineage-specific transposition is essentially balanced by the amount of DNA lost over the same time frame. Is then straightforward to segment the human genome in regions of different age, that means which were inserted in different human ancestors and never deleted from our genome (see Chapter 2). Some of these regions might even have been inserted after the split between Human and Chimpanzee (or gained before and deleted elsewhere than Human) emerging as Human-Specific genomic regions.

These changes, altering the genome on a wide scale, very often break apart some cellular mechanism, disfavouring or killing the bearer, tending therefore to be preserved relatively seldom during evolution. The rate of chromosome rearrangement along the vertebrate lineage has been estimated to be between 0.1 to 2.3 changes per million years (My) [27] while, in terms of quantity, an analysis on Amniotes evaluate the insertion of sequence from 0.1 to 11 Mb gained per My and genomic loss at a rate of 2 to 13 Mb lost per My, mainly through large scale deletions (> 10kb) [26]. This

does not necessarily mean that these changes do not happen as frequently as lower order mutations, only that they are likely to be discarded at a higher rate.

Lastly, since mutations at this spatial scale have been historically studied comparing genomes of different species, researchers relied on a single genome representing each one of them, i.e. the reference genome, thus losing intra-species nuances. The advent of whole genome intra-species variation studies already began to change this, as these projects already catalogued these large scale events, defined Structural Variants, in several human populations (see Section 1.3.2).

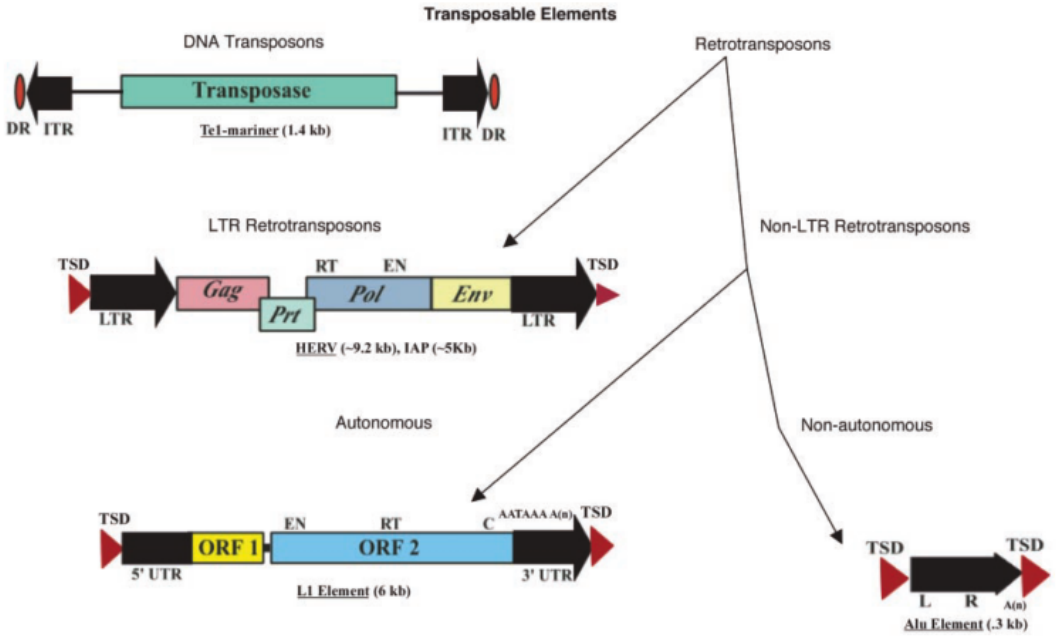

FIGURE 1.1: ***Classes of transposable elements.*** *From Kazazian et al. [25].*

1.2.2 Evolution at Single Positions

As we zoom in, reducing the spatial scope, we enter the realm of single nucleotide mutations, which evolutionary speaking can be subdivided into substitutions or polymorphisms.

Substitutions

A substitution is a point mutation that, after appearing in a population as novel allele, reaches fixation replacing the original allele. On the other hand single nucleotide polymorphism (SNP) refers to mutations which are variable within the population of interest. It goes without saying that the substitution is the smallest difference observable when comparing different species using a reference genome, as there is no knowledge of the intra-species frequency of that mutation. Consequently some of the substitutions identified by Comparative Genomics may instead be polymorphic within a species. Nonetheless, we can exclude that these account for the majority of the cases, given that the chance of these variations of being incorporated in the reference genome increases linearly with their frequency [28], and that the vast majority of SNPs has low frequency. Furthermore, as the average time to fixation is

often much smaller than the evolutionary distance between two species, the presence of heterospecific SNPs, or SNPs in common between two species, is negligible [29]. The study of substitutions had countless implications, e.g. the popular theory of the Molecular Clock: under the classical Wright-Fisher model the substitution rate is dependent only on the mutation rate μ, so substitutions should happen regularly in time, although recent experimental advances suggest that the substitution rate and the mutation rate do not always trivially coincide [30].

Polymorphisms

If we focus on smaller evolutionary times, then, the fundamental process to look at is how we go from the emergence of a mutation to its fixation within a species. How two principal mechanisms, genetic drift and natural selection, have shaped genetic variation during this process remains a pivotal question at the crossroads of Evolutionary and Population Genomics. Given that the average time to fixation of neutral alleles in a diploid population, conditional on the allele fixing, is approximately $4N_e$ generations [31], if we assume human $N_e = 10,000$, this scope translates for humans into the last million of years, time-wise speaking. Nevertheless, in the presence of selection, this process can be much faster, proportionally to the selection coefficient, which in turn depends on how much an allele is beneficial, see figure 1.2 A extracted from *Otto et al.* [32].

The main quantity studied at this end of the spectrum is the allele frequency (AF), which can be compared across populations and time points, to infer distances between samples and strength of selection coefficients at single positions. AF represents the basis for a great deal of summary statistics as the Site Frequency Spectrum (SFS; see Section 1.3.2), because is one of the quantities that are more affected by natural selection. To have an intuitive idea of how selective regimes affect the frequency spectrum, see figure 1.2 B extracted from *Nielsen* [33]. Unfortunately a very well known problem in using SFS and other measures mentioned below, is that demographic events like bottlenecks and rapid population expansions dramatically influence them, making difficult to disentangle selection from demographic effects.

Recombination and Linkage Disequilibrium

Thus far we considered the study of evolution as based only on the evolutionary mutations themselves, but within small time scales the appearance and rise in frequency of a mutation is accompanied by unequivocal effects at nearby positions. Indeed a mutation will be inherited together with its surroundings and recombination takes time to break them down, defining haplotypes, which are arrangements of specific alleles occurring in the same chromosome within a given genetic segment. The size of ancestral haplotypes around a mutation is inversely correlated to the time, in generations to the common ancestor. The rationale is apparently simple: small haplotypes indicate ancient mutations, even if some substantial difficulties are encountered in deciphering past events [34, 35]. This phenomenon can be used to infer very recent ancestry or pedigree relations, as in Fu et al [12] where the authors used it to validate recent Neanderthal ancestry in a 40,000 years old modern human sample.

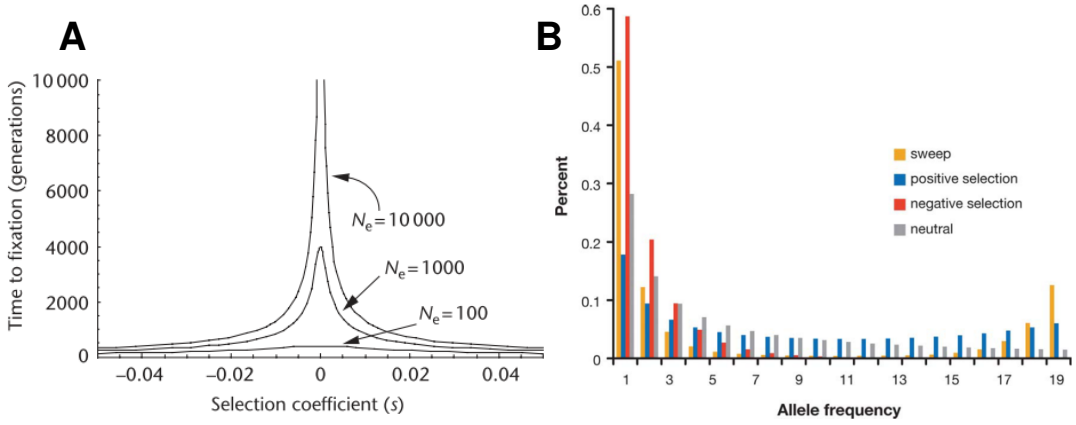

FIGURE 1.2: *(A)* ***Mean time to fixation in a diploid population.*** *From Otto et al. [32]. The average time to fixation, conditioned on the fact that the allele does fix, is a decreasing function of $|s|$ where s is the selection coefficient. Alleles will also fix faster, on average, in populations of smaller effective size. Note that the mean time in generations, to fix neutral alleles ($s = 0$) is $4N_e$. (B)* ***Frequency spectrum under a selective sweep, negative selection, neutrality, and positive selection.*** *From Nielsen [33]. For the selective sweep, the frequency spectrum is calculated in a window around the location of the adaptive mutation immediately after it has reached fixation in the population. In all cases, a demographic model of a population of constant size with no population subdivision is assumed.*

The measurable quantity often used in exploring this phenomenon is the Linkage Disequilibrium, which is simply the correlation between two nearby alleles: LD has the property to decay geometrically with the relative distance of the two markers, and the extent of this decay can be altered by demographic events and natural selection, generating wide regions with many SNPs at high frequency showing highly correlated allelic states. In fact, dips in genetic diversity along the genome have been associated with the effects of both negative and positive selection, delivering the basis for many summary statistics for selection [33, 36]. It is noteworthy that as the LD is a useful clue for detecting recent evolutionary events, it represents a great complication in functional studies as well: see Section 1.4.3 for a brief introduction to this issue.

Genetic Flow and Admixture

A species is not a well defined group of individuals that suddenly, when speciation occurs, split forming two independent phylogenetic branches [37]. It has instead an internal structure composed by populations that can interbreed (hybridization), exchange genetic material (genetic flow), fuse together introducing new genetic lineages (admixture), incorporate each other alleles (introgression). It is only when these process are impossible, due to physical separation or genetic divergence (often resulting in sterile offspring), that two populations can be considered separate species [38]. These phenomena leave traces as well, resulting for example in local phylogenetic trees that differ from the species phylogeny [39] or in stretches of alleles that segregate at a frequency consistent with one population followed by sequences more consistent with another one. This idea is exploited in many methods for ancestry inference, where we interpret each chromosome in an individual genome as a mosaic of segments that originate from different ancestral populations [40]. On the other hand, isolated segments deriving from one entity (population or species)

found into the gene pool of a second, divergent entity, might be better described as introgressions, as shown for archaic human introgressions in modern humans [10].

1.3 Tools to Investigate Evolution at Different Scales

1.3.1 Genome Wide Alignments

The most important and basic tool used in evolutionary biology and probably the most accomplished in the whole bioinformatics field, is sequence alignment. Nowadays this is frequently declined in the form of Genome Wide alignment, as it allows A) to perform scans on the whole genome to focus *a posteriori* on the most interesting features and B) capture biological effects on the genome in their completeness.

One of the most popular publicly available alignments across vertebrate species is the Multiz 100-way [41], also accessible through the UCSC genome browser. This particular case is a multiple alignment, but also pairwise alignments between many published reference genomes are publicly available in the form of NET alignments [42] on the same Genome Browser.

The main use of genome wide alignments is the same of classical local alignments, i.e. to identify regions of similarity that may be a consequence of functional, structural, and moreover evolutionary relationships between the sequences. In this way we can compare the behaviour of orthologous regions and ultimately of the single nucleotidic bases in different species, a task that allows to identify substitutions and that poses the basis for comparative genomics. Accordingly, countless applications focused on the regions that can be aligned, e.g. infer strength of natural selection on different lineages [43], mutation rates [44], to compute conservation scores [45], to identify ultra-conserved sequences of putative biological significance [46, 47].

It is also possible to focus on those region that cannot be aligned between species. This is what we did in our first case study [48], where we used gaps in multiz 100 way alignment to develop a segmentation of the human genome based on sequence age. Each region can be present or absent in each of the extant aligned species and one can use, let us say, a parsimony algorithm to infer present/absent/unknown state in each ancestral node, thus defining a likely time of birth for any window of the genome considered. We used this method to analyse the traces of sequence expansion in the human genome since the common ancestor of all Vertebrates, attempting to reconstruct its role in rewiring regulatory networks. If we are interested in species-specific regions we can simply consider those regions that are gaps in all the other aligned species, ignoring any deeper stratification, as we did in *Marnetto et al.* [49].

Lastly, alignments can be coupled with functional annotation, to compare elements with same function, e.g. genes, as is done in phylostratigraphy [50], where alignments of genes across species and within the same genome are used to describe gene phylogeny and age.

1.3.2 Variation Studies

As we can align genome of different species to identify orthologous regions, we can trivially do so among individuals of the same species, using the same reference genome as scaffold. In this case the small differences encountered identify intra-species polymorphic traits, which can be small scale, SNPs and short insertions/deletions (indels), or large-scale, i. e. structural variants (SV).

These variation studies gained popularity as the cost of whole genome sequencing decreased and represent nowadays a milestone in the whole biology field. It is worth remembering that before sequencing, other techniques as SNP genotyping arrays [51] allowed variations studies of remarkable proportions, like the HapMap project [52]. The gold standard in human variation studies is currently represented by the 1000 Genomes project [53] which reconstructed the genomes of 2,504 individuals from 26 populations using a combination of low-coverage whole-genome sequencing, deep exome sequencing, and dense microarray genotyping, discovering in total 84.7 million SNPs, 3.6 million indels and 60,000 SVs. Lastly, as these projects now focus on rare mutations [54] and previously excluded populations in humans [13, 14], variations analyses are being applied other organisms like the Great Apes [28]

Methods in Population Genetics

Determining the contributions of natural selection in the evolution of a population is one of the central questions in human evolutionary biology since way before the access to these data. As result a vast body of theoretical literature available for over 50 years [1, 56, 57] is now applied on empirical data under the form of diverse summary statistics that exploit haplotype or frequency information. For example, haplotype-based statistics such as iHS (integrated haplotype score [58]), EHH (extended haplotype homozygosity [59]), $XP-EHH$ (cross- population extended haplotype homozygosity [60]), and the nSL statistic (number of segregating sites by length; [61]) (see Table 1, extracted from [55]) are based on the observation that haplotype homozygosity should be greater around a positively selected locus than in a neutrally evolving locus. In parallel, statistics depending on allele frequencies such as the SFS (site frequency spectrum; [33, 62]), F_{ST} (fixation index [63]), and PBS (population branch statistic; [64]) (Table 2, extracted from [55]) can be applied to one, two, and three populations respectively to identify regions under positive selection. Noteworthy higher level summary statistics based on the SFS are the Tajima's D, Fay and Wu's H and similars [65]. These methods keep updating to investigate even the most recent human history: a good example is the Singleton Density Score (SDS), a method to infer very recent (2,000-3,000 years ago) changes in allele frequencies from contemporary genome sequences [66].

These summary statistics are greatly valuable as they provide meaningful knowledge, but come at the cost of sacrificing the incredible amount of basic information encoded in the genetic sequences. The alternative to this is having to deal with raw data, i.e complex patterns of genetic variation, thus tools to examine these patterns in their completeness are also needed to further elucidate the underlying evolutionary processes. As our second case study we present *Haplostrips* [67], a tool to assist researchers in visualizing the polymorphisms of a given region of the genome, providing the user with a few options to reveal hidden haplotype structure that may

Method	Explanation	Main Information	Reference
EHH (extended haplotype homozygosity): $\mathrm{EHH}(X_i) = \sum_{h \in C(x_i)} \frac{n_h/2}{n/2}$ $\mathrm{EHH}_c(X_i) = \sum_{h \in H_c(x_i)} \frac{n_h/2}{n/2}$	The EHH score is calculated given a locus of interest and the *i*th marker upstream or downstream x_i. h is all the observed types (unique combinations of SNPs) from SNP x_0 to x_i. n_h is the number of haplotypes of type h, and n is the total number of haplotypes. EHH_c is very similar to EHH but is calculated only for samples carrying the core haplotype c.	Linkage disequilibrium and haplotype	Sabeti et al. 2002
iHS (integrated haplotype score): $\mathrm{iHH}_c = \sum_{i=1}^{\lvert D\rvert} \frac{1}{2[\mathrm{EHH}_c(x_{i-1}) + \mathrm{EHH}_c(x_i)]\, g(x_{i-1}, x_i)} + \sum_{i=1}^{\lvert U\rvert} \frac{1}{2[\mathrm{EHH}_c(x_{i-1}) + \mathrm{EHH}_c(x_i)]\, g(x_{i-1}, x_i)}$ $\text{unstandardized iHS} = \ln\left[\frac{\mathrm{iHH}_0}{\mathrm{iHH}_1}\right]$ $\text{standardized iHS} = \frac{\ln\left[\frac{\mathrm{iHH}_0}{\mathrm{iHH}_1}\right] - Ep\left[\ln\left(\frac{\mathrm{iHH}_0}{\mathrm{iHH}_1}\right)\right]}{SDp\left[\ln\left(\frac{\mathrm{iHH}_0}{\mathrm{iHH}_1}\right)\right]}$	The iHS score calculates the decay of EHH for the ancestral ($c = 0$) and derived ($c = 1$) haplotype extending from a query site. D is a set of markers downstream from the marker of interest, and U is a set of markers upstream from the marker of interest. $g(x_{i-1}, x_i)$ is the genetic distance between x_{i-1} and x_i. E and SD denote the expected value and standard deviation respectively.	Linkage disequilibrium and haplotype	Voight et al. 2006
XP-EHH (cross-population EHH): $\mathrm{XP\text{–}EHH} = \frac{\ln\left[\frac{\mathrm{iHH}_A}{\mathrm{iHH}_B}\right] - \ln\left[\frac{\mathrm{iHH}_A}{\mathrm{iHH}_B}\right]}{SD\left[\ln\left(\frac{\mathrm{iHH}_A}{\mathrm{iHH}_B}\right)\right]}$	XP-EHH is a cross-population EHH score that looks at the ratio of the iHH score in two different populations.	Linkage disequilibrium and haplotype, cross-population	Sabeti et al. 2007
nSL (number of segregating sites by length): $\mathrm{SL}_c = \sum_{i=1}^{\lvert D\rvert} \left(\frac{1}{2[\mathrm{EHH}_c(x_{i-1})]} + \mathrm{EHH}_c(x_i)\right) g(x_{i-1}, x_i) \times \sum_{i=1}^{\lvert U\rvert} \left(\frac{1}{2[\mathrm{EHH}_c(x_{i-1})]} + \mathrm{EHH}_c(x_i)\right) g(x_{i-1}, x_i)$ $\mathrm{nSL} = \frac{\ln\left[\frac{\mathrm{SL}_1}{\mathrm{SL}_0}\right] - E_p\left[\ln\left(\frac{\mathrm{SL}_1}{\mathrm{SL}_0}\right)\right]}{SD_p\left[\ln\left(\frac{\mathrm{SL}_1}{\mathrm{SL}_0}\right)\right]}$	nSL is very similar to the iHS score, and the only difference is that $g(x_{i-1}, x_i)$ correspond to the distance in base pairs rather than the recombination distance, so $g(x_{i-1}, x_i) = 1$.	Linkage disequilibrium and haplotype	Ferrer-Admetlla et al. 2014

Table 1. Haplotype-Based Methods to Detect Positive Selection *From Antelope et al. [55].*

not be apparent when the haplotypes are plotted in a random order. We will then briefly review two applications of this tool from [55] and [19].

Method	Explanation	Main Information	Reference
π (average number of pairwise differences): $\pi = \frac{1}{n/2}\sum_{i=1}^{n-1} i(n-i)\xi_i$	π is the mean pairwise sequence differences among a sample of n haplotypes. ξ_1 is the unfolded site frequency spectrum.	Allele frequency	Nei 1987
F_{ST} (fixation index): $F_{ST} = \frac{H_T - H_S}{H_T}$ $H_S = 1 = \frac{1}{I}\sum_{i=1}^{I}\frac{1}{J}\sum_{j=1}^{J} p_{ij}^2$ $H_T = 1 = \frac{1}{I}\sum_{i=1}^{I}\left(\frac{1}{J}\sum_{j=1}^{J} p_{ij}\right)^2$	H_T is the mean pairwise sequence differences in the total population, and H_S is the same statistic in a subpopulation. It can be used to detect population structure, and local outliers of high values of F_{ST} suggest selection specific to a subpopulation. p_{ij} is the frequency of allele i in population j.	Allele frequency, cross-population	Nei 1973
PBS (population branch statistic): $T^{A,B} = -\log(1 - F_{ST}^{A,B})$ $PBS_A = \frac{T^{A,B} + T^{A,0} - T^{B,0}}{2}$	$T^{A,B}$ is an estimate of the divergence time between populations A and B as a function of genetic differentiation, F_{ST}.	Allele frequency, cross-population	Yi et al. 2010

Table 2. Frequency-Based Methods to Detect Positive Selection *From Antelope et al. [55].*

1.3.3 Ancient Genomics

The advance of sequencing and amplification techniques has made possible retrieving ancient DNA (aDNA) from old specimens: after the death of an organism, normally sequestered catabolic enzymes, bacteria, fungi, degrade all macromolecules, but when a tissue becomes rapidly desiccated or the DNA becomes adsorbed to a mineral matrix, it may escape this degradation. Still, the aDNA analysis presents great complications due to damage and contamination [68].

An obvious example of damage is the high fragmentation of the sequence. Estimates of decay rates of mtDNA in bone under the best preservation conditions predict that no intact bonds (average length = 1 bp) will remain in the DNA strand after 6.8 Myr, and that nuclear DNA should degrade twice as fast [69]. Nevertheless this degradation estimate is subject to great individual variability and is dramatically reduced when considering conservation at higher temperatures. Another example of damage is the deamination that can occur on DNA bases causing misincorporations of nucleotides during the amplification process, typically A instead of G, and C instead of T [68].

Contamination in turn represent another extremely serious concern in the study of aDNA. Many ancient samples contain no endogenous DNA detectable with current techniques, but exposure to other organisms, as well as excavators, museum personnel, or laboratory researchers, fill the sample with exogenous DNA. This is even more serious in the study of human aDNA, where primers that amplify contemporary human DNA are used to perform amplifications. The weapons to face this complication are to date: the evolution of techniques to avoid contamination, bioinformatics pipeline to asses its degree [70], the adoption of severe criteria of authenticity [68].

All these obstacles do not reduce the profit of using ancient samples: these snapshots from the past provide an unprecedented level of resolution to delineate in detail how genetic variation changed through time. We can look for signatures of positive selection in the ancient populations and ask whether we also observe those in the current populations or vice-versa [55]. For example, [71] sequenced 101 ancient humans from across Eurasia from the Bronze age and compared them to populations from different time periods (Paleolithic, Mesolithic, and Neolithic) to study population migrations, as well as the temporal dynamics of selected genetic variants. Another study [72] investigated data of 230 ancient individuals from West Eurasia from 6,500 to 1,000 BCE. These and other studies [73] resolved previous highly variable estimates on the timing of selection, revealed the potential origin of the putative beneficial genetic variants and underlined how present-day populations were created by a complex history of admixture and population movement. Although Improved methods for DNA extraction have now started to yield some ancient data sets from regions until now left aside due to challenges with the rate of DNA deterioration, as the African continent [11].

1.4 A Focus on the Regulation of Gene Expression

The investigation of regulatory evolution has been of pivotal importance in the human evolution since the observation that, considering Human and Chimpanzee *"Their macromolecules are so alike that regulatory mutations may account for their biological differences"* as stated in 1975 with a seminal paper by King and Wilson [74]. Empirical evidence and theoretical arguments suggest indeed that the rewiring of gene regulatory networks plays an important role in the evolution of metazoan anatomy [75, 76]. Such arguments are supported by a large body of experimental evidence demonstrating, in specific cases, how the evolution of anatomical traits is triggered by adding or subtracting targets to a trans-acting regulatory element [77–82]. Genetic events affecting gene regulation can be classified between two extremes: exaptation of existing sequence through the accumulation of small-scale mutations, and de-novo appearance of regulatory DNA through genome expansion driven for example by transposable elements (TE). Both mechanisms have been shown to be relevant in the evolution of human regulatory DNA[83–90].

1.4.1 Regulatory Features

What are the features on which we want to focus? A very well known operative unit in the task of regulating gene expression is the Transcription Factor Binding Site, which can be identified with two main approaches. The first, that we can define "theoretical" is the adoption of tools that recognise binding motifs on the genome sequence, a classical example of which are the Positional Weight Matrices [91]. This method is inferring binding sites, which can optionally be validated *a posteriori*, but allows to investigate the molecular nature and affinity of the binding: an extension of this is the Total Binding Affinity approach [92]. The second is the "empirical" way: the use of Chromatin Immuno Precipitation, coupled with sequencing (ChIPseq) to target a Transcription Factor of interest and obtain the cross-linked positions. this result in less precise, but validated and cell-line specific, binding "peaks": the ENCODE ChIPseq datasets are the gold standard for such analysis.

Is very well known that TFBS cluster together in defining regulatory regions: this genomic feature can be studied as well with the use of ChIPseq, but targeting chromatin markers known to be present in active regulatory elements. These include H3K27ac for active regulatory states, H3K4me3 for promoters of active genes, H3K4me1 for enhancers: the NIH Roadmap Epigenomics Mapping Consortium [93] is providing an incredible dataset of such and other markers (e.g. Dnase Hypersensitive Sites which describe windows of open chromatin). In the past downstream analyses were made to combine these data [94], but has been difficult to summarize them. Now is more common to see them used independently, although is known that these and other markers (coregulatory complexes, nucleosomes, DNA methylation...) are largely redundant and likely predictive of each other [95].

These data are available for human and some model organism (mouse, drosophila) but remain generally missing for other species. To perform comparative studies the researchers can project human data on other species, taking into account the uncertainty about the regulatory activity in these species (this is what we did in Chapter 2) or perform *ad hoc* data generation. A beautiful example of this approach is presented in *Villar et al.* [87] where they compared regulatory histone marks in 20 mammals.

Lastly it is of the utmost importance to underline that not only regulatory features can be identified and then projected onto the evolutionary scope: historically the opposite has been much more common. Indeed, the identification of non-coding conserved elements [47, 96] has led the researchers to consider these important in the regulation of nearby genes [97].

1.4.2 Effects of Regulatory Evolution

As the reason-of-being of coding elements, that is to code for proteins, is important in the study of their evolution, the effects on gene expression of regulatory elements are key in investigating theirs. Under the comparative point of view this generated some brilliant studies as the one published in [98] where they use the expression of orthologous genes in 9 amniotes to infer dramatic expression shifts at precise ancestors in specific tissues. To correlate these shifts to the evolution of nearby regulatory regions is not trivial, as the relation between a regulatory element and its target genes remains somewhat elusive (see a partial success in Chapter 2). By the way, this task is likely to improve markedly, given the revolutionary importance of recent studies on topologically associated domains, insulator binding factors and highly dimensional functional correlations [99].

On the population genetics side, the study of the functional effects of regulatory variation was recently revolutionized by the systematic identification of expression Quantitative Trait Loci (eQTL). As Genome Wide Association Studies (GWAS) for disease traits were burgeoning, the researchers realized that many of the variants associated with phenotypes or diseases were falling in non-coding sequences. Therefore the same principle of GWAS has been adopted to find variants associated with a more basic trait: the expression of a gene in a particular cell line [100, 101]. More specifically, the expression level of a gene is tested against the allelic state of all SNPs or at least, to reduce the combinations to be tested, of the SNPs nearby (*in cis*) in order to find a correlation. The rationale is that if they correlate, the SNP tested might regulate the expression level of the gene, even if correlation alone cannot prove the

causal relationship. In this case again an effect of evolution (intra-species variation) has been used to identify regulatory features, that can in turn be used to investigate the evolution of expression of target genes.

1.4.3 Critical Points at Different Time Scales

Traditionally, GWAS have been hampered by the Linkage Disequilibrium: intuitively, relying on correlation between genetic variant and phenotypic trait, GWAS cannot discern among causal variants and other markers correlated with the causal one, i.e. in high LD. eQTLs claim to be slightly better, perhaps because they correlate with a clearly quantitative and simpler trait (gene expression), their association Pvalue has often enough power to discern the causative variant, nevertheless this is not always true and can be confounded by several factors[100]. Moreover, even in this case, when the causal variant is at low frequency in the analysed population or is in nearly perfect association with other variants, is virtually unrecognisable. Variants which have undergone interesting evolutionary stories are even more difficult to pinpoint: regions under adaptive selection are known for their low diversity stretches that can extend for hundreds of kilobases (see above). Notwithstanding, when we learn that the median length of completely independent blocks is over 1MB [102] and that the LD decays reaches plateau at about 60kb in Africans and 100kb in Europeans and Asians [53], we can grasp how problematic can be this issue also for human variants with common evolutionary histories.

Especially when comparing the average gene length (10-15kb) to the distances listed above, we can understand why the difficulty of discerning the feature that is under selection for recent evolutionary times is crucial. Indeed, we cannot exclude that the traces of recent selection are due to the coding part of the genes, unless we develop specific methods to focus on regulatory features, e.g. TFBS [103–105]. Another difficulty is tied to the fact that, contrary to coding sequences, where synonymous positions can be used as neutral reference [106], identifying a neutral region to compare with the selected regulatory feature is not trivial. The practice usual for other evolutionary scales (see below) is to use putatively neutral flanking regions, but LD can force us to go so far that these flanks have no more biological meaning.

Over larger evolutionary times, many of these problems are absent, thus explaining the large number of studies about regulatory evolution in comparative genomics (see introduction to this Section (1.4)). One of the problems that remain is the small size of some regulatory features, e.g. TFBS, that can be overcome when considering sets of small coherent elements, an approach exploited by INSIGHT [107]. This approach relies on the comparison with flanking regions, thus having difficult application on intra-species distances, but also makes not trivial the process of extrapolating conclusions obtained genome wide for the application to a single regulatory element.

Chapter 2

Evolutionary rewiring of human regulatory networks by waves of genome expansion

2.1 Introduction

Evolution of regulatory networks is believed to underlie a significant fraction of the phenotypic divergence between vertebrates [75, 76, 79]. Genetic events affecting gene regulation can be classified into two classes: exaptation of existing sequence through the accumulation of small-scale mutations, and de-novo appearance of regulatory DNA through genome expansion driven for example by transposable elements (TE). Both mechanisms have been shown to be relevant in the evolution of human regulatory DNA [83–89].

In particular, information-rich binding sites (BS) such as the one recognized by *CTCF* [MIM 604167] are much less likely to arise through the accumulation of random point mutations than simpler binding motifs: indeed it was shown [85] that the expansion of lineage specific transposable elements efficiently remodeled the *CTCF* regulome. The activity of TEs in generating transcription factor binding sites (TFBS) was studied more generally by *Sundaram et al.* [86], where it was observed that about 20% of BS were embedded within TEs, thus revealing the latent regulatory potential of these elements [84]. The role of a specific class of TEs in generating transcription factor binding sites was recently investigated by *Ito et al.* [89]. On the other hand, it was shown [87] that recent enhancer evolution in mammals is largely explained by exaptation of existing, ancestral sequence rather than by the expansion of lineage-specific repeated elements. A systematic investigation of the role of genomic sequence expansion in rewiring regulatory networks is however still missing.

In previous works we investigated the most recent evolution of human regulatory networks by looking at both promoter sequence divergence [108] and genomic expansion after the split from the chimpanzee [49]. Here we attempt to reconstruct a much longer evolutionary history, focusing on regulatory evolution through genome expansion since the common ancestor of all Vertebrates. To this aim we develop a segmentation of the human genome based on sequence age. By overlaying this segmentation onto Transcription Factor (TF) binding data we can reconstruct how successive waves of genome expansion modified the regulome of each TF.

We then examine some signatures that can help determine whether the binding sites were present at the time of the appearance of the new sequence, or were created

later by progressive accumulation of point mutations. For repeated elements, these signatures include the specificity of the TFs binding each class of repeated elements, the existence of preferred locations of the binding sites within the repeated elements, and of distinctive motif words. Finally, we use comparative transcriptomics data to determine the effect of such waves on the evolution of gene expression.

2.2 Methods

2.2.1 Classifying the human genome based on sequence age

For each base of the human genome (hg19 release) we assigned presence/absence in modern vertebrates using the Multiz 100-way aligment [41]. A base of the human genome is present in another species if it aligns to sequence (regardless of matching) rather than a gap in such species. We defined regions of the genome as maximal stretches of consecutive bases sharing the same presence/absence vector. The state of each region in ancestral nodes was reconstructed through a parsimony algorithm on a fixed tree. The tree was obtained from the Multiz 100-way documentation, and led to the definition of 19 ancestral nodes ranging from *Homo sapiens* to the ur-Vertebrate.

For each region, the parsimony algorithm uses as input the present/absent state of the region in each of the extant aligned species and returns a present/absent/unknown state in each ancestral node. We defined the age of the region as the oldest ancestor where the algorithm returned a non-absent state (present or unknown). All analyses were repeated with the opposite choice (age as the oldest ancestor with a present state) to check that our results do not depend on such bias.

For most of the genome regions (99.5% of the sequence) the reconstructed history is consistent with a single birth event (that is the region is reconstructed as absent in all ancestors older than its assigned age). The regions for which this did not happen were discarded. We also removed from our genome all regions annotated within the Gap track (downloaded on 11/22/13) in the UCSC database.

2.2.2 Overlap with repeated sequence and preferential insertions

For Repeat Elements (RE) annotation we used the Repeat Masker (downloaded on 11/19/12) and Simple Repeat (downloaded on 10/20/15) tracks, from UCSC database. We labeled as 'Transposable Elements' those belonging to DNA, LINE, LTR, Other, and SINE classes. The overlap of inconsistently reconstructed regions to repeat classes was tested against 1000 randomizations obtained by shuffling their genome positions.

To test whether insertions happen preferentially into recent genome we removed all regions smaller than 50bp and we defined as insertions all regions such that the two flanking regions are older and of equal age. We then counted the insertions happening inside regions of all possible ages, and tested their distribution against the null model in which the probability of insertion into a given age is proportional to the total genomic sequence of that age, using a χ^2 test.

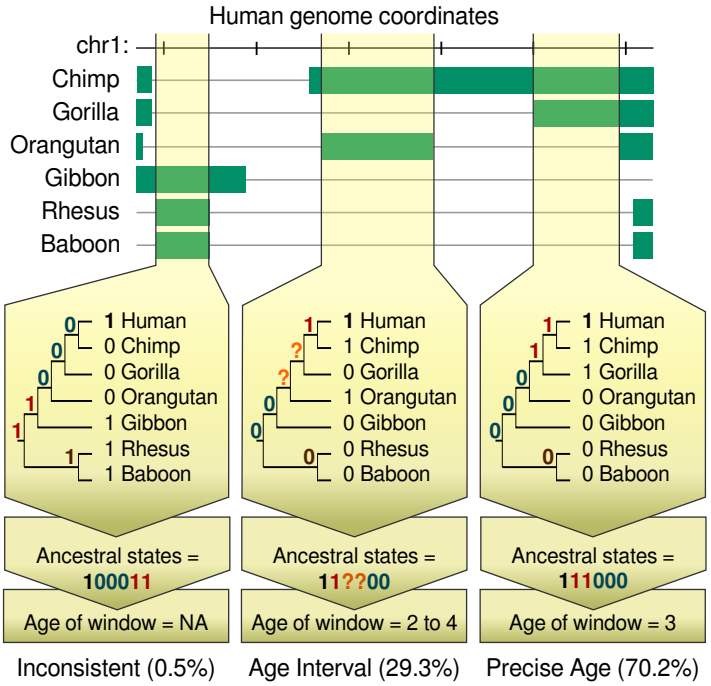

FIGURE 2.1: Genome segmentation in ages. *Maximum parsimony applied to a multiple alignment of the human genome to 100 vertebrates is used to infer the presence of a genomic region in each ancestor and thus the evolutionary age of the region. A green bar indicates that the human genome region aligns to sequence, rather than to a gap, in the non-human genome. Each region can be classified into one of three classes. "Inconsistent": The reconstructed ancestral states are incompatible with a single birth event for the region; "Age Interval": The reconstruction is compatible with single birth but only a time interval can be estimated; "Precise Age": The precise time of the birth of the region can be estimated. The fraction of genome sequence falling into each class is shown.*

2.2.3 Evolutionary age of functional classes of sequence

To each RefSeq gene we attributed five functional classes of sequence: coding exons, non-coding exons and introns were defined based on RefSeq annotation. To define regulatory classes we used the genome segmentation of *Ernst et al.* [94]: we consider as regulatory sequence all sequence that is not inside an exon and falls into one of the regulatory classes (classes 1-8) in at least one cell line, after removing the cancer-derived cell lines Hepg2 and K562. We then classified as *promoters* all regulatory sequence within -5kbp and +1kbp of a RefSeq transcription start site and *extended regulatory* all regulatory sequence within the extended regulatory region defined associated by GREAT [109] to the gene.

2.2.4 Evolutionary age and gene expression

We used the expression profiling of human tissues from the RNA-seq Atlas [110] and we defined a gene as expressed in a tissue if its RPKM was > 1. We then compared the mean genomic age of tissue-specific or ubiquitary genes vs. all other genes with a Wilcoxon-Mann-Withney test, separately for each sequence functional class. We also evaluated the Spearman correlation between mean genomic age and gene expression expressed as RPKM.

2.2.5 TFBS distribution in age

To investigate the age distribution of TFBS we collected all ENCODE ChIP-seq datasets, merged all the binding sites of the same TF from different cell lines and removed data from non-sequence-specific binding events (PolII and general transcription machinery) or non endogenous (GFP-conjugated proteins).

We removed peaks wider than 5kb and set a minimum peak width of 100bp, enlarging narrower peaks up to this size. We repeated the analysis with 0 or 500 minimum peak width cutoff. To each TFBS we attributed the age of its median point, after discarding all peaks of the same TF overlapping a narrower peak. We then compared the age distribution of the binding sites of each TF to a null model defined by the age distribution of all TFBSs taken together, using a χ^2 test. The chi squared residuals are visualized in Figure 2.3 and Supp. Fig. A.6 and represent the enrichment of binding sites of a given TF in regions of each evolutionary age. To generate an empirical *P*-value, we performed 5000 times the following randomization: divide the genome into (unequal) windows, each containing exactly 200 TFBS; each window defines a sequence of 200 TF names; randomly permute these sequences of names among the windows. In this way the TFs are assigned randomly to genomic regions while conserving the local clustering of binding sites of the same TF.

2.2.6 TFBS enrichment in Gene Ontology categories

TFBS were associated to genes using GREAT [109]. We evaluated the enrichment of TFBS of each evolutionary age in genes belonging to GOs where specific peaks of regulatory innovations had been identified [83]: GO:0003700 (transcription factor activity), GO:0032502 (developmental process), GO:0005102 (receptor binding) and GO:0043687 (post-translational protein modification). The enrichment was defined as the fold enrichment of the number of TFBS of each age associated to genes in each GO category compared to the number of TFBSs of the same age regardless of GO association.

2.2.7 TFBS enrichment in RE classes

To evaluate the enrichment of RE classes we used Fisher's exact tests: for each TF/age/RE combination, we tested whether the binding sites of the specific TF tend to overlap instances of the RE more often than binding sites of the same age irrespective of the identitiy of the bound TF. *P*-values were Bonferroni corrected. Repeat coordinates were obtained from RepeatMasker. We also explored the possibility of testing the overrepresentation across ages, rather than across TFs: that is testing whether the binding sites of a given TF and age tend to overlap instances of the RE more often than binding sites of the same TF irrespective of the age. However the age-specificity of most REs drives the significance of this test in most cases, so that we find a much larger number of significant results which include virtually all results obtained with the the test across TFs. We thus deemed safer to use the test across TFs.

2.2.8 TFBS position in REs

We kept all REs overlapping a TFBS, of length between 80% and 120% of the RE model length obtained from RepeatMasker metadata, discarding shorter and longer instances. We tested only those TF/age/RE combinations which after this filter retained at least 10 instances: only transposon classes survived after this process. We annotated the positions of the TFBS peak median point, normalized them over the length of the TE instance, and built a distribution with bins of about 50bp: the precise size of the bin ensured that the TE model could be divided into bins of equal length. For each TF/age/TE combination, call A the set of the binding sites of any TF falling in the specified age and overlapping the RE, and S its subset where the TF is the one under investigation. We computed the entropy of the binned position distribution of S and compared it to 10,000 random subsamplings of A, each of size equal to $\#S$.

2.2.9 TFBS word composition

Using the JASPAR Core Vertebrates set [111], we were able to associate a Positional Weight Matrix (PWM) to 66 of the 127 TFs to be tested. For each TF/age/RE triplet to be tested, we searched each Chip-seq peak with the corresponding PWM, keeping the top scoring site and discarding peaks where such site scored less than half of the highest possible score. Accounting separately for TFBS falling on instances of the enriched RE class and for TFBS placed elsewhere, we computed a contingency matrix for all occurring words and obtained a P-value for it with a Monte Carlo simulation with 100,000 replicates, using χ^2 as the test statistic.

2.2.10 TFBS and gene expression shifts

TFBSs determined as above were associated to genes using GREAT [109]. For all the genes which underwent an expression shift in exactly one human ancestor (in any tissue) according to *Brawand et al.* [98] we counted the associated TFBSs of each evolutionary age. These counts were compared to those obtained in the same way after randomly shuffling 5,000 times the age of the expression shift and thus transformed into z-scores shown in Figure 2.7. Statistical significance of the enrichment of the diagonal element of each row was determined empirically by comparing the fraction of TFBSs of the same age as the expression shift to the distribution of the same number in the 5,000 randomizations.

2.2.11 Repeat-driven targets of *FOXP2* and *FOXA2*

Putative TFBS targets were determined using GREAT [109], which was also used to determine the functional enrichemnt of repeat-driven targets compared to all targets. Gene expression data in human tissues were obtained from the GTEx consortium [101]: we used the median expression of each gene in a given tissue across all GTEx samples of that tissue.

2.3 Results

2.3.1 Segmentation of the human genome by sequence age

To estimate the age of each region of the human genome we used a published multiple alignment of 100 vertebrate genomes [41]. Each region was classified as present or absent in each non-human species depending on whether it aligns to sequence or to a gap in the multiple alignment: therefore each region is characterized by a present/absent binary vector of length equal to the number of non-human species in the alignment (see Methods for details).

For each region the present/absent vector was used as input to a maximum parsimony algorithm to determine the most likely (most parsimonious) history of appearance/disappearance of the region during vertebrate evolution that explains what is observed in extant species. Note that parsimony was not used to reconstruct the phylogenetic tree, which was instead fixed, but only to reconstruct the presence or absence of the sequence in each ancestral node. We thus obtained a new vector expressing the presence of the region in progressively older ancestors, from the human-chimpanzee common ancestor to the common ancestor of all vertebrates (see Figure 2.1).

Similar principles have been used in previous works [83, 87, 88]. All of them evaluated the age of windows of interest using the most distantly related species with an alignable sequence. The use of a parsimony algorithm allowed us to exploit in a controlled way genomic alignments with a large number of species and thus provide a segmentation of the human genome by age that is more robust and dense in terms of the human ancestors considered.

For 70% of the genome this method allowed us to determine a precise age of birth and for another 29.5% an age interval (when the parsimony algorithm reported the presence of the sequence in some ancestors as uncertain). For this latter fraction we defined as age the upper end of the interval. While this leads to a systematic overestimation of genomic ages, all results reported in the following were essentially unchanged when the opposite choice was made.

Only for 0.5% of the genome the reconstructed history was inconsistent with a single birth event, and this fraction was enriched in Low Complexity, Simple and tRNA repeats ($P < 0.001$, permutation test), possibly reflecting sequencing and alignment problems; this part of the genome was excluded from further analysis. The age segmentation of the genome obtained in this way turned out to be robust with respect to the choice of the initial alignment data: using a collection of 47 pairwise "net" alignments obtained from the UCSC Genome Browser in place of the multiple alignment gave very similar results (A.1).

2.3.2 The age distribution of the human genome and the role of transposable elements

The age distribution of the human genome is shown in Figure 2.2A. Most of the human genome appeared after the split between placentals and marsupials: indeed only 13.5% of the human genome aligns with the opossum genome, while 43% aligns with the elephant (among the farthest eutherians from humans). The figure also

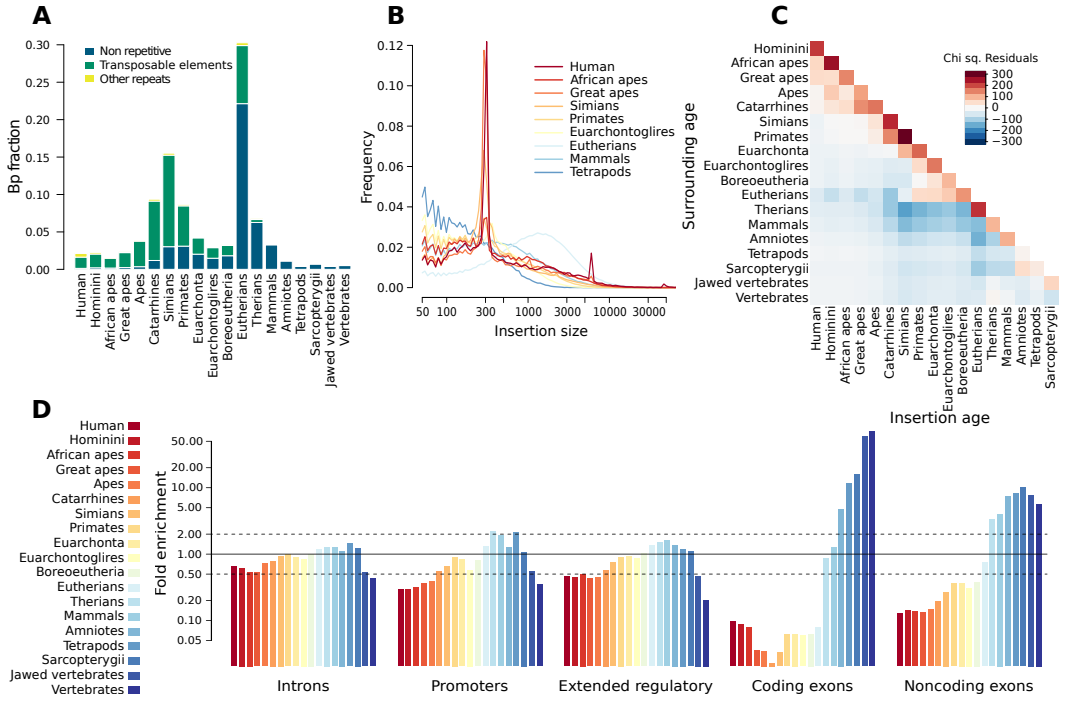

FIGURE 2.2: Genomic age distribution and properties. *(A) Age distribution of the genome. Most of the human genome appeared after the split between placentals and marsupials. Note the increasing relative importance of TEs as we move towards younger regions. This is partly due to the difficulty in recognizing TEs that have been integrated in the genome for very long times. (B) Length distribution of reconstructed insertions of each age. Colors are associated to ages as in (A). The major peaks at length $\sim$ 300 and $\sim$ 6000 correspond to ALU and L1 insertions respectively (C) Comparison between the age of genomic regions (x axis) to the age of the surrounding genome (y axis) shows that new genomic sequence is preferentially inserted in younger regions, possibly because these are subjected to less selective pressure. The heatmap shows the chi squared residuals with respect to a null hypothesis of no preferential insertion, in which the insertion probability in a region of a certain age is simply proportional to the genome fraction of that age (D) Enrichment of genomic ages in functional sequence classes. We show the coverage fold enrichment compared to what expected if all functional classes shared the genome-wide age distribution.*

shows the fraction of newly created sequence overlapping known TEs, which increases as we get closer to the present time. Older TEs are difficult to recognize today, and this probably explains at least in part their lesser prevalence in older regions of genome; nevertheless, since the ur-Boreoeutheria, the majority of newly gained sequence is still identifiable as TE. This indicates that TEs are an important driver of new sequence acquisition at least since then, and possibly further back in time. As expected, TEs are generally constant in age and their boundaries are close to age breaks, as seen in A.2.

If, as expected, most newly acquired genomic sequence is neutral or almost neutral we expect it to appear at an approximately constant rate in time. To verify whether this is the case we plotted the amount of human sequence of each age with the estimated time interval between the speciation events defining such age. The molecular clock hypothesis following from neutrality appears to hold up to the common ancestor of all Eutherians (Supp. Fig. A.3). For older ancestors it seems to break down, or at least to hold with a much lower acquisition rate, possibly because sequence that is conserved at very large evolutionary distances is unlikely to be neutral.

The length distribution of reconstructed insertions of each age is shown in Figure 2.2 B and is driven, in particular, by waves of expansion of Alu (length $\sim$ 300) and L1 ($\sim$ 6,000) retrotransposons in the ur-Primate and ur-Homininae respectively. New insertions happen preferentially in younger regions (Figure 2.2 C), presumably because younger regions are subjected to weaker selective constraints, leading to the appearance of insertion hotspots within young TEs [112].

When looking at the age distribution of various functional classes of the genome we see, as expected, that age increases with functional constraint (Figure 2.2 D): coding exons are made by the oldest sequence, while introns are the newest. Moreover we confirmed some known results relating age to expression patterns: older genes are more expressed than younger ones [113] (see Supp. Fig. A.4 A), and the coding exons of ubiquitously expressed genes are older than tissue-specific ones [114] (Supp. Fig. A.4 B). However the promoters of ubiquitously expressed genes are younger than those of tissue-specific ones, perhaps due to the relaxed constraints on their fine regulation. Indeed it was recently shown [115] that ubiquitously expressed genes have broader promoters and greater variation among individuals, suggestive of relaxed selective pressure, compared to the promoters of tissue-specific genes. Within tissue-specific genes the newest are expressed in testes and the oldest in the central nervous system. Importantly, the strategy by which we re-obtained these known results is completely independent from gene annotation, in contrast with the methods commonly used in classic phylostratigraphy [50, 116]. A comparison of our gene dating results with the age classes defined by *Neme et al.* [117], and those derived from the GeneTrees provided by Ensembl [118] is shown in Supp. Fig. A.5.

2.3.3 Genomic age enrichment of Transcription Factor Binding Sites

Newly acquired genomic sequence can contribute to the evolution of regulatory networks by creating binding sites for TFs [84–86]. To investigate this phenomenon in a systematic way we superimposed the results of ChIP-seq experiments performed on many TFs to the age segmentation and, for each TF, we asked whether significant age preferences could be discerned. Specifically we used a χ^2 test to compare, for each TF, the number of TF binding sites (TFBS) found in each genomic age to what expected under the null hypothesis where age preference is the same for all TFs. The null model thus incorporates any deviation from the uniform distribution displayed by TFBSs as a whole, and the test reveals the specific deviations of each TF.

Out of 139 TFs, 137 showed an age distribution significantly different than the null model ($P < 0.05$ after Bonferroni correction for multiple testing), with only *PPARGC1A* [MIM 604517] and *STAT2* [MIM 600556] not significant. TFBS local clustering could inflate the χ^2 *P*-values, and can be due to technical reasons (e.g. a single binding site interpreted as multiple peaks by the peak-calling software) or biological reasons (e.g. the accumulation of multiple binding sites of the same TF in regulatory regions). These effects can be controlled by counting as a single BS peaks closer than a given cutoff. Reassuringly, the enrichment results are essentially unchanged whether we use or not a cutoff, and with cutoffs of 100 bps and 500 bps. As an alternative strategy, we replaced the χ^2 *P*-values with empirical ones, by shuffling the TFBS in a block-wise manner: within each block the succession of TFBS is maintained, so as to maintain their local clustering, while blocks are randomly shuffled on the genome. This results in 135 significant TFs, excluding only two more TFs, *SIRT6* [MIM 606211] and *SMARCB1* [MIM 601607], compared to the χ^2 analysis. These results show that

most TFs show specific preferences in the age of the genomic sequence they bind. Such preferences can be visualized using the chi squared residuals, and are shown in Figure 2.3 and Supp. Figures A.6 and A.7.

Notably, such age enrichment in TFBS corresponds to specific functional enrichments of their target genes. In agreement with what reported by [83] for conserved non-coding regions, we observe significant TFBS enrichment near developmental and transcription factor genes in ages preceding the appearance of mammals; near receptor-binding proteins between the ur-Amniote and the ur-Eutherian; and enrichment in more recent ages near genes involved in post-translational modifications (see Supp. Fig. A.8).

2.3.4 Age enrichment suggests waves of TFBS expansions

The enrichment of binding sites of a given TF inside genomic sequence of a given age suggests an evolutionary process in which new genomic sequence extensively rewires transcriptional regulatory networks by providing existing TFs with abundant new targets. This mechanism was shown to have operated in the evolution of the *CTCF* regulome in mammals [85]. However, the fact that a human TFBS resides in a region that appeared at a certain time in evolution does not necessarily mean that the binding sites have the same evolutionary age. Indeed it is well known that TFBS can be generated within pre-existing sequence [87, 119–121]. In this mechanism new genomic sequence could simply provide raw material for evolution to act upon by accumulation of point mutations, possibly aided by relaxed negative selection, and create TFBS that were not there when the sequence entered the genome. Other effects, such as functional characterization of genomic regions with same origin and age, could also contribute to the age enrichments shown above. In the following we will use various signatures to identify the cases in which waves of genomic expansion indeed generated an immediate rewiring of regulatory networks.

Repeated elements carry specific motifs for specific TFs in specific portions of their sequence

If repeated elements (RE) carried TFBS at the time of their insertion in the genome, we expect to detect three signatures that are, instead, difficult to reconcile with TFBS creation by accumulation of point mutations. First, we expect each class of repetitive elements to be enriched with binding sites of just one or a few specific TFs; second, we expect such binding sites to be preferentially located in a specific portion of the repeated elements; and third, we expect the TFBS located in the RE to use a specific subset of the set of all possible motif-words (DNA *k*-mers compatible with the binding [85]).

We thus asked which of the age enrichments shown in Figure 2.3 could be ascribed to the expansion of a specific, recognizable repetitive element. We considered all TF/age pairs (cells in the heatmap shown in A.6) and for each of them we evaluated the number of TFBS overlapping each repeat class. We then tested whether such overlap was significantly enriched with respect to all TFBSs of the same age irrespective of the identity of the TF. That is, we asked whether the instances of a repetitive element appearing at a certain time tended to be associated to TFBSs of a specific TF.

We found a total of 3625 significant TF/RE/age triplets, involving 888 TF/age pairs. In Figure 2.3 these are shown as bordered cells. In most cases (2887 involving 600 cells) the enriched RE class is a TE. In Figure 2.4 we show the TFs whose binding sites are significantly enriched in each RE class, and the distribution of the number of TFs associated to each RE class. For most classes the enrichment in binding sites are restricted to one or a few TFs.

To determine whether TFBS occur in specific positions within repeated elements we considered all the enriched TF/RE/age triplets determined above and computed the entropy of the position distribution of the TFBS (represented by the median point of the ChIP-seq peak) within instances of the RE of the appropriate age. We then used a permutation test (see Methods) to determine whether such entropy was significantly

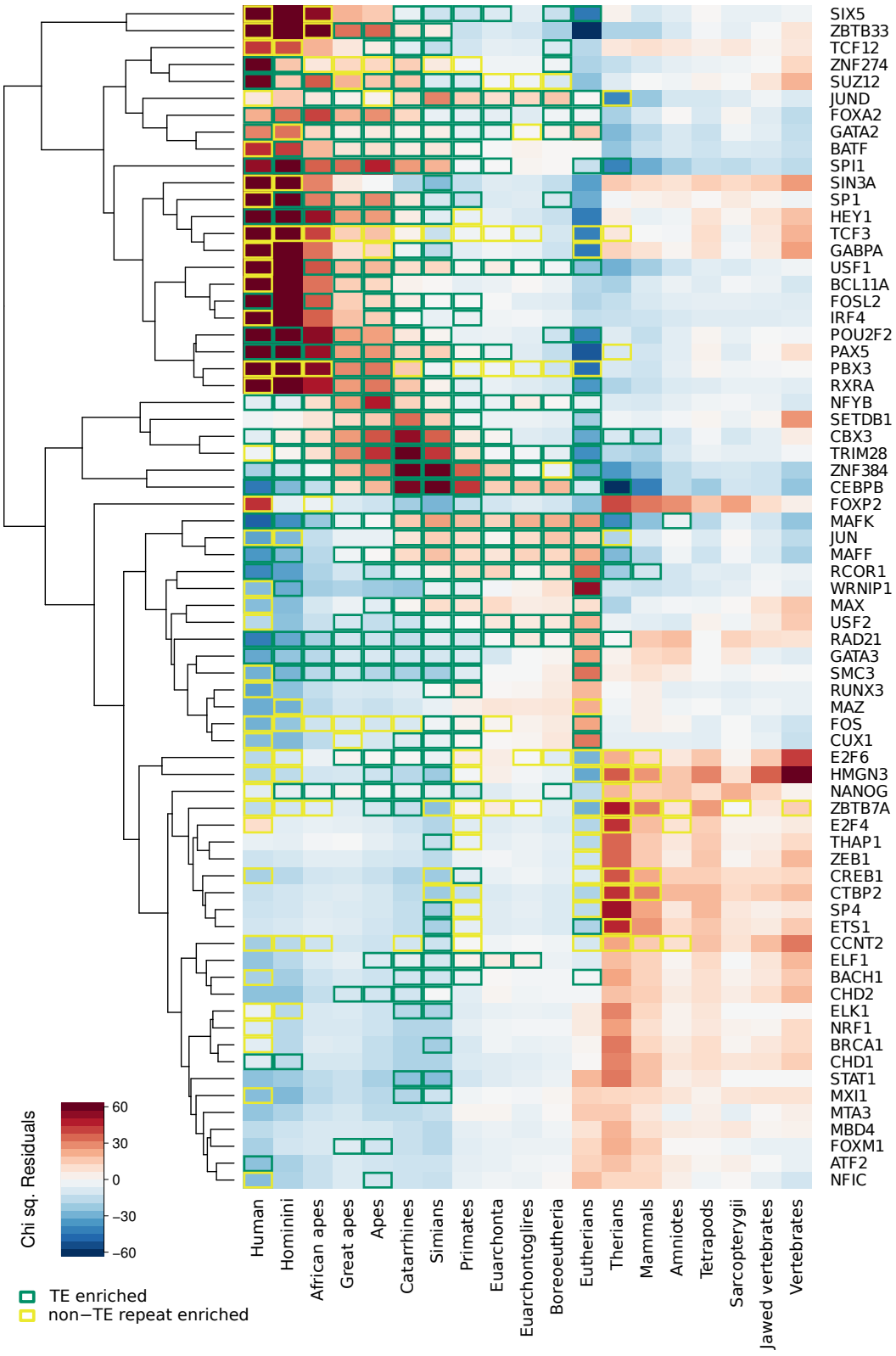

FIGURE 2.3: Enrichment of ChIP-seq peaks for each TF in genomic regions of different evolutionary age. *The heatmap represents the deviations of the age distribution of the binding sites of each TF from the overall TFBS distribution. Chi squared residuals are shown, so that positive values (red) correspond to enrichment and negative values (blue) to depletion. Only the top 50% TFs by significance are shown: a complete figure is available as Supp. Fig. A.6. TF/age combinations with significant repeat class enrichments are bordered, separately for non-transposon repeats (yellow) and transposon repeats (green).*

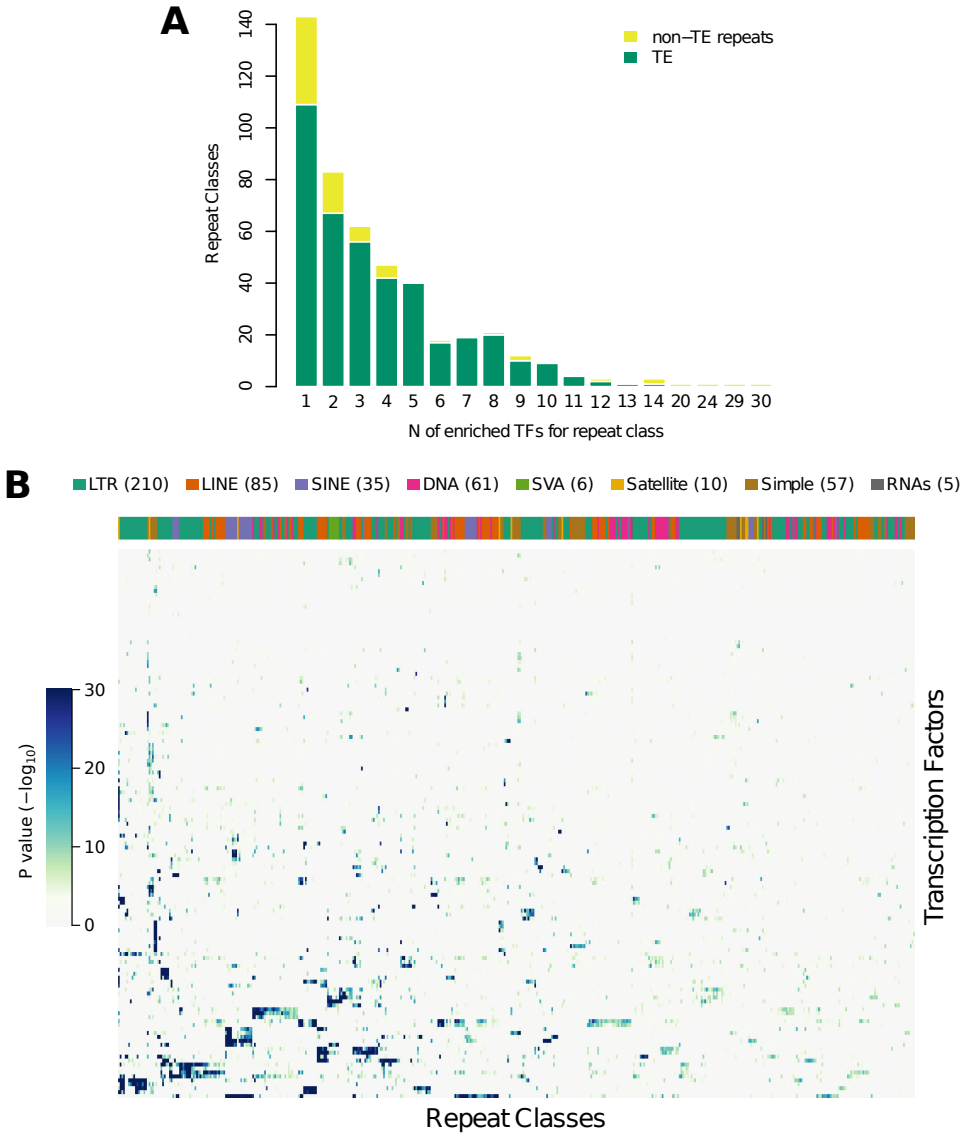

FIGURE 2.4: Classes of repeated elements carry specific TFs. *(A) distribution across RE classes of the number of enriched TFs. (B) The heatmap shows which TFs are enriched in each class of repeated elements: for each TF-RE class pair the age with the strongest enrichment is shown, colour intensity corresponds to $-log_{10}(P)$ (Fisher's exact test). RE classes are labeled by the top coloured bar, where each colour corresponds to a RE family. See the legend for a count of the RE classes included, divided by family.*

lower than that of the position distribution of all TFBS (of all TFs) on the instances of the same RE of the same age. The test thus shows whether the enriched TF is more precisely localized inside the RE compared to all TFs that bind the RE. In 1377 out of 1970 cases tested (involving 399 out of 495 TF/age combinations), we found a significantly lower entropy in the enriched TF/age/TE combination (Benjamini FDR 0.05, see Material and Methods for further details), as shown in Figure 2.5.

Considering again the set of enriched TF/RE/age triplets, we then examined if binding sites lying on the enriched RE used a specific set of motif-words with respect to binding motifs placed elsewhere, similarly to what was shown for *CTCF* [85]. To do so we identified the top scoring motif within all ChIP-seq peaks with a Positional Weight Matrix corresponding to the TF of interest and annotated whether the peak was falling on the enriched RE class or not. We then asked wether the two sets of TFBS differed in word composition: 523 out of 1707 TF/RE/age triplets tested (199 out of 366 TF/age combinations) were found significant against 100,000 simulations (Benjamini FDR 0.05, see Material and Methods for further details), as shown in Figure 2.6 A.

Altogether, in most cases of RE enriched in the binding sites of a TF we observe that the enrichment is not only specific for the identity of the TF, but also for the position of the binding within the RE and/or the word composition of the binding motif. Such specificity is difficult to reconcile with a process in which point mutations create binding sites in the ages after the appearance of the new sequence, and suggests instead that at least the genetic component of the regulatory rewiring was effected directly by the insertion of new sequence.

Network rewiring by genome expansion causes gene expression evolution

An important question is whether the TFBSs created by waves of genomic expansion are functional, that is whether they lead to changes in gene expression. *Brawand et al.* [98] used comparative RNA-seq data to identify genes whose expression underwent a shift at a certain point of the evolutionary history of mammals. By comparing the timing of expression shifts to the age of surrounding TFBSs we can investigate the role of genomic expansion in effecting gene expression evolution.

We considered all genes which underwent an expression shift [98] in exactly one human ancestor (except ur-Mammal and ur-Hominoidea for which the shift cannot be attributed to a branch [98]). We determined their associated regulatory region using GREAT [109] and, in such region, counted the TFBSs falling within each genomic age (Figure 2.7).

Such counts were then compared to 5,000 randomizations of the age of the gene expression shift, and thus transformed into z scores, shown in Figure 2.7. Given a genomic age a_G and an expression shift age a_E, the z score $z(a_G, a_E)$ is positive when TFBSs of age a_G are enriched in the regulatory region of genes which shifted their expression in a_E. Conversely, the z score $z(a_G, a_E)$ is negative when TFBSs of age a_G are depleted in the regulatory region of genes which shifted their expression in a_E. For the three evolutionary ages with the largest number of gene expression shifts (Human, Primates and Therians) we observed a statistically significant enrichment of TFBS with $a_E = a_G$ (permutation test based on the randomizations described above). These results show that the rewiring of regulatory networks by newly acquired genomic sequence results in an immediate, detectable change in the transcriptome. The

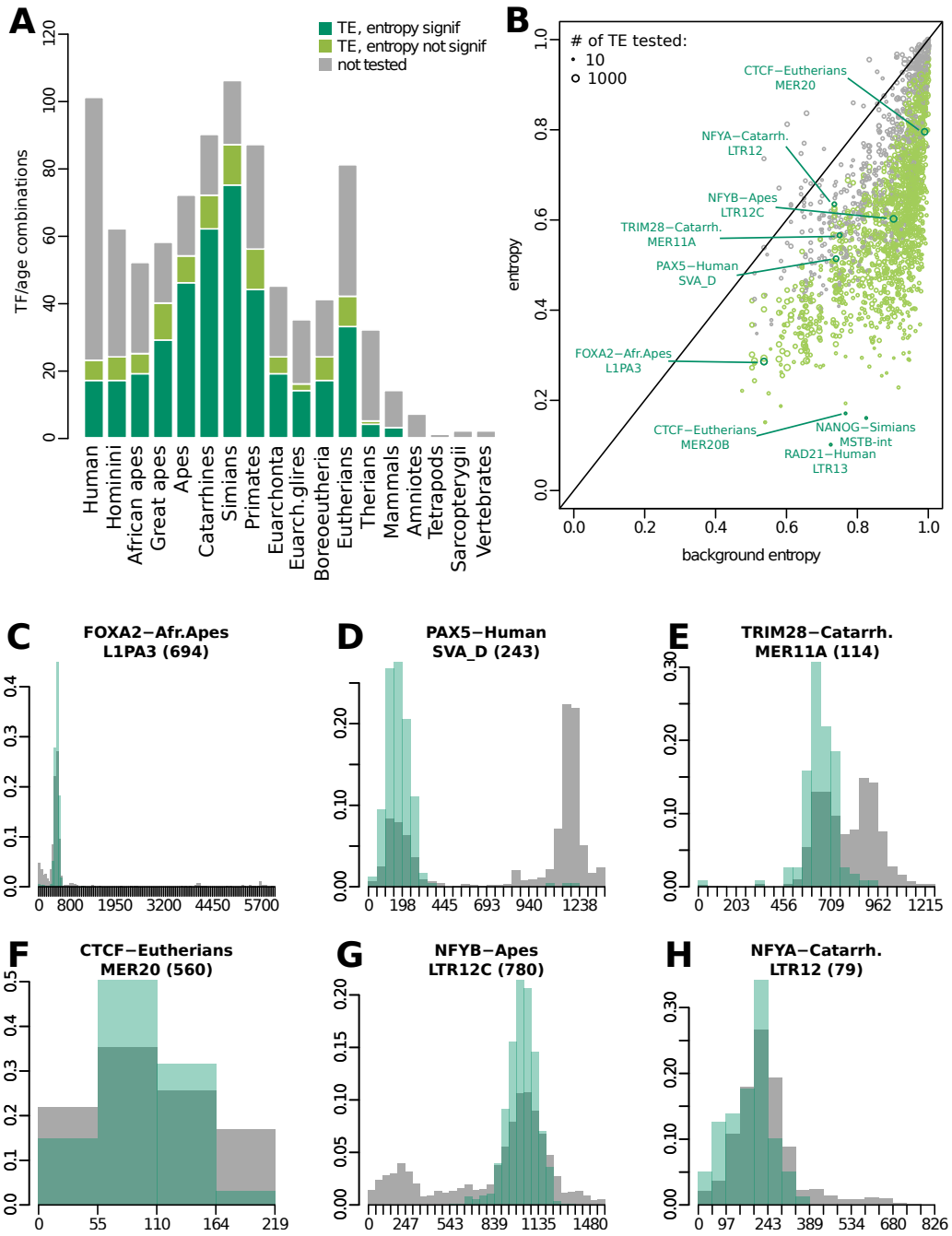

FIGURE 2.5: Enriched TFBS that have a constant position within Repeated elements. *(A) Cases with significantly lower entropy in the position distribution with respect to the background. Each bar represents the number of TF/age pairs that resulted significant (dark green), not significant (light green), not tested (grey), for each age. (B) Entropies for each tested TF/TE/age triplet: on the vertical axis is reported the entropy of the position distribution of BS for the enriched TF within the corresponding TE, whereas on the horizontal axis the entropy of the background, that is the same distribution but considering all TF binding to the TE. Point size and colour represent respectively number of tested instances and test result, with green being significant. (C-H) Significant examples: green bars represent the position distribution of BS considering only the enriched TF, grey bars represent the background. In particular (G) is the case with most instances and (H) is being considered a borderline case with FDR 0.05.*

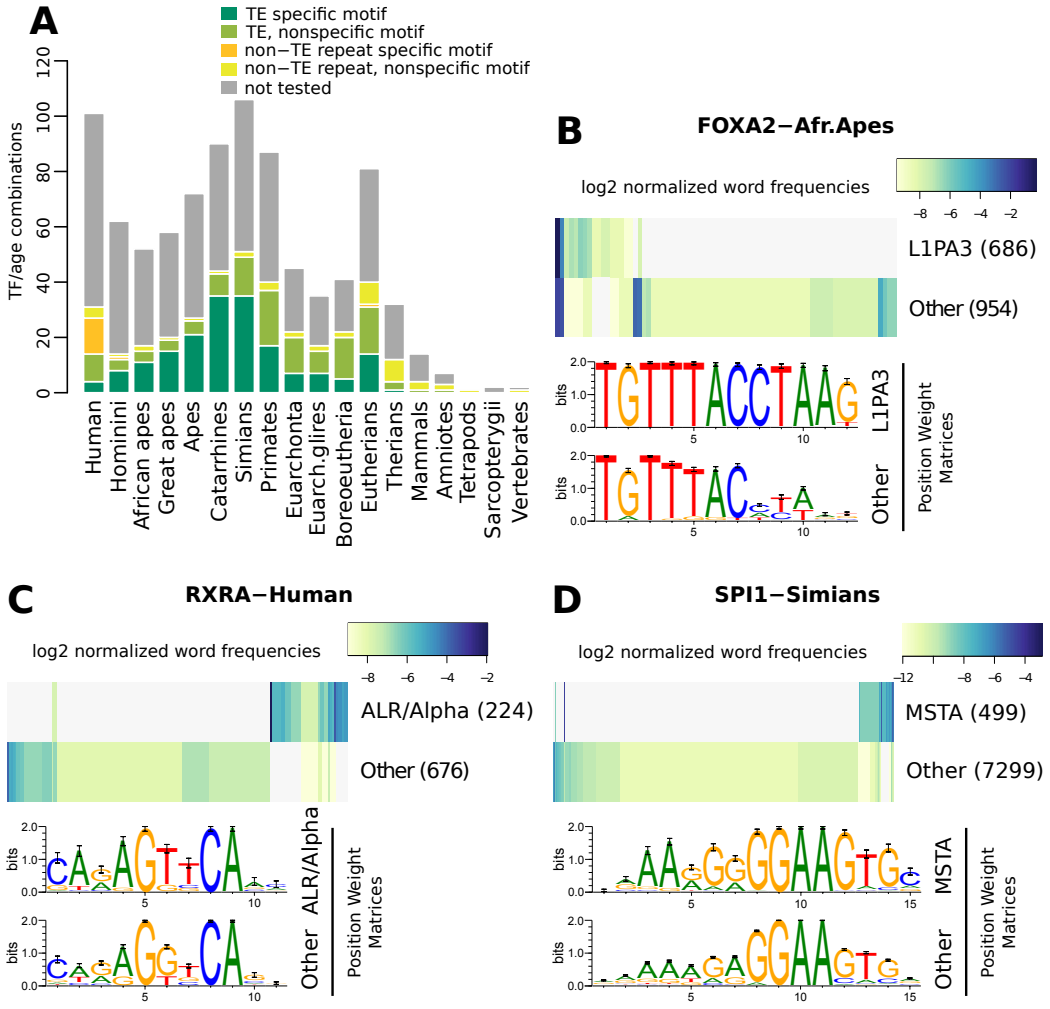

FIGURE 2.6: TFBS with RE-specific motif-word distribution. *(A) Cases with word occurrences in the enriched RE class significantly different than elsewhere. Each bar represents for each age the number of TF/RE pairs that resulted significant (dark colors), not significant (light colors), not tested (grey), separating for transposon (green) and non-transposon (yellow) repeats. (B-D) Significant examples: the word frequency matrix is represented by the heatmap (colors are in log2 scale, white means 0 occurrences); the Positional Weight Matrix logo obtained from binding site motifs occurring on the tested RE class is compared with the logo from all other binding sites.*

result holds when considering tissue-specific expression shifts, and ChIP-seq results limited to relevant cell lines, at least when the number of shifted genes is enough to provide reasonable statistical power, as in brain and liver (see Supp. Fig. A.10).

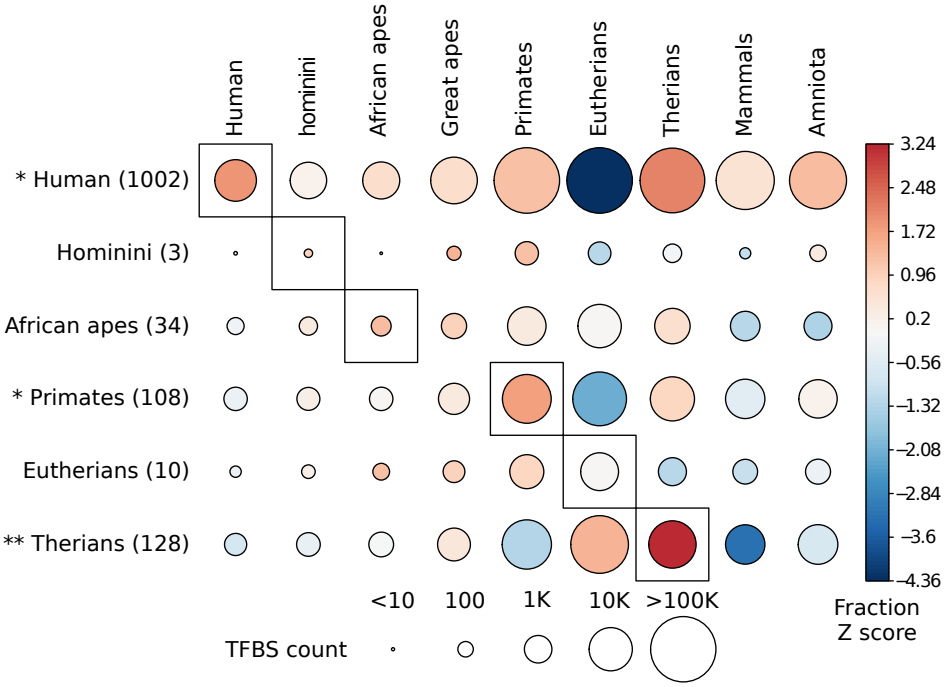

FIGURE 2.7: Age distribution of TFBSs in the regulatory region of genes which underwent an expression shift in a human ancestor. *The heatmap represents, for each ancestor in which the shift occurred, the fraction of TFBS with a certain age transformed into z-scores. The size of the circle represents the number of TFBS in each cell. Next to each ancestor name is annotated in parentheses the number of genes whose expression shifted. Asterisks indicate significant enrichment of TFBSs of the same age as the expression shift (*: P<0.05; **: P<0.01). z-scores and enrichment P-values were based on 5,000 randomizations of the age of the expression shift.*

2.3.5 Transcription factors that underwent binding site expansion

By looking at Figure 2.3 we can observe several examples of transcription factors which gained targets through RE expansion. A sizable cluster of TFs preferentially bind Hominini-specific regions (that is, originating in the human-chimpanzee common ancestor). These include immune-related TFs such as *IRF4* [122] [MIM 601900], *PAX5* [123] [MIM 167414] and *POU2F2* [124] [MIM 164176] and TFs involved in the regulation of metabolism, such as *RXRA* [125, 126] [MIM 180245] (see Figure 2.6 C) and *USF1* [127] [MIM 191523]. A few of these TFs gained new targets through TE expansion, including *PAX5* (see Figure 2.5 D), *FOXA2* [MIM 600288] (see Figures 2.5 C and 2.6 B), *SPI1* [MIM 165170] (see Figure 2.6 D), *HEY1* [MIM 602953], *GATA2* [MIM 137295].

The expansion of *TRIM28* (aka *KAP1*) [MIM 601742] binding sites during primate evolution is also evident: its newly formed binding sites in the ur-Simiformes, ur-Catarrhine and ur-Hominoidea are preferentially found within TEs of various classes (especially LTR) in agreement with its role in repressing newly arising TEs [128], see

also Figure 2.5 E. Among the TFs showing a marked target expansion in the ur-Eutherian we find *CTCF* and *MAFF* [MIM 604877], which is involved in parturition by regulating the oxytocin receptor [129]. In agreement with *Schmidt et al.* [85], *CTCF* shows both ancestral binding sites predating the origin of vertebrates and several waves of expansion in mammals. Our results confirm in particular a role of MER20 and MER20B in the expansion of *CTCF* sites in the ur-Eutherian, as previously reported [130], see Figure 2.5 F.

To gain some insight in the possible functional impact of TFBS creation by genomic expansion we examined in more detail the cases of *FOXP2* [MIM 605317], a TF involved in speech and language which underwent significant mutations in the human lineage after the split with the chimpanzee [131], and which shows a large number of human-specific sites inside ALR/Alpha and HSATII satellite repeats; and *FOXA2*, a TF involved mostly in regulating gene expression in liver, that provides one of the clearest examples of regulome expansion through transposable elements, specifically L1PA3 (Figure 2.5 C). In particular we evaluated, using GREAT [109], possible functional enrichment of the putative gene targets created by the repeated element (repeat-driven targets) compared to all putative targets.

Very few genes were associated by GREAT to the *FOXP2* binding sites, and these genes did not show any functional characterization, suggesting that these new binding sites might have very limited functional effect. On the other hand many repeat-driven targets could be identified for *FOXA2*: compared to all *FOXA2* target genes, these show a strong enrichment in neural function, represented by GO terms such as "serotonin receptor" activity (fold enrichment 5.3, FDR $\sim 4 \cdot 10^{-4}$) and Disease Ontology terms including "Intellectual disability" (fold enrichment 2.1, FDR $\sim 4 \cdot 10^{-4}$). While *FOXA2* is mostly known as a liver-specific TF, it is known to have a role in neuron differentiation [132, 133]. These results suggest that L1PA3 expansion could have contributed to the rewiring of the *FOXA2* regulome specifically in the brain. Analysis of the expression pattern of the repeat-driven targets in human tissues shows that brain tissues are indeed the ones where the expression ratio between L1PA3-driven and other *FOXA2* targets is the highest (see Supp. Fig. A.11).

2.4 Discussion and conclusions

We classified the human genomic sequence based on its evolutionary age, that is the ancestor in which the sequence first appeared. We then examined the age of regulatory sequences, defined as transcription factor binding sites determined by ChIP-seq experiments. Many transcription factors appear to have acquired new binding sites through genomic expansion, a fact that was known for some of them [85] but that we could establish in a systematic way.

Transposable elements play a crucial role in generating these waves of regulatory expansions, and in many cases specific families of them can be associated to specific waves of expansion, especially when these are relatively recent so that the originating TE can still be recognized. However our approach does not rely on databases of TE sequence, and thus is able to identify ancient waves of expansion such as the ones involving several transcription factors in the ur-Therian.

Several features of the TFBS located in repeated elements suggest they were already present at the time of genomic insertion: binding sites of specific TFs appear in

fixed positions inside each class of repeated elements and use a distinctive set of motif-words, a pattern difficult to reconcile with a process in which binding sites are formed by the gradual accumulation of point mutations.

Between the exaptation of pre-existing sequence and the immediate recruitment of new sequence for regulatory rewiring, various intermediate scenarios are possible, including for example new sequence carrying quasi-binding sites needing some point mutations to become effective, or binding sites that are not immediately effective because they reside within closed chromatin. Our age-enrichment map (Figure 2.3) undoubtedly reflects also some of these intermediate cases.

While our results suggest that most transcription factors obtained a relevant part of their binding sites during specific waves of genomic expansion, they do not imply that *most* binding sites are generated in this way. For example, only about 4.9% of all the TFBSs used to build Figure 2.3 contribute to the enrichment of a TF/RE/age triplet, and can thus be specifically attributed to the expansion of the RE in a specific evolutionary age. This must be considered as a lower bound since our strict control of false positives in evaluating enrichment certainly leads to many false negatives. However, this relatively low percentage shows that our results are not in contradiction with those of *Villar et al.* [87] where it is shown that most regulatory elements are created by exaptation of existing sequence: while this is the dominant mechanism, most human transcription factors have also undergone waves of rapid target expansion driven by newly acquired genomic sequence.

It is also worth stressing that the binding of transcription factors to DNA is by no means always functional (see e.g. [104]), so that we do not expect all the TFBSs generated through genomic expansion to significantly alter the transcriptome. For example the lack of functional characterization of the repeat-driven *FOXP2* targets might suggest that these binding sites have little if any effect on the regulatory network. On the other hand the results on *FOXA2* and, more generally, the significant concordance between age of regulatory elements and age of gene expression shifts suggest that the effect of genomic expansion on the human transcriptome is real and measurable.

Chapter 3

Investigating Adaptation In Modern Humans Through Haplotype Visualization

3.1 Introduction

A haplotype is an arrangement of specific alleles occurring in the same chromosome within a given genetic segment. Often genetic variation is studied through summaries of single nucleotide polymorphisms (e.g. frequency of mutations). However, haplotypes provide more information because we can see the combination of alleles that are present on a single chromosome. Access to tools to examine the complete patterns of genetic variation in these regions, and not just summary statistics, will help further elucidate the underlying evolutionary processes.

To our knowledge, the first tool to fulfill this need is *inPHAP* [134]. It features a graphical interface to show sequences of alleles and to aggregate them interactively and according to meta information. Although it is useful for basic haplotype visualization and analysis, it requires manually supplying groupings to observe summaries, and does not provide allele polarization, dataset merging, and a command-line environment. Therefore, *inPHAP* is less suitable for automated hypothesis generation in population genetics.

Other methods to visualize haplotype structure provide insights on the abundance of several haplotypes in a population [135], or on the linkage disequilibrium between variants [136]. However, these methods generate a data summary, and do not show the full sequence of variants directly. Here we present a tool to assist researchers in visualizing the polymorphisms of a given region of the genome. In particular, this tool provides the user a few options to reveal hidden haplotype structure that may not be apparent when the haplotypes are plotted in a random order. Therefore, beyond being a visualization software, *Haplostrips* can be used to gain information about the evolutionary processes responsible for the observed haplotype patterns (e.g. positive selection, introgression etc.).

We will then briefly showcase how *Haplostrips* has been applied in two different projects: the first, *Antelope et al.* [55], features LCT as quintessential example of positive selection and the second, *Racimo et al.* [19] is one of the first investigations into the joint dynamics of archaic introgression and positive selection.

Adaptive Selection: the evolution of Lactase Persistence

One of the key features of mammals is the ability to produce and digest milk, in order to feed newborn offspring. The need of breaking down lactose, the only carbohydrate found in milk, is fulfilled by the lactase enzyme (LCT), whose expression level starts declining after weaning and is then very low in all adult mammals [137], with the notable exception of humans. Instead, in approximately a third of humans [138], the expression of lactase persists throughout life, a phenotype known as lactase persistence (LP). The frequency of LP varies greatly among populations, ranging from 5% to almost 100%, with the highest frequencies found in people of northern European descent and some populations from West Africa, East Africa, and the Middle East [137, 138]. In 2002, a study of Finnish families by *Enattah et al.* [139] identified the first mutation associated with the LP phenotype: *-13.910:C>T* (*rs4988235*), located in a intron of *MCM6*, a gene immediately upstream of *LCT*. Soon after, *Bersaglieri et al.* [140] showed that haplotypes carrying the *-13.910:T* variant present typical characteristics of recent and local positive selection in Europeans. The time of the onset of the LP allele has been evaluated with different methods [141],[140] [142], also integrating data from ancient genomes in Europe [71, 72]. Despite huge confidence intervals and minor differences, the consensus is that an appreciable frequency of this lactase persistence allele in Europe only dates to the last 4,000 years, and possibly later. By observing the haplotype pattern with *haplostrips*, measuring the haplotype shared tract length and the F_{ST} between North Western European and Han Chinese populations, we draw attention to a background haplotype, previously observed by *Bersaglieri et al.* [140] and *Enattah et al.* [143], the length of which was never analyzed thoroughly. This observation triggered further study, published in *Antelope et al.* [55], which showed that the observed haplotype structure among the NW European and Han populations cannot be explained by selection on *de novo* mutation or standing variation under current demographic models.

Archaic Adaptive Introgression in Present-Day Human Populations

There is now a large body of evidence supporting the idea that certain modern human populations admixed with archaic groups of humans after expanding out of Africa. In particular, non-African populations have $1-2\%$ Neanderthal ancestry [10, 15], while Melanesians and East Asians have 3% and 0.2% ancestry, respectively, from Denisovans [10, 16, 17]. Recently, it has become possible to identify the fragments of the human genome that were introgressed and survive in present-day individuals [10, 144, 145]. Researchers have also detected which of these introgressed regions are present at high frequencies in certain present-day non-African populations. Some of these regions are likely to have undergone positive selection in those populations after they were introgressed, a phenomenon known as adaptive introgression (AI). One particularly striking example of AI is the gene *EPAS1* [146] which confers a selective advantage in Tibetans by making them less prone to hypoxia at high altitudes [64, 147, 148]. The selected Tibetan haplotype is likely to have been introduced in the human gene pool by Denisovans or a population closely related to them [18].

In this study, we first use simulations to assess the power to detect AI using different exploratory summary statistics that do not require the introgressed fragments to be

identified *a priori*: U and $Q95$. They rely on two signatures that proved to be the hallmark of AI, and that were observed in *EPAS1*. First, in a region under AI, one would expect the sequence divergence between an individual from the source population and an admixed individual to be smaller than the sequence divergence between an individual from the source population and a non-admixed individual. Second, we would expect a large number of sites containing archaic alleles at high frequency in the admixed population, but absent or at low frequency in a non-admixed population., We then apply these statistics to real human genomic data from phase 3 of the 1000 Genomes Project [53], to detect AI in human populations, and find candidate genes. While these statistics are sensitive to adaptive introgression, they may also be sensitive to other phenomena that generate genomic patterns similar to those generated by AI, like ancestral population structure and incomplete lineage sorting. These processes, however, should not generate long regions of the genome where haplotypes from the source and the recipient population are highly similar. As additional confirmation that the candidates we found with our statistics are generated by AI, we explored the haplotype structure of some of the most promising candidates with *Haplostrips*. Finally, to have a grasp on how much uniquely shared archaic alleles at high frequencies in non-Africans were affecting functional regions of the genome, we tested their enrichment in genic versus intergenic human DNA.

3.2 Methods

3.2.1 Haplostrips: Revealing Population Structure Through Haplotype Visualization

Haplostrips handles variation data, selecting the window of interest, extracting the haplotype data from the phased genotypes, polarizing variant sites and filtering them for mapping and genotype qualities. It keeps the samples belonging to populations of interest and chooses only the most informative sites, eliminating variations with very low frequency in all the populations to be plotted. Finally it produces a heatmap plot that displays the haplotypes in rows while each column represents a SNP within a region of interest. Haplotypes are labeled with a color defined by metadata, e.g. populations, from a file supplied by the user. Derived alleles are represented as black spots and ancestral alleles are represented as white spots (see Figure 3.1).

A key feature of *Haplostrips* is being able to sort and cluster haplotypes using only the distance between the genetic sequences, regardless of the meta information supplied. This turns the disorganized heatmap, of which an example is represented in Figure 3.1 A, into an informative plot that reveals hidden haplotype structures, as seen in Figure 3.1 B.

*Haplostrips*is a command-line tool written in Python and R. It takes advantage of the preexisting Python package *Pandas* [149] and the R package *gplots* to manage input data and draw the plot, respectively.

The software is downloadable at *https://bitbucket.org/dmarnetto/haplostrips* .

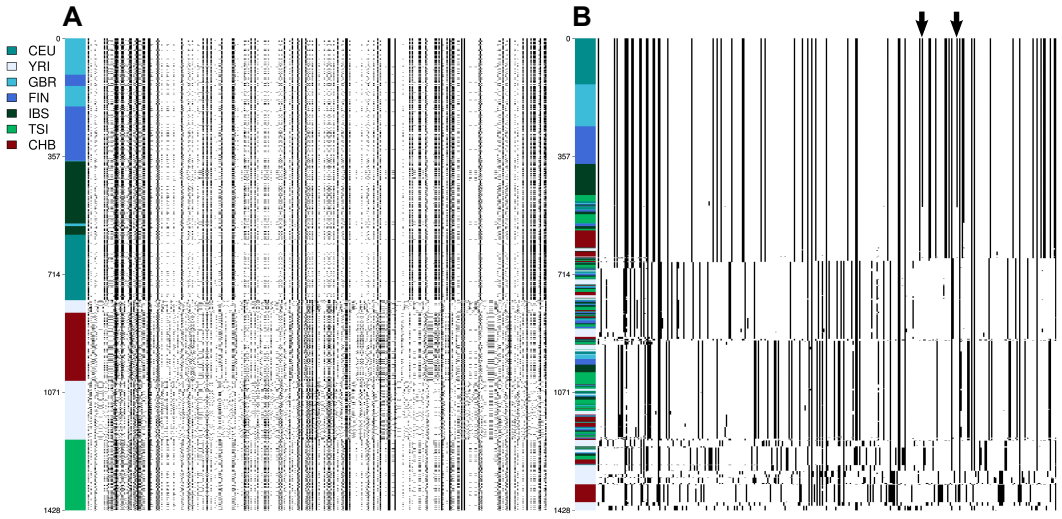

FIGURE 3.1: ***Haplostrips plot of LCT and MCM6.*** *(A) unsorted haplotypes, (B) haplotypes clustered and sorted by increasing distance with CEU consensus. Rows are haplotypes and columns are variable sites. Black denotes derived alleles, and white denotes ancestral alleles. The arrows indicate the position of rs4988235 and rs182549, associated with lactase persistence. This region spans 88-kb encompassing LCT and MCM6 in their entirety. CEU = Utah residents with northwestern European ancestry, GBR = British, FIN = Finnish, TSI = Toscani in Italia, IBS= Iberian in Spain, CHB = Han Chinese in Beijing, YRI = Yoruba in Ibadan, Nigeria.*

Input

The user can supply as input a VCF genotype file, similar to those produced by the 1000 genome project [53]. This format has become a standard for genetic variation data, and this makes our tool portable, versatile and simple to use. In addition, where Tabix indexes are available, *Haplostrips* uses the *pysam* package [150] to perform a fast retrieval of the region of interest. More VCF files can be supplied to the tool, which is capable of merging them using the reference allele of variants present in one VCF to infer missing data in others, or simply working on the intersection. *Haplostrips* can run iteratively over many windows of interest supplied with a file that contains the genomic intervals and populations to be plotted: this can be useful to visualize windows resulting from genome wide scans, e.g. GWAS. Another accepted input is the format generated by *ms* [151], a widely used software to generate samples under a variety of neutral models. This feature allows one to observe the direct effects of particular demographic histories without any further parsing or coding.

Clustering, sorting and other options

The clustering is optional and is performed hierarchically via the single agglomerative method based on Manhattan distances, using the *stats* library in R. The Manhattan distance is simply the number of SNPs with different alleles in two sequences. The clustering brings together similar haplotypes, generating a thicker row in the plot for those that are more abundant, and a thicker row in the label column for the populations where they are more represented. The resulting dendrogram, which can be visualized as well, is re-ordered by decreasing similarity with a reference

haplotype or a consensus sequence of a defined population. The reordering is performed using the minimum distance method, to ensure that the closest haplotype to the reference is always shown at the top. Available ordering options also include (1) performing the clustering after the population grouping, (2) sorting the haplotypes for increasing differences from the reference one or (3) keeping the input order. The first option can be used to investigate population specific Linkage Disequilibrium or other effects, while the second one allows the comparison the heatmap with a plot showing the increasing number of differences to the reference haplotype, also reported in a separate file. This method lets the user deal with a simpler quantity, avoiding the clustering step. The heatmap and the distances to the reference haplotype can be optionally output to tab-delimited files.

The user can define the populations or groups of interest to be plotted, which are associated to the samples by another input file. Alleles at variant sites can be polarized for ancestral/derived status, using the ancestral allele provided in the INFO field of the VCF file. Knowing the ancestral or derived state of the allele is important for understanding the time of arrival of the mutation in the human lineage, and informs what the correct evolutionary models need to be applied in analysing a dataset. Sites can be filtered for genotype and mapping qualities or for having a low intra-population minor allele frequency in all populations plotted, as cited above. This last filter is particularly important because usually only a small portion of the polymorphic sites is informative, while most of them have frequency so low that would result in nearly uniform columns (white or black) in the plot. As an example only 344 out of 2463 polymorphic sites were plotted in Figure 3.1, while all sites with a maximum within-population MAF below 0.05 were removed.

Choice of the region to be plotted

It is worthwhile to point out that *Haplostrips* is useful for inspecting local patterns. Selecting a region that is too long can make the interpretation and haplotype clustering difficult, to the point where the plot loses its meaning. Consequently *Haplostrips* is not optimized for large regions and the RAM usage is dependent on their dimension. Windows of around 1000 sites before the filtering steps tend to provide good resolution, though this depends on the nature of the region that will be plotted and on the SNP density of your dataset.

3.2.2 Calculating the Length of the Shared Track of Homozygosity

We defined the pairwise shared track of homozygosity length around a position x (hereafter "track length" around x) as the sum of the maximum number of base pairs to the left and right of position x until the two chromosomes differ by a base pair. We filtered out all SNPs under 5% frequency and calculated the track length by lexicographically sorting the chromosomes from the edge up to the base pair next to site x and calculating the shared length between every adjacent pair [152]. For the LCT locus, we first partitioned all sampled chromosomes into three types: the ones carrying the derived allele in NW Europeans, the ones carrying the ancestral allele in NW Europeans and the Han haplotypes, which all carry the ancestral allele. We then calculated the within-type shared track length and the between-type shared track length.

3.2.3 Summary Statistics to detect Adaptive Introgression

According to [19], for a window of arbitrary size, let $U_{A,B,C}(w,x,y)$ be defined as the number of sites where a sample C (the "bait") from an archaic source population (which could be as small as a single diploid individual) has a particular allele at frequency y, and that allele is at a frequency smaller than w in a sample A (the "outgroup") of a population but larger than x in a sample B (the "target") of another population. In other words, we are looking for sites that contain alleles shared between an archaic human genome and a test population, but absent or at very low frequencies in an outgroup (usually non-admixed) population. For example, suppose we are looking for Neanderthal adaptive introgression in the Han Chinese (CHB). In that case, we can consider CHB as our target panel, and use Africans as the outgroup panel and a single Neanderthal genome as the bait. If $U_{AFR,CHB,Nea}(1\%, 20\%, 100\%) = 4$ in a window of the genome, that means there are 4 sites that are shared at more than 20% frequency in Han Chinese and at 100% Neanderthal, but at less than 1% in Africans.

Another statistic introduced in [19] is $Q95_{A,B,C}(w,y)$, and is defined for a window of arbitrary size, as the 95^{th} percentile of derived frequencies in an admixed sample B of all SNPs in that window that have a derived allele frequency y in the archaic sample C, but where the derived allele is at a frequency smaller than w in a sample A of a non-admixed population. For example, $Q95_{AFR,CHB,Nea}(1\%, 100\%) = 0.65$ means that if one computes the 95% quantile of all the Han Chinese derived allele frequencies of SNPs where the Neanderthal genome is homozygous derived and the derived allele has frequency smaller than 1% in Africans, that quantile will be equal to 0.65. In other words, it is a summary of the allele frequency spectrum in the introgressed population, conditional on only looking at alleles uniquely shared with the source population and at low frequency in the non-admixed population.

If we have samples from two different archaic populations (for example, a Neanderthal genome and a Denisova genome), we can define $U_{A,B,C,D}(w,x,y,z)$ and $Q95_{A,B,C,D}(w,y,z)$. In this way we can filter for sites where the archaic sample C has a particular allele at frequency y and the archaic sample D has that allele at frequency z. For example we could set $y = 100\%$ and $z = 0\%$ to find alleles uniquely shared with Neanderthal, but not Denisova. If we were interested in archaic alleles shared with both Neanderthal and Denisova, we could set $y = 100\%$ and $z = 100\%$.

In [19] we use simulations to assess the power of these summary statistics to detect AI.

3.2.4 Testing for enrichment in genic regions

We used two different Linkage Disequilibrium pruning methods. In one (called "LD-1"), we downloaded the approximately independent European LD blocks published in ref. [102]. For each set of high frequency derived sites, we randomly sampled one SNP from each block. In a different approach (called "LD-2"), for each set of high frequency derived sites, we subsampled SNPs such that each SNP was at least 200 kb apart from each other. We then tested these two types of LD-pruned SNP sets against 1,000 SNP sets of equal length obtained permuting allele frequencies, pruning for LD and collecting SNPs in the same ways as described above.

3.3 Adaptive Selection: the evolution of Lactase Persistence

We applied *Haplostrips* to the locus responsible for lactase persistence in Europeans, that includes the genes *LCT* and *MCM6*. Data are from 1000 Genomes phase 3 [53]. Figure 3.1 B shows a sequence that encompasses both genes: this plot allows us to visually distinguish a haplotype at high frequency in all Europeans populations, but at very low frequency in Africans. Interestingly, the Italian Toscani population possess a different and more variable set of haplotypes, consistent with a higher incidence of lactose intolerance in this population. We can observe that Han Chinese carry an almost exact copy of the North Western European haplotype at moderate frequencies. However 2 out of 3 sites where they differ are *rs4988235* and *rs182549*, which have previously been associated with lactose intolerance [139] and are featured only by Europeans. We thus set out to analyze more thoroughly the differences between North Western Europeans and Han Chinese.

3.3.1 NW Europeans and Han Haplotype Pattern, Shared Length and Diversity

If we inspect the haplotype patterns restricting to NW Europeans and Han, as in Figure 3.2 A, we can isolate 4 main families of haplotypes present in in both NW Europeans and Han, plus one exclusive of Han. The most common haplotype, is subdivided in two versions: one that carries the ancestral allele of *rs4988235* (LP ancestral), and one with the derived allele at the same site (LP derived). The latter is absent in Han but vastly predominant in Europe, in fact, the haplotype patterns in Europeans at this locus are often used as the canonical example of what happens under a selective sweep [58, 140]. However, this comparison shows that the Han also carry a haplotype with striking similarity to the NW European population at intermediate frequencies even though the Han haplotype is missing the two putatively beneficial mutations. In addition, we can observe that these haplotypes never recombine, in concordance with this being a region of positive selection [33, 36]. We therefore set out to find how long this similarity region is.

One way to summarize the observed haplotype similarity between NW European and Han individuals, and evaluate how long it continues, is by computing the haplotype shared tract length at this locus, a pairwise measure that is simply the number of shared base pairs from the site of interest until a mismatch is encountered (see Methods). Therefore, a “Derived-Derived” shared tract length is generated from comparing two chromosomes both with the derived allele at the site of interest (*rs4988235*, blue arrow in Figure 3.2). On the other hand, “Derived-Ancestral” is the shared tract length generated from comparing two chromosomes, one with the derived allele and the other with the ancestral allele at the site of interest. If on one hand the derived allele is present only on NW Europeans, we computed the shared tract length separately for "Han Ancestral" haplotypes and "NW European Ancestral" haplotypes. We found that most haplotypes with the derived allele are identical (see Figure 3.2 B) at least within 200 kb of distance and often more than 500 kb (this is the maximum given the 1 Mb size of the region considered). Indeed, "Derived-NW European Ancestral" shared no similarity,as shown in Fig. 3.2 C, whereas a proportion of the "Han Ancestral" haplotypes resulted similar to the putative selected haplotype in the NW Europeans up to 100 kb of distance (see Figure 3.2D).

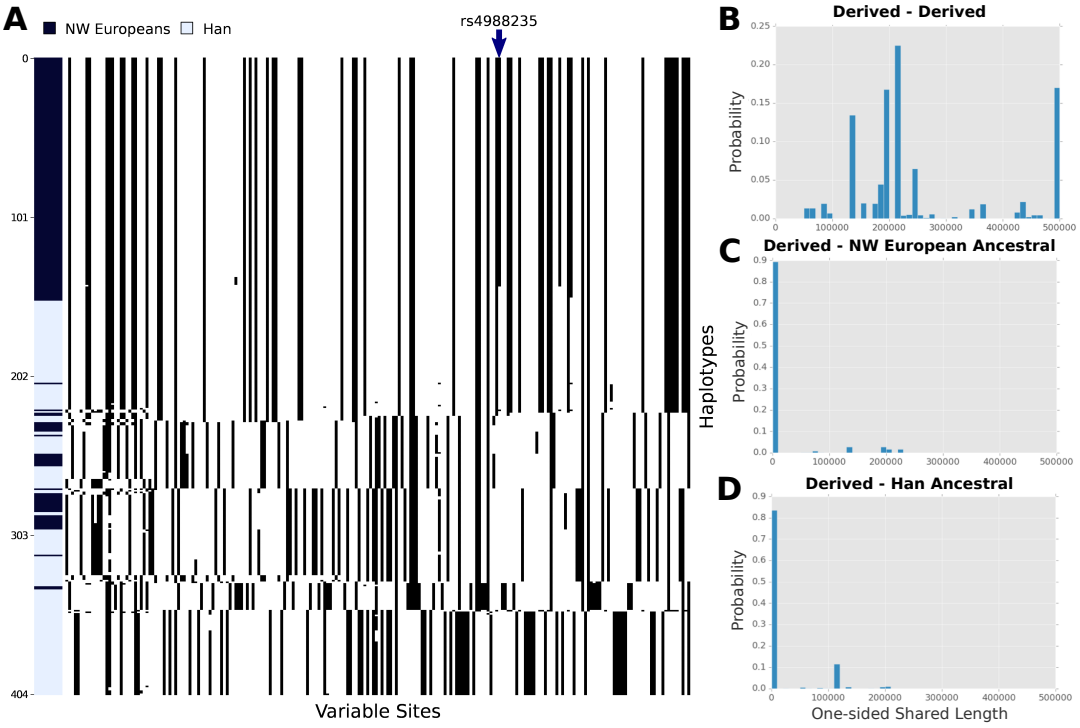

FIGURE 3.2: *(A)* ***Haplotype structure at LCT and MCM6.*** *Similar to Figure 3.1 but restricting to 103 Han individuals and 99 NW European individuals. The red arrow points to rs4988235, the site with the derived allele present at high frequency in NW European haplotypes and absent in Han haplotypes. Data are from 1000 Genomes phase 3. Sample sizes are 103 Han individuals and 99 NW European individuals. (B-D)* ***Haplotype-sharing comparisons.*** *The x-axis shows the one-sided shared length, with pairwise comparisons starting at the selected locus until 500 kb to the left or right of the locus. (Note that the length of sharing could be longer than 500 kb.) The y-axis indicates the proportion of pairs of haplotypes in a 10-kb bin of length sharing. "Han Ancestral" is a haplotype in the Han population which do not have the ancestral allele in the putatively selected position, "NW European Ancestral" is the NW European haplotype with ancestral allele in the selected position, and "Derived" is the NW European haplotype with the selected allele.*

Inspecting the patterns of genetic differentiation (as measured by F_{ST} between the NW European and Han populations) in a region of 1 Mb (see Figure 3.3) shows a region of size 200-400 kb around the putative selected site that exhibits many variants with constant level of genetic differentiation (F_{ST}) around 0.3 that is maintained for at least $\pm$ 100 kb (200 kb total length). Secondly, we can observe a small sequence of 100 kb withF_{ST} similar to *rs4988235*, therefore embedded within the NW European LP haplotype and, next to it, a low diversity region that extends until a position around 136.4 Mb.

These results, taken together, suggest that the haplotype background may have been an ancestral haplotype, i.e. preceding the Han-NW European split, and that after this split a mutation arose on that haplotype background in the NW European populations only. Previous haplotype studies of LCT have examined haplotype lengths and the relationship to the lactase persistence allele. [143] remarked on this observation of the similarity and frequency of selected and nonselected haplotypes (within a 30-kb region) and hypothesized that a Central Asian haplotype (which does not have the lactase persistence allele) may in fact be the haplotype background of the current European selected haplotype. [140] also summarized LCT haplotypes from

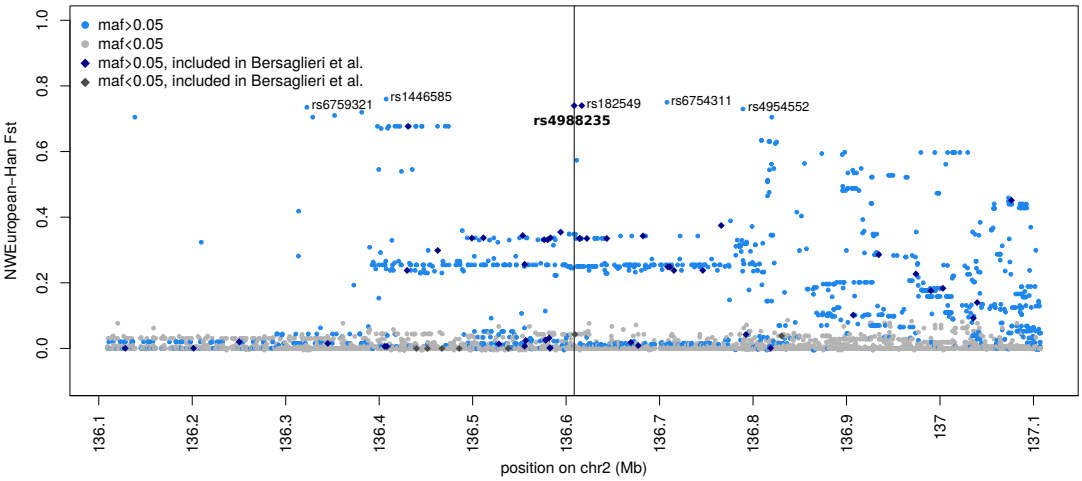

FIGURE 3.3: ***F_{ST} between NW European and Han populations.*** *F_{ST} for all SNPs in the region around rs4988235 (±500 kb; position of SNP denoted by black vertical line). Each SNP is a dot/point with x-coordinate given by the position on chromosome 2 and y-coordinate representing the F_{ST} computed according to Weir and Cockerham (1984). Blue SNPs have a minor allele frequency >0.05; gray SNPs have a minor allele frequency ≤0.05. Diamonds (in dark blue or black) represent the 101 SNPs included by Bersaglieri et al. [140].*

101 genotyped SNPs across a 3.2-Mb region, noting that the parental core haplotype is present in Asians, but did not make the observations described above, which challenges the underlying assumption that the extended high-frequency haplotypes in this locus are solely due to a selective sweep on the putatively selected lactase persistence allele.

To understand if the haplotype similarity between populations is to be expected under our current demographic out-of-Africa model [153], in [55] we simulated chromosomes of length 1 Mb under the reference demographic model and applied selection on the European branch consistent with the estimated selection parameters on LCT from other studies. The simulated statistic values were then compared with the real ones, finding that a standard model of demographic history coupled with strong selection parameters does not reflect the high level of shared similarity between the selected European haplotype and a subset of the Han haplotypes. Therefore, other scenarios for LP evolution are possible, for instance the LP Ancestral haplotype may already have had some positive effect with respect to lactase persistence or some other adaptive phenotype.

3.4 Candidates of Archaic Adaptive Introgression in Present-Day Human Populations

Bringing together similar haplotypes and ordering them with respect to a reference has been proven productive in previous work [18] where visualization of the data led to the observation that the haplotype at high frequency in Tibetans originated from another population, a conclusion that was not evident from statistical summaries of the data. Furthermore, *Haplostrips* has been applied in a recent project [19], to substantiate Denisovan and Neanderthal adaptive introgressions in modern human populations.

3.4.1 Genome wide identification of adaptively introgressed regions

To identify adaptively introgressed regions of the genome, we computed $U_{A,B,C,D}(w,x,y,z)$ and $Q95_{A,B,C,D}(w,y,z)$ (see Methods for details) in 40 kb non-overlapping windows along the genome, using the low-coverage sequencing data from phase 3 of the 1000 Genomes Project [53]. We used this window size because the mean length of introgressed haplotypes in ref. [10] was 44,078 bp. We conditioned on observing the archaic allele at less than 1% frequency in a non-admixed outgroup panel (A) composed of all the African panels (YRI, LWK, GWD, MSL, ESN), excluding African-Americans, and then looked for archaic alleles at high frequency in particular non-African populations (B). We used the high-coverage Altai Neanderthal genome [10] as bait panel C and the high-coverage Denisova genome [17] as bait panel D. We deployed these statistics in three ways: a) to look for Neanderthal-specific AI, we set $y = 100\%$ and $z = 0\%$; b) to look for Denisova-specific AI, we set $y = 0\%$ and $z = 100\%$; c) to look for AI matching both of the archaic genomes, we set $y = 100\%$ and $z = 100\%$. We can see in Figure 3.4 A the 40kb regions in the 99.9% highest quantiles of both the $U_{Afr,Pop,Nea,Den}(1\%, 20\%, y, z)$ and $Q95_{Afr,Pop,Nea,Den}(1\%, y, z)$ for different choices of target introgressed population (Pop).

3.4.2 Inspecting candidate loci

Below we analyze in detail three candidates extracted from the outliers in Figure 3.4 A , inspecting their haplotype patterns with the help of *Haplostrips*. The plots in Figure 3.4 B, C, D cover continental populations that show a large number of uniquely shared archaic alleles, and include YRI as a representative African population. The haplotypes are clustered and ordered by similarity to the closest archaic genome (Altai Neanderthal or Denisova).

As can be observed all these regions tend to show sharp distinctions between the putatively introgressed haplotypes and the non-introgressed ones. This is also evident when looking at the cumulative number of differences of each haplotype to the closest archaic haplotype, where we see a sharp rise in the number of differences, indicating strong differentiation between the two sets of haplotypes. Additionally, the YRI haplotypes tend to predominantly belong to the non-introgressed group, as expected.

The most extreme example according to the U statistic is a 120 kb region containing the *LARS* gene, with 76 uniquely shared Neanderthal alleles at $< 1\%$ frequency in Africans and $> 50\%$ frequency in Peruvians, which are also at $> 20\%$ frequency in Mexicans. *LARS* codes for a leucin-tRNA synthetase [154], and is associated with liver failure syndrome [155].

Another case with an extreme U statistic, is the one containing genes *OAS1* and *OAS3*, involved in innate immunity [156]. This region was previously identified as a candidate for AI from Neanderthals in non-Africans [157].

Looking at the haplotype patterns of these candidate loci in Figure 3.4 B, C, we can appreciate that the introgressed haplotype is covering less than half of the panel of individuals included, and is quite evenly distributed across populations of the same continent, except for *LARS* in Peruvians, where is slightly enriched (this might be related to the higher proportion of native american ancestry in Peruvians [53]). The distance with the Neanderthal haplotype is minimal, thus giving a very high

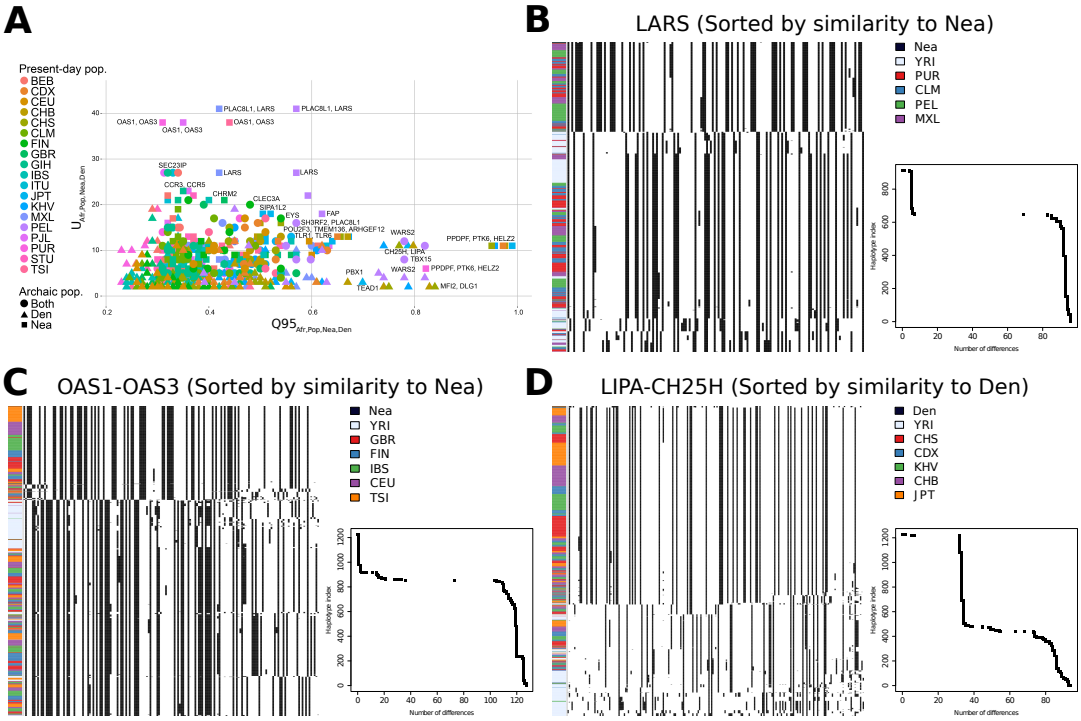

FIGURE 3.4: ***Candidates of Adaptive Introgression.*** *(A) 40kb regions in the* 99.9% *highest quantiles of both the* $Q95_{Out,Pop,Nea,Den}(1\%, y, z)$ *and* $U_{Out,Pop,Nea,Den}(1\%, x, y, z)$ *statistics for individual non-African populations within the 1000 Genomes data, using all African populations (excluding African-Americans) as the outgroup, and a cutoff* x *of* 20%*. We used three different symbols for Nea-only, Den-only, and Nea-Den-shared configurations of the statistics. (B-D) Haplotype structure of three candidate regions. For each region, we applied a clustering algorithm to the haplotypes of particular human populations and then ordered the clusters by decreasing similarity to the archaic human genome with the larger number of uniquely shared sites (see Section 3.2.1). We also plotted the number of differences to the archaic genome for each human haplotype and sorted them simply by decreasing similarity. In the latter case, no clustering was performed, so the rows in the cumulative difference plots do not necessarily correspond to the rows in the adjacent haplotype structure plots. LARS: chr5:145480001-145520000. OAS1-OAS3: chr12:113360001-113400000. LIPA-CH25H: chr10:90920001-90980000.*

number of uniquely shared alleles (hence a very high U), but the fact that only 30% of the individuals carry this haplotype causes $Q95$ to be less than extreme, resulting in these candidates being in the left side of the plot in 3.4 A.

Another interesting candidate region contains two genes involved in lipid metabolism: *LIPA* and *CH25H*. We find a 40 kb region with 11 uniquely shared Denisovan alleles that are at low ($< 1\%$) frequency in Africans and at very high ($> 50\%$) frequency in various South and East Asian populations (JPT, KHV, CHB, CHS, CDX and BEB). The $Q95$ statistic in this region is very high across all of these populations, and we also find this region to have extreme values of these statistics in a broader Eurasian scan, see [19]. This is also visible from 3.4 D, where the introgressed haplotype is covering the plot nearly to its entirety. In fact, $Q95$, being the 95th percentile of the SNP distribution in terms of frequency, is correlated to the frequency of this haplotype, whereas U uses the frequency as cutoff thus having a more ambiguous interaction with this quantity. Furthermore, with *Haplostrips* we can realize that the introgressed haplotype is quite far to the Denisovan one, thus reducing the number of shared alleles in this region. These two facts are encapsulated in a not particularly striking U statistic in this case. A possible cause of this distance between the bait and the

admixed population might be the fact that the bait used here (Denisova) was not the real source of the introgressed haplotype, but only related to it: the actual source could be another Denisovan subpopulation, as yet unsampled. Another hypothesis could be that the introgressed region started as neutral, accumulating some mutations before that this slightly mutated haplotype became adaptive.

The *LIPA* gene codes for a lipase [158] and is associated with cholesterol ester storage disease [159] and Wolman disease [160]. In turn, the *CH25H* gene codes for a membrane hydroxylase involved in the metabolism of cholesterol [161] and associated with Alzheimer's disease [162] and antiviral activity [163].

It is worth to note that this analysis identified the region containing genes *TBX15* and *WARS2* as a very strong candidate of adaptive introgression in whole Eurasia. This region has been associated with a variety of traits, (adipose tissue differentiation [164], body fat distribution [165], skeletal development [166],...) and was previously identified as positively selected in Greenlanders [167]: this finding triggered a separate study devoted to its analysis [168].

3.4.3 Testing for enrichment in genic regions

We aimed to test whether uniquely shared archaic alleles at high frequencies were enriched in genic regions of the genome. We looked at archaic alleles at high frequency in any of the Non-African panels that were also at low frequency ($< 1\%$) in Africans. As background, we used all archaic alleles that were at any frequency equal or larger than 1% in the same Non-African populations, and that were also at low frequency in Africans. We then tested whether the high-frequency archaic alleles tended to occur in genic regions more often than expected.

SNPs in introgressed blocks will tend to cluster together and have similar allele frequencies, which could cause a spurious enrichment signal. To correct for the fact that SNPs at similar allele frequencies will cluster together (as they will tend to co-occur in the same haplotypes), we performed linkage disequilibrium (LD) pruning using two processes, see Methods for details. Regardless of which LD method we used, we find no significant enrichment in genic regions for high-frequency ($> 50\%$) Neanderthal alleles (LD-1 P=0.706, LD-2 P=0.161) or Denisovan alleles (LD-1 P=0.348, LD-2 P=0.192). Similarly, we find no enrichment for medium-to-high-frequency ($> 20\%$) Neanderthal alleles (LD-1 P=0.553, LD-2 P=0.874) or Denisovan alleles (LD-1 P=0.838, LD-2 P=0.44).

3.5 Discussion

We presented a tool to visualize the polymorphisms of a given region of the genome in order to reveal the haplotype structure within or across populations. We followed briefly exposing two examples where the use of *Haplostrips* proved proficient.

The first was about one of the most celebrated examples of human adaptation: the evolution of lactase persistence. Many of the insights about of this case of adaptation come from considering the frequency of the putative selected SNP (*rs4988235*) that carries the derived allele [71, 72]. However, the inspection of the genetic variation around the putative beneficial alleles provided additional information regarding the

adaptive history of the region, even in the case of a highly studied example such as *LCT*. With *Haplostrips* we clearly restated that the LP haplotype in Europeans also exists in a version carrying the LP ancestral allele, a fact previously introduced [140, 143]. The presence at a relatively high frequency and the length in an Asian population of such a haplotype, that triggered the subsequent analyses shown in [55], caught our attention thanks to this visualization approach. Note that these observations would not be derivable from a haplotype network view [135], for example, as the SNPs contributing to the haplotypes are not presented, and one cannot distinguish clearly between the selected European haplotype and the very similar moderate frequency Han haplotype.

A second work here presented that involved the use of *Haplostrips* was one of the first investigations into the joint dynamics of archaic introgression and positive selection, which developed statistics that are informative of AI and found candidates in the human genome, [19]. Here we look at both the number and allele frequency of mutations that are uniquely shared between the introgressed and the archaic populations. Such mutations should be abundant and at high-frequencies in the introgressed population if AI occurred. In particular, two novel summaries of the data that capture this pattern quite well have been identified: the statistics $Q95$ and U. Identified candidates mostly include genes involved in lipid metabolism, pigmentation and innate immunity, as observed in previous studies [144]. Phenotypic changes in these systems may have allowed archaic humans to survive in Eurasia during the Pleistocene, and may have been passed on to present-day human populations during their expansion out of Africa. *Haplostrips* has been used to inspect selected candidates, of which we report 3 cases: *LARS*, *OAS1/OAS3* and *LIPA/CH25H*. Using this visualization we were able to verify that the statistics cited above were working as expected, identifying tracts at high frequency with strong similarity between archaic humans and modern populations, but absent from a modern outgroup. In addition, these plots also explained the values of these statistics in terms of haplotype structure, distinguishing which patterns could give a high U and relatively low $Q95$ or vice-versa.

On a functional genomics side, with a specific focus on gene regulation, the low spatial resolution of these phenomena and therefore analyses (40kb for each introgressed window, see Section 3.4.1), make difficult to investigate anything at a subgene level (average gene size in human: 10-15 kb; BioNumbers, see Web Resources). Therefore, we tested whether uniquely shared archaic alleles at high frequencies in non-Africans were significantly more likely to be found in genic regions, relative to all shared archaic alleles, but did not find a significant enrichment. Though this suggests archaic haplotypes subject to AI may not be preferentially found near or inside genes, it may also be a product of a lack of power, or of the fact that not all uniquely shared archaic alleles may be truly introgressed. Some of these alleles may be present due to incomplete lineage sorting, which could add noise to the test signal. However, in this study, we did not pursue this line of research further.

To conclude, *Haplostrips* can be used to conduct exploratory analyses, confirm hypotheses about candidate regions, or even substantiate findings in scientific publications. It can be applied in all living systems for which haploid or phased diploid genotype datasets are available to visualize complex effects of, among others: introgression, domestication, selection and demographic events. Although existing tools already address the task of visualizing haplotypes, *Haplostrips* includes the ability of

independent haplotype clustering and providing meaningful plots without sacrificing the basic information encoded in the genetic sequences.

Chapter 4

Conclusions

We started this thesis briefly reviewing the human evolutionary history. On one hand this gives us the possibility to introduce the reader to our evolutionary history, allowing everyone to start on common ground despite their diverse experience in this field. On the other hand allows us to emphasize that peculiar aspects of the evolution of our species introduce precise issues and that the study of other organisms might rise different questions and complications. We briefly introduced the modes and times through which the evolution shapes genomes, sometimes defining spheres of application of different approaches, sometimes underlining questions that are peculiar to the human evolution as we know it, e.g: rapid radiations in the great apes family, introgressions from archaic genomes...

Then we focused on the tools that can be used to investigate molecular evolution and variation, underlining methodological differences necessary to study different time scales. We mentioned the contribution that classical population genetics delivered to today's questions and the advantages that technological improvements granted e.g. in the study of intra-species variation and ancient genomics.

We presented two projects at both ends of the time span that we set out to review [48, 67]. Both focus on investigating evolution at a molecular level, trying to evaluate the effects that genetic changes might have on the human phenotype. But the most noticeable take-home message form the comparison of these works come from their impressive methodological differences, that come primarily because of the time-scale that they set out to investigate. If on one end we compare reference genomes, focusing on huge gaps that the alignment might reveal and completely ignoring single nucleotide variants, on the other end we are conscious that considering only reference genomes is a serious oversight, therefore we use variation datasets and investigate the traces that recombination did not already break. This comes at the huge cost of reducing our resolution, with implications in the study of regulatory evolution.

4.1 A Focus on the Regulation of Gene Expression

In our analysis presented in Chapter 2, as well in previous literature (see *Carroll* [75]) we can observe how changes at a non-coding level are of paramount relevance in explaining diversity throughout the whole vertebrate history. Nevertheless, issues like low conservation on one end and Linkage Disequilibrium on the other, need to be addressed properly.

In Chapter 2 we performed a systematic investigation of the role of genomic sequence expansion in rewiring regulatory networks, focusing on TFBS as our regulatory units. The same would have needed a specific approach if transferred to the time span studied in Chapter 3. Using the study of Structural Variations to inform us on the insertions/deletions in the human genome we can build a phylogeny to trace expansions, but would be much more difficult given the continuous admixtures among human populations, also we do not expect many highly functional genome expansion at high frequency, even less than the human-specific ones presented in [49] and the modern human ones in [169], thus reducing our power in finding genome wide signals. Furthermore, if we were looking for signals of selection at TFBS, using the methods in Chapter 3 we would have impacted into low resolution issues. The selected haplotype at the *LCT* locus in Europeans extends for over 1 Mb, and including Asians where supposedly no selection happened, the same haplotype extends up to 100 kb (see Chapter 3). Also introgressd haplotypes from archaic humans have a median lenght, 44 kb [10], well over the median gene length. Thus special methods need to be developed in order to analyze the same features at different evolutionary scales [103–105].

4.2 Concluding Remarks

The study of molecular evolution at different scales in time and space involves the use of different tools and approaches, which are not trivially transferable to study the same functional features, as in the case of gene regulation analysis. I focused on several aspects, inevitably going quickly over some others, without the expectation of being fully exhaustive. Conversely, with this thesis I tried to give a glimpse of how diverse the approaches in evolutionary biology could be, to explain why two working examples at both ends of the temporal and spatial scales could be so methodologically different.

Appendix A

Supplementary Material for "Evolutionary rewiring of human regulatory networks by waves of genome expansion"

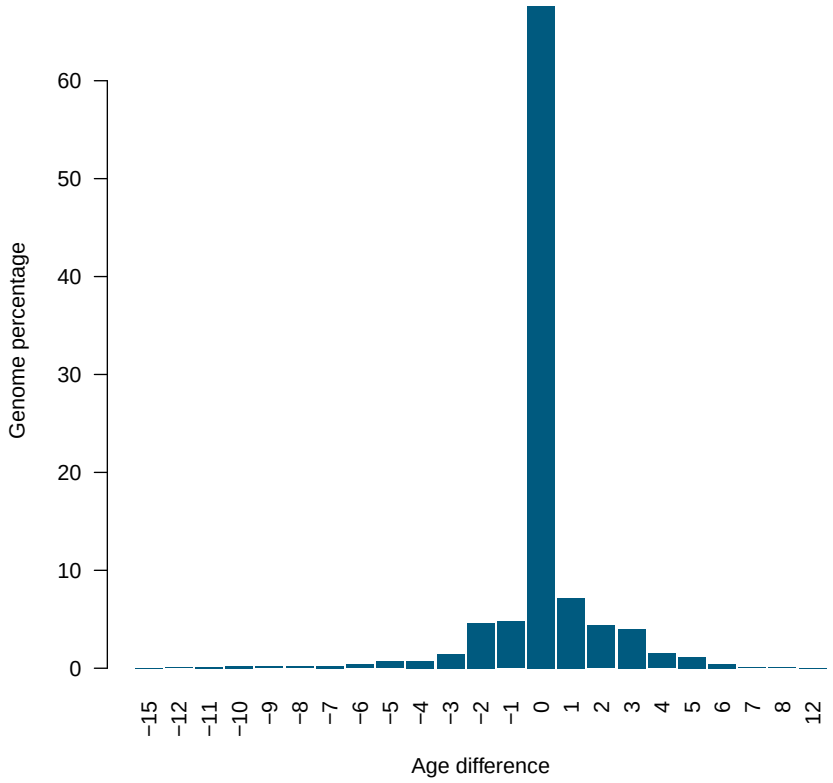

FIGURE A.1: Comparison of ages obtained with different aligments *Distribution of the difference between genomic ages evaluated starting with a multiple alignment of 100 vertebrate genomes or pairwise alignments between the human genome and 47 vertebrate genomes.*

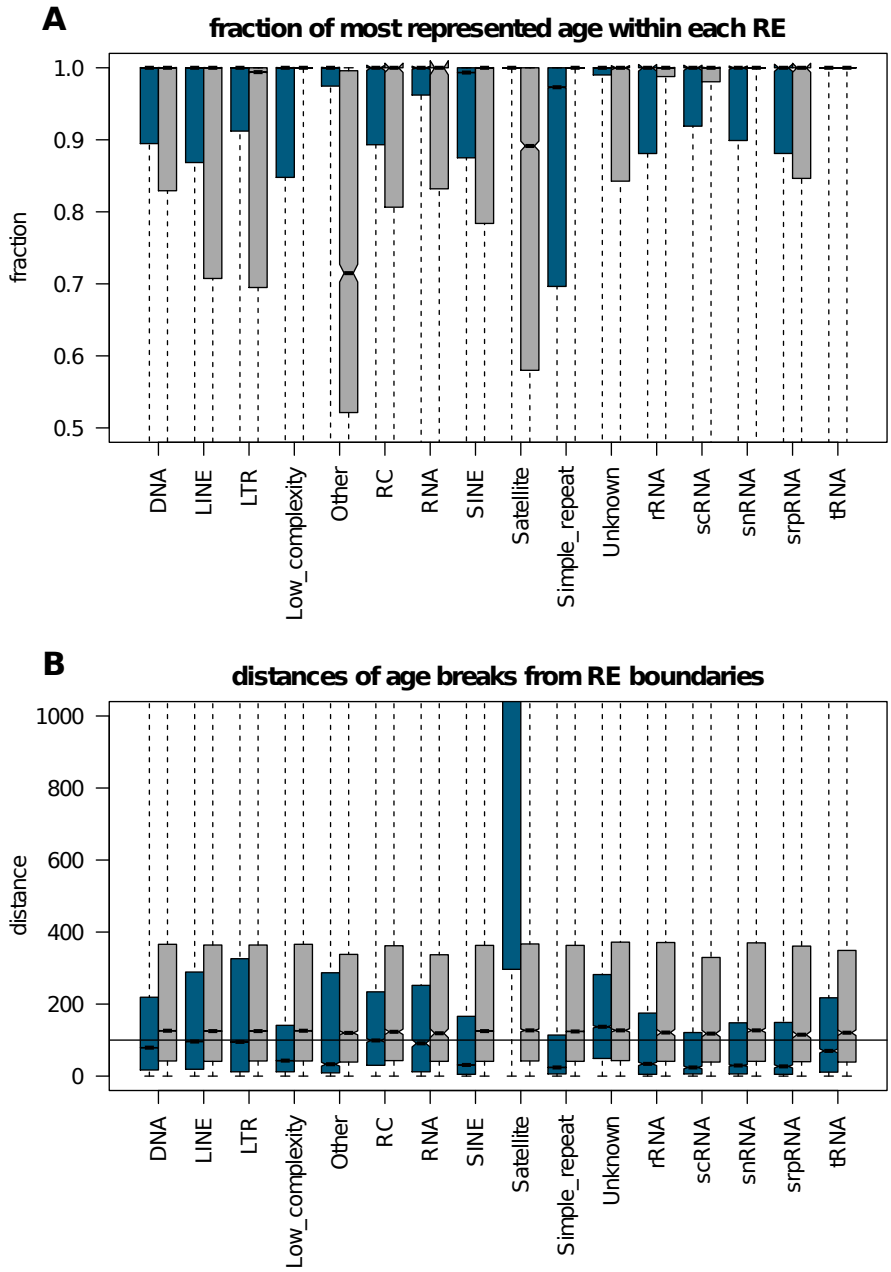

FIGURE A.2: Identity of Repetitive Elements and age blocks. *Repeats are grouped for class. Each repeat class (blue) is compared to a random set of regions with same size and length distribution (grey). (A) Distributions of fraction covered by the most represented age within each element: RE tend to have constant age. (B) Distributions of distances between repeats boundaries and age breaks. Except for Satellites and Unknown repeats, age breaks and RE borders tend to appear closer than 100 bp and with respect to their randomized counterpart.*

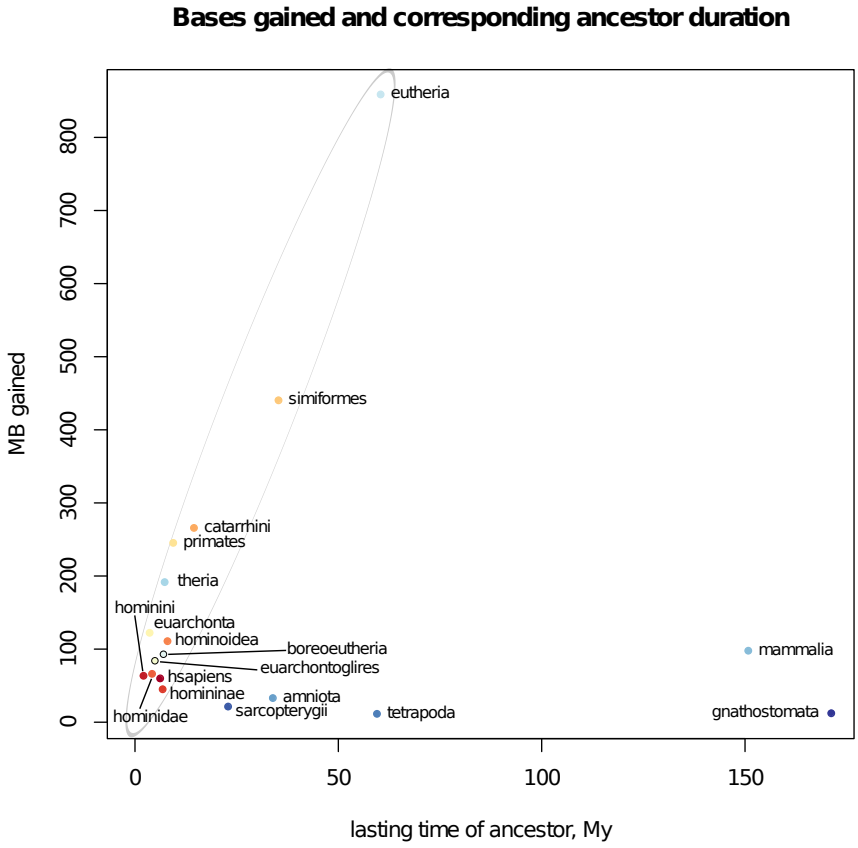

FIGURE A.3: Bases gained and corresponding ancestor duration. *Amount of human genomic sequence acquired at different time as a function of the time elapsed between successive speciation events. Elapsed times were obtained from the Timetree website (see Web Resources)*

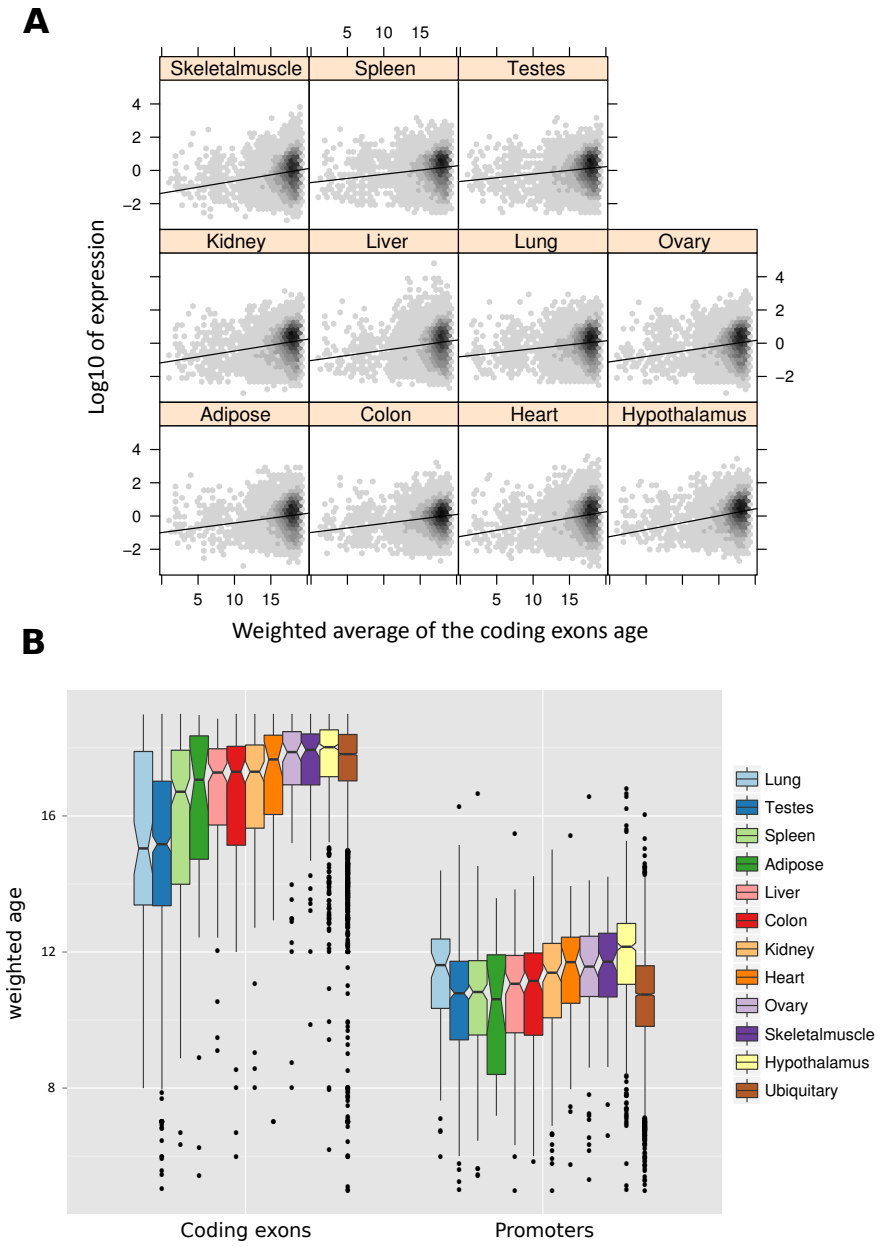

FIGURE A.4: Gene expression and gene age. *The age of a sequence is defined as the average age of the regions overlapping the sequence, weigthed by overlap length. (A) Older genes are more expressed: the panels show the dependence of gene expression in a collection of human tissues [16] from the age of their exonic sequence. (B) Ubiquitously expressed genes are older in their coding exons but younger in their promoters than tissue-specific genes.*

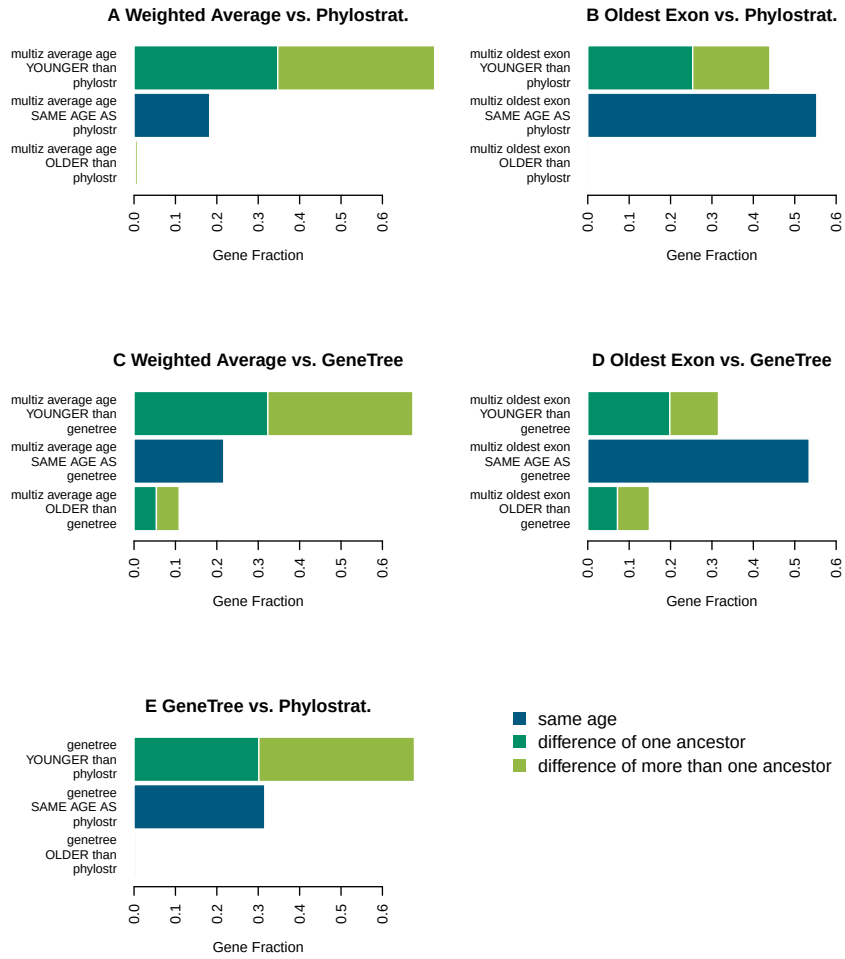

FIGURE A.5: Comparisons between gene age assigned by different dating methods. *Weighted Average is defined as the average age of the regions overlapping the gene, weigthed by overlap length, while Oldest Exon is defined as the age of the oldest sequence at least 100 bp long within the gene. Only coding regions are considered.* **Phylostrat***: Age is obtained from [26] by translating older age classes into our oldest age (Vertebrates).* **GeneTree***: age is obtained from Ensemble GeneTrees, dating each gene to the last common ancestor of human and the most distant species in which any type of ortholog was detectable. When we used the Oldest Exon approach the majority of genes showed the same age attributed by Phylostrat or GeneTree, reflecting the ancestor in which the core of the gene was generated de novo. On the other hand the Weighted Average tend to assign more recent ages: this age definition decreases when either new sequences are acquired or existing exons are lost.*

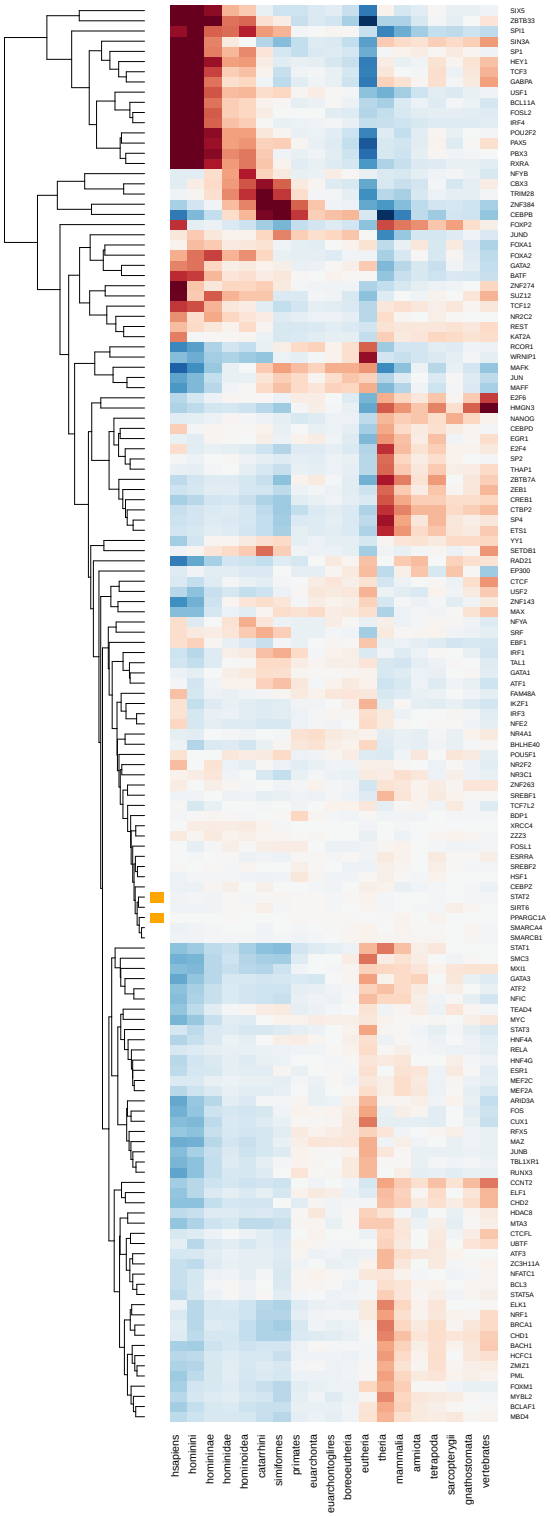

FIGURE A.6: Enrichment of ChIP-seq peaks for each TF in genomic regions of different evolutionary age, full. *Same as Fig. 3, including all TFs studied*

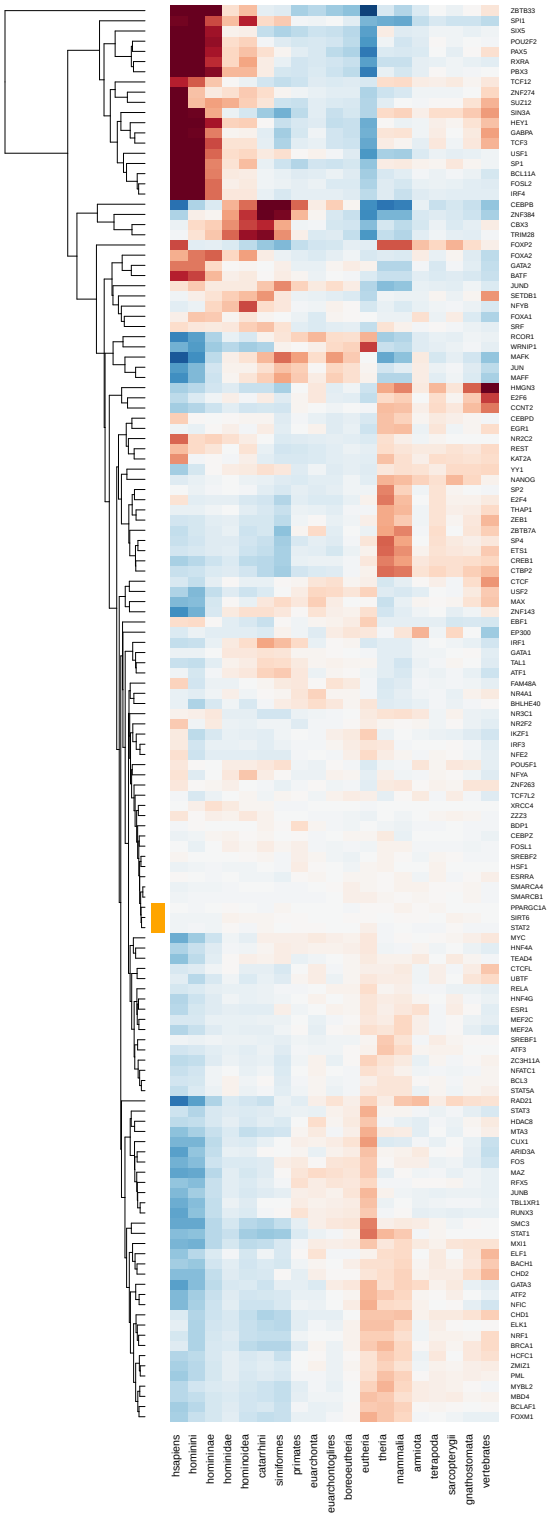

FIGURE A.7: Enrichment of ChIP-seq peaks for each TF in genomic regions of different evolutionary age, full, minimum ages. *Same as Fig. A.6, but based on minimum genomic ages*

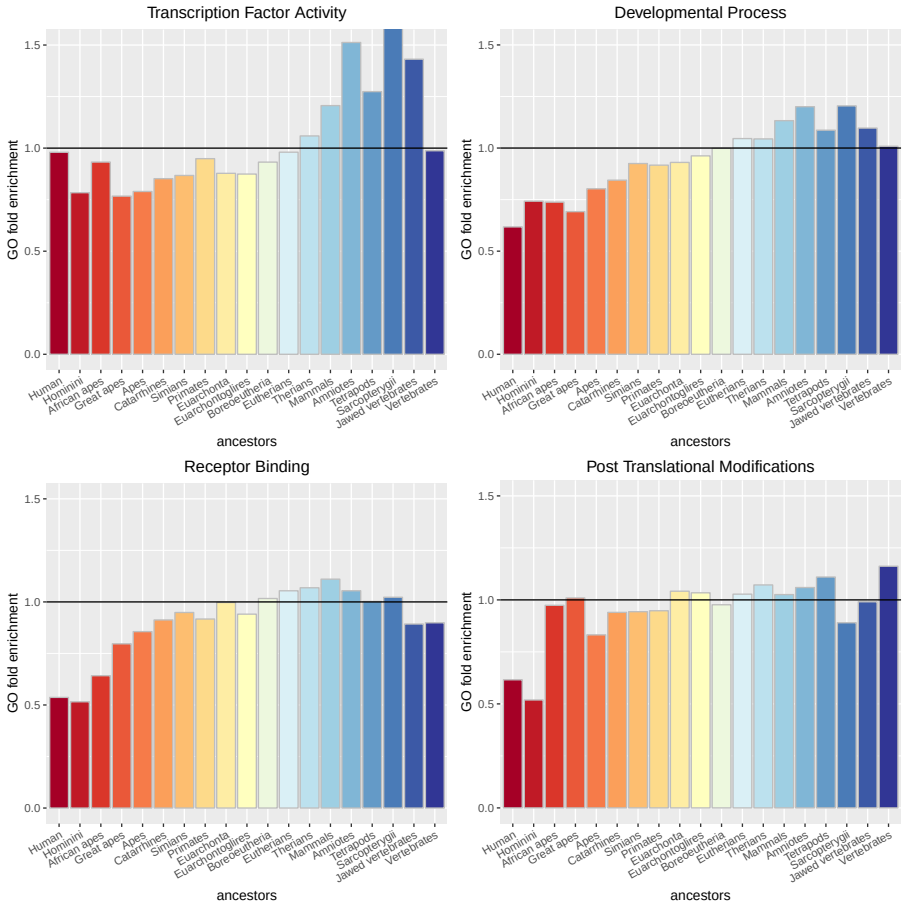

FIGURE A.8: Age enrichment of TFBSs targeting genes in the GO categories shown in Ref.[4]. *We show the ratio between the number of TFBS of each age targeting genes annotated to the GO category and the expected number based on the total number of TFBS of the same age.*

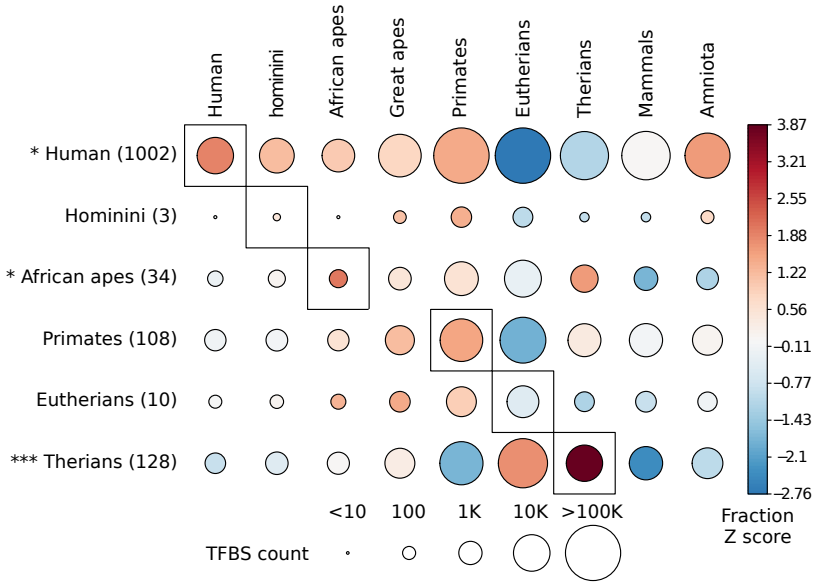

FIGURE A.9: Age distribution of TFBSs in the regulatory region of genes which underwent an expression shift in a human ancestor, minimum ages. *Same as Fig. 7, but based on minimum genomic ages*

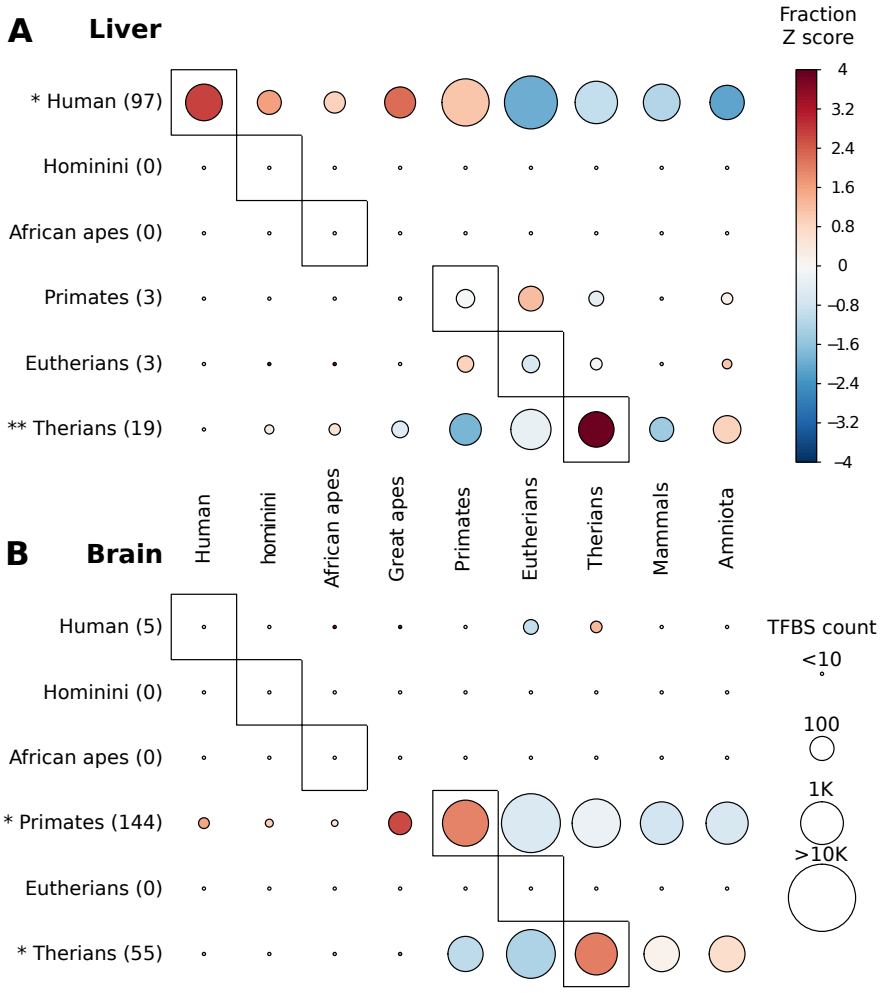

FIGURE A.10: Age distribution of TFBSs in the regulatory region of genes which underwent an expression shift in a human ancestor, in Liver or Brain. *Same as Fig. 7, but using only expression shifts and TFBS ChIP-seq peaks from selected tissues/cell lines. (A) Genes shifted in liver and ChIP-seq peaks from Hepg2. (B) Genes shifted in brain and ChIP-seq peaks from several brain derived cell lines: PFSK-1, SK-N-MC, SK-N-SH, SK-N-SH_RA, U87.*

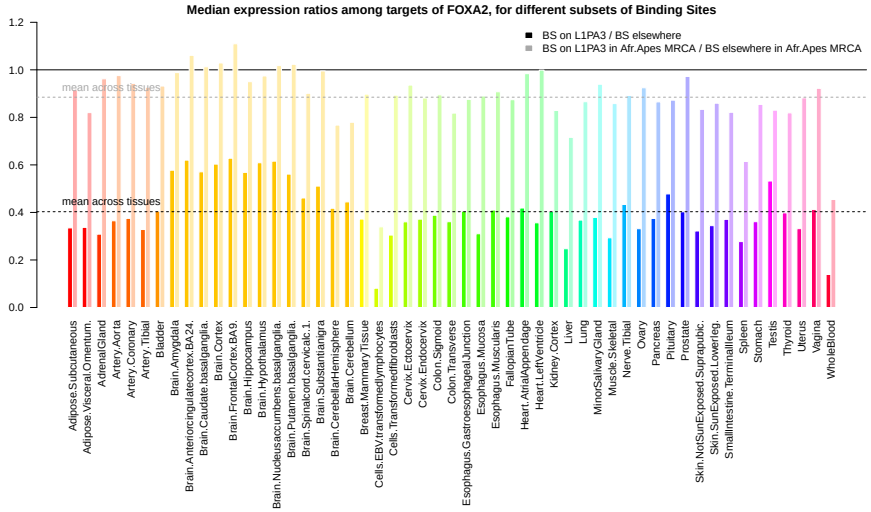

FIGURE A.11: Expression of L1PA3-driven FOXA2 targets. *Ratio of median expression of L1PA3-driven FOXA2 targets versus other targets in human tissues. Dark bars: ratio of median expression of all L1PA3-driven targets and median expression of all other FOXA2 targets, regardless of genomic age of the TFBS. Light bars: the same considering only TFBS in genomic regions acquired in the most recent common ancestor of African Apes, the genomic age in which the strongest expansion of L1PA3 took place.*

Acknowledgements

I want to acknowledge the work of:

- Chenling X. Antelope on the tables at Section 1.3.2 and on the shared track of homozygosity at Section 3.3.1
- Fernando Racimo on the Section 3.4.1 and on the plot in Figure 3.4.
- All my collaborators in the projects here presented [19, 48, 49, 55, 67] with special mentions for Ivan Molineris, Elena Grassi and Federica Mantica.
- My supervisors in the projects here presented: Prof. Emilia Huerta-Sànchez and moreover Prof. Paolo Provero. Their generous support and mentorship has always been available when I needed it, constituting the basis of these projects and becoming a cornerstone of my academic experience until now.

I would like to thank Ugo Ala, Elisa Mariella, Ettore Zapparoli, Christian Damasco, Antonio Lembo, Mattia Forneris for discussions, comments and suggestions: even if not direct collaborators on the projects here presented, their know-how and support has been always crucial.

I would like to thank Tyler Linderoth, Stefan Prost, Fernando Racimo, Fergal Casey and Peter Wilton for their insightful comments and software testing on *Haplostrips* and Prof Rasmus Nielsen for hosting me at UC Berkeley. Thanks also to Christoph Theunert, Maria Eugenia Lopez Dinamarca and again Tyler Lynderoth for making me feel at home in Berkeley.

Finally, I would like to thank my fellow researchers that shared with me a long way through the academia: Amerigo Pagoto, Lorena Consolino, Jean Piero Margaria, Roberto Ruiu, Maria Rosaria Ruggiero. And now, for more than 5 years, Arianna Marengo.

Web Resources

- **UCSC genome browser**: http://genome.ucsc.edu
- **The 1000 Genomes Project**: http://www.internationalgenome.org/data.
- **The GTEx consortium**: http://www.gtexportal.org
- **R Language website**: http:// www.R-project.org
- **Bioconductor**: http://www.bioconductor.org
- **BioNumbers**: http://bionumbers.hms.harvard.edu
- **BEDtools**: http://bedtools.readthedocs.org
- **GREAT**: http://great.stanford.edu
- **Online Mendelian Inheritance in Man**: http://www.omim.org
- **Timetree**: http://www.timetree.org

Bibliography

[1]J. L. King and T. H. Jukes, "Non-Darwinian evolution.", Science **164**, 788 (1969).

[2]D. G. Shu, H. L. Luo, S. Conway Morris, et al., "Lower Cambrian vertebrates from south China", Nature **402**, 42 (1999).

[3]K. E. Langergraber, K. Prufer, C. Rowney, et al., "Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution", Proceedings of the National Academy of Sciences **109**, 15716 (2012).

[4]B. Wood, "Reconstructing human evolution: Achievements, challenges, and opportunities", Proceedings of the National Academy of Sciences **107**, 8902 (2010).

[5]B. Villmoare, W. H. Kimbel, C. Seyoum, et al., "Early Homo at 2.8 Ma from Ledi-Geraru, Afar, Ethiopia", Science **347**, 1352 (2015).

[6]M. Y. Dennis, X. Nuttle, P. H. Sudmant, et al., "Evolution of Human-Specific Neural SRGAP2 Genes by Incomplete Segmental Duplication", Cell **149**, 912 (2012).

[7]T. Garcia, G. Féraud, C. Falguères, et al., "Earliest human remains in Eurasia: New 40Ar/39Ar dating of the Dmanisi hominid-bearing levels, Georgia", Quaternary Geochronology **5**, 443 (2010).

[8]G. Shen, X. Gao, B. Gao, and D. E. Granger, "Age of Zhoukoudian Homo erectus determined with 26Al/ 10Be burial dating", Nature **458**, 198 (2009).

[9]G. D. van den Bergh, Y. Kaifu, I. Kurniawan, et al., "Homo floresiensis-like fossils from the early Middle Pleistocene of Flores", Nature **534**, 245 (2016).

[10]K. Prüfer, F. Racimo, N. Patterson, et al., "The complete genome sequence of a Neanderthal from the Altai Mountains", Nature **505**, 43 (2013).

[11]C. M. Schlebusch, H. Malmström, T. Günther, et al., "Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago", Science **358**, 652 (2017).

[12]Q. Fu, H. Li, P. Moorjani, et al., "Genome sequence of a 45,000-year-old modern human from western Siberia.", Nature **514**, 445 (2014).

[13]L. Pagani, D. J. Lawson, E. Jagoda, et al., "Genomic analyses inform on migration events during the peopling of Eurasia", Nature **538**, 238 (2016).

[14]S. Mallick, H. Li, M. Lipson, et al., "The Simons Genome Diversity Project: 300 genomes from 142 diverse populations", Nature **538**, 201 (2016).

[15]R. E. Green, J. Krause, A. W. Briggs, et al., "A Draft Sequence of the Neandertal Genome", Science **328**, 710 (2010).

[16]D. Reich, R. E. Green, M. Kircher, et al., "Genetic history of an archaic hominin group from Denisova Cave in Siberia", Nature **468**, 1053 (2010).

[17]M. Meyer, M. Kircher, M.-T. Gansauge, et al., "A High-Coverage Genome Sequence from an Archaic Denisovan Individual", Science **338**, 222 (2012).

[18]E. Huerta-Sánchez, X. Jin, Asan, et al., "Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA.", Nature **512**, 194 (2014).

[19]F. Racimo, D. Marnetto, and E. Huerta-Sánchez, "Signatures of archaic adaptive introgression in present-day human populations", Molecular Biology and Evolution **34**, msw216 (2016).

[20]K. Harris and R. Nielsen, "The Genetic Cost of Neanderthal Introgression", Genetics **203**, 881 (2016).

[21]K. Harris and R. Nielsen, "Q&A: Where did the Neanderthals go?", BMC Biology **15**, 73 (2017).

[22]E. E. Eichler, "Structural Dynamics of Eukaryotic Chromosome Evolution", Science **301**, 793 (2003).

[23]M. R. Stratton, P. J. Campbell, and P. A. Futreal, "The cancer genome", Nature **458**, 719 (2009).

[24]B. Chénais, A. Caruso, S. Hiard, and N. Casse, "The impact of transposable elements on eukaryotic genomes: From genome size increase to genetic adaptation to stressful environments", Gene **509**, 7 (2012).

[25]H. H. Kazazian, "Mobile Elements: Drivers of Genome Evolution", Science **303**, 1626 (2004).

[26]A. Kapusta, A. Suh, and C. Feschotte, "Dynamics of genome size evolution in birds and mammals", Proceedings of the National Academy of Sciences **114**, E1460 (2017).

[27]D. W. Burt, C. Bruley, I. C. Dunn, et al., "The dynamics of chromosome evolution in birds and mammals", Nature **402**, 411 (1999).

[28]J. Prado-Martinez, P. H. Sudmant, J. M. Kidd, et al., "Great ape genetic diversity and population history.", Nature **499**, 471 (2013).

[29]J. Ng, J. S. Trask, D. G. Smith, and S. Kanthaswamy, "Heterospecific SNP diversity in humans and rhesus macaque (Macaca mulatta)", Journal of Medical Primatology **44**, 194 (2015).

[30]P. Moorjani, Z. Gao, and M. Przeworski, "Human Germline Mutation and the Erratic Evolutionary Clock", PLoS Biology **14**, 11 (2016).

[31]M. Kimura and T. Ohta, "The Average Number of Generations until Fixation of a Mutant Gene in a Finite Population.", Genetics **61**, 763 (1969).

[32]S. P. Otto and M. C. Whitlock, "Fixation Probabilities and Times", eLS, 1 (2008).

[33]R. Nielsen, C. Bustamante, A. G. Clark, et al., "A scan for positively selected genes in the genomes of humans and chimpanzees", PLoS Biology **3**, 0976 (2005).

[34]M. Slatkin and B. Rannala, "Estimating allele age.", Annual review of genomics and human genetics **1**, 225 (2000).

[35]S. Hanein, I. Perrault, S. Gerber, et al., "Population history and infrequent mutations: How old is a rare mutation? GUCY2D as a worked example", European Journal of Human Genetics **16**, 115 (2008).

[36]R. Nielsen, I. Hellmann, M. Hubisz, C. Bustamante, and A. G. Clark, "Recent and ongoing selection in the human genome", Nature Reviews Genetics **8**, 857 (2007).

[37]J. B. W. Wolf, J. Lindell, and N. Backström, "Speciation genetics: current status and evolving approaches.", Philosophical transactions of the Royal Society of London. Series B, Biological sciences **365**, 1717 (2010).

[38]R. G. Harrison and E. L. Larson, "Hybridization, introgression, and the nature of species boundaries", Journal of Heredity **105**, 795 (2014).

[39]A. Siepel, "Phylogenomics of primates and their ancestral populations", Genome Research **19**, 1929 (2009).

[40]B. Padhukasahasram, "Inferring ancestry from population genomic data and its applications", Frontiers in Genetics **5**, 1 (2014).

[41]M. Blanchette, W. J. Kent, C. Riemer, et al., "Aligning multiple genomic sequences with the threaded blockset aligner", Genome Research **14**, 708 (2004).

[42]W. J. Kent, R. Baertsch, A. Hinrichs, W. Miller, and D. Haussler, "Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes", Proceedings of the National Academy of Sciences **100**, 11484 (2003).

[43]R. Haygood, O. Fedrigo, B. Hanson, K.-D. Yokoyama, and G. A. Wray, "Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution", Nature Genetics **39**, 1140 (2007).

[44]S. Kumar and S. Subramanian, "Mutation rates in mammalian genomes", Proceedings of the National Academy of Sciences **99**, 803 (2002).

[45]G. M. Cooper, E. A. Stone, G. Asimenos, et al., "Distribution and intensity of constraint in mammalian genomic sequence.", Genome research **15**, 901 (2005).

[46]G. Bejerano, "Ultraconserved Elements in the Human Genome", Science **304**, 1321 (2004).

[47]K. Lindblad-Toh, M. Garber, O. Zuk, et al., "A high-resolution map of human evolutionary constraint using 29 mammals.", Nature **478**, 476 (2011).

[48]D. Marnetto, I. Molineris, F. Mantica, et al., "Evolutionary rewiring of human regulatory networks by waves of genome expansion", American Journal of Human Genetics **In Press** (2017).

[49]D. Marnetto, I. Molineris, E. Grassi, and P. Provero, "Genome-wide identification and characterization of fixed human-specific regulatory regions", American Journal of Human Genetics **95**, 39 (2014).

[50]T. Domazet-Lošo, J. Brajković, and D. Tautz, "A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages", Trends in Genetics **23**, 533 (2007).

[51]S. Kim and A. Misra, "SNP Genotyping: Technologies and Biomedical Applications", Annual Review of Biomedical Engineering **9**, 289 (2007).

[52]K. A. Frazer, D. G. Ballinger, D. R. Cox, et al., "A second generation human haplotype map of over 3.1 million SNPs", Nature **449**, 851 (2007).

[53]The 1000 Genomes Project Consortium, A. Auton, G. R. Abecasis, et al., "A global reference for human genetic variation", Nature **526**, 68 (2015).

[54]K. Walter, J. L. Min, J. Huang, et al., "The UK10K project identifies rare variants in health and disease", Nature **526**, 82 (2015).

[55]C. X. Antelope, D. Marnetto, F. Casey, and E. Huerta-sanchez, "Leveraging Multiple Populations across Time Helps Defi ne Accurate Models of Human Evolution: A Reanalysis of the Lactase Persistence Adaptation", (2017).

[56]R. Nielsen and M. Slatkin, *An Introduction to Population Genetics: Theory and Applications* (Sinauer Associates, 2013), p. 287.

[57]M. Kimura, "Evolutionary Rate at the Molecular Level", Nature **217**, 624 (1968).

[58]B. F. Voight, S. Kudaravalli, X. Wen, and J. K. Pritchard, "A map of recent positive selection in the human genome", PLoS Biology **4**, 0446 (2006).

[59]P. C. Sabeti, D. E. Reich, J. M. Higgins, et al., "Detecting recent positive selection in the human genome from haplotype structure.", Nature **419**, 832 (2002).

[60]P. C. Sabeti, P. Varilly, B. Fry, et al., "Genome-wide detection and characterization of positive selection in human populations.", Nature **449**, 913 (2007).

[61]A. Ferrer-Admetlla, M. Liang, T. Korneliussen, and R. Nielsen, "On Detecting Incomplete Soft or Hard Selective Sweeps Using Haplotype Structure", Molecular Biology and Evolution **31**, 1275 (2014).

[62]C. D. Bustamante, J Wakeley, S Sawyer, and D. L. Hartl, "Directional selection and the site-frequency spectrum.", Genetics **159**, 1779 (2001).

[63]J. Reynolds, B. S. Weir, and C. C. Cockerham, "Estimation of the coancestry coefficient: Basis for a short-term genetic distance", Genetics **105**, 767 (1983).

[64]X. Yi, Y. Liang, E. Huerta-Sanchez, et al., "Sequencing of 50 human exomes reveals adaptation to high altitude.", Science (New York, N.Y.) **329**, 75 (2010).

[65]K. Zeng, Y. X. Fu, S. Shi, and C. I. Wu, "Statistical tests for detecting positive selection by utilizing high-frequency variants", Genetics **174**, 1431 (2006).

[66]B. F. Voight, S. Kudaravalli, X. Wen, et al., "Detection of human adaptation during the past 2,000 years", bioRxiv pre-print **0776**, 1 (2016).

[67]D. Marnetto and E. Huerta-Sánchez, "Haplostrips : revealing population structure through haplotype visualization", Methods in Ecology and Evolution **8**, edited by S. Price, 1389 (2017).

[68]S. Pääbo, H. Poinar, D. Serre, et al., "Genetic Analyses from Ancient DNA", Annual Review of Genetics **38**, 645 (2004).

[69]M. E. Allentoft, M. Collins, D. Harker, et al., "The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils.", Proceedings. Biological sciences **279**, 4724 (2012).

[70]F. Racimo, G. Renaud, and M. Slatkin, "Joint Estimation of Contamination, Error and Demography for Nuclear DNA from Ancient Humans", PLoS Genetics **12**, 1 (2016).

[71]M. E. Allentoft, M. Sikora, K.-G. Sjögren, et al., "Population genomics of Bronze Age Eurasia", Nature **522**, 167 (2015).

[72]I. Mathieson, I. Lazaridis, N. Rohland, et al., "Eight thousand years of natural selection in Europe", bioRxiv, 016477 (2015).

[73]M. Slatkin and F. Racimo, "Ancient DNA and human history", Proceedings of the National Academy of Sciences **113**, 6380 (2016).

[74]M. King and A. Wilson, *Evolution at two levels in humans and chimpanzees*, 1975.

[75]S. B. Carroll, "Evolution at two levels: On genes and form", PLoS Biology **3**, 1159 (2005).

[76]D. Villar, P. Flicek, and D. T. Odom, "Evolution of transcription factor binding in metazoans - mechanisms and functional implications", Nat Rev Genet **15**, 221 (2014).

[77]S. A. Tishkoff, F. A. Reed, A. Ranciaro, et al., "Convergent adaptation of human lactase persistence in Africa and Europe.", Nature genetics **39**, 31 (2007).

[78]C. T. Miller, S. Beleza, A. A. Pollen, et al., "cis-Regulatory changes in Kit ligand expression and parallel evolution of pigmentation in sticklebacks and humans.", Cell **131**, 1179 (2007).

[79]G. a. Wray, "The evolutionary significance of cis-regulatory mutations.", Nature reviews. Genetics **8**, 206 (2007).

[80]E. Chan, G. Quon, G. Chua, et al., "Conservation of core gene expression in vertebrate tissues", Journal of Biology **8**, 33 (2009).

[81]I. S. Peter and E. H. Davidson, "Evolution of gene regulatory networks controlling body plan development", Cell **144**, 970 (2011).

[82]M. Rebeiz, N. Jikomes, V. A. Kassner, and S. B. Carroll, "Evolutionary origin of a novel gene expression pattern through co-option of the latent activities of existing regulatory sequences.", Proceedings of the National Academy of Sciences of the United States of America **108**, 10036 (2011).

[83]C. B. Lowe, M. Kellis, A. Siepel, et al., "Innovation During Vertebrate Evolution", Science **333** (2011) 10.1126/science.1202702.

[84]L. D. Ward and M. Kellis, "Evidence of abundant purifying selection in humans for recently acquired regulatory functions.", Science (New York, N.Y.) **337**, 1675 (2012).

[85]D. Schmidt, P. C. Schwalie, M. D. Wilson, et al., "Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages", Cell **148**, 335 (2012).

[86]V. Sundaram, Y. Cheng, Z. Ma, et al., "Widespread contribution of transposable elements to the innovation of gene regulatory networks", Genome Research **24**, 1963 (2014).

[87]D. Villar, C. Berthelot, S. Aldridge, et al., "Enhancer Evolution across 20 Mammalian Species", Cell **160**, 554 (2015).

[88]D. Emera, J. Yin, S. K. Reilly, J. Gockley, and J. P. Noonan, "Origin and evolution of developmental enhancers in the mammalian neocortex", (2016) 10.1073/pnas.1603718113.

[89]J. Ito, R. Sugimoto, H. Nakaoka, et al., "Systematic identification and characterization of regulatory elements derived from human endogenous retroviruses.", PLoS genetics **13**, e1006883 (2017).

[90]D. G. Torgerson, A. R. Boyko, R. D. Hernandez, et al., "Evolutionary processes acting on candidate cis-regulatory regions in humans inferred from patterns of polymorphism and divergence", PLoS Genetics **5**, e1000592 (2009).

[91]N. Jayaram, D. Usvyat, and A. C. R. Martin, "Evaluating tools for transcription factor binding site prediction", BMC Bioinformatics (2016) 10.1186/s12859-016-1298-9.

[92]E. Grassi, E. Zapparoli, I. Molineris, and P. Provero, "Total Binding Affinity Profiles of Regulatory Regions Predict Transcription Factor Binding and Gene Expression in Human Cells", PLOS ONE **10**, edited by S. Aerts, e0143627 (2015).

[93]A. Kundaje, W. Meuleman, J. Ernst, et al., "Integrative analysis of 111 reference human epigenomes", Nature **518**, 317 (2015).

[94]J. Ernst, P. Kheradpour, T. S. Mikkelsen, et al., "Mapping and analysis of chromatin state dynamics in nine human cell types.", Nature **473**, 43 (2011).

[95]T. Ahsendorf, F.-J. Müller, V. Topkar, J. Gunawardena, and R. Eils, "Transcription factors, coregulators, and epigenetic marks are linearly correlated and highly redundant", Plos One **12**, e0186324 (2017).

[96]D. L. Halligan, F. Oliver, J. Guthrie, et al., "Positive and negative selection in murine ultraconserved noncoding elements", Molecular Biology and Evolution **28**, 2651 (2011).

[97]M. Hemberg, J. M. Gray, N. Cloonan, et al., "Integrated genome analysis suggests that most conserved non-coding sequences are regulatory factor binding sites", Nucleic Acids Research **40**, 7858 (2012).

[98]D. Brawand, M. Soumillon, A. Necsulea, et al., "The evolution of gene expression levels in mammalian organs.", Nature **478**, 343 (2011).

[99]O. Delaneau, M. Zazhytska, C. Borel, et al., "Intra- and inter-chromosomal chromatin interactions mediate genetic effects on regulatory networks", bioRxiv, 171694 (2017).

[100]T. Lappalainen, M. Sammeth, M. R. Friedländer, et al., "Transcriptome and genome sequencing uncovers functional variation in humans.", Nature **501**, 506 (2013).

[101]K. G. Ardlie, D. S. Deluca, A. V. Segre, et al., "The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans", Science **348**, 648 (2015).

[102]T. Berisa and J. K. Pickrell, "Approximately independent linkage disequilibrium blocks in human populations.", Bioinformatics (Oxford, England) **32**, 283 (2016).

[103]B. Deplancke, D. Alpern, and V. Gardeux, "The Genetics of Transcription Factor DNA Binding Variation", Cell **166**, 538 (2016).

[104]D. A. Cusanovich, B. Pavlovic, J. K. Pritchard, and Y. Gilad, "The Functional Consequences of Variation in Transcription Factor Binding", PLoS Genetics **10** (2014) 10.1371/journal.pgen.1004226.

[105]K. Beal, M. Spivakov, E. E. Furlong, et al., "Analysis of variation at transcription factor binding sites in Drosophila and humans", Genome Biology **13**, R49 (2012).

[106]J. H. McDonald and M. Kreitman, "Adaptive protein evolution at the Adh locus in Drosophila", Nature **354**, 293 (1991).

[107]I. Gronau, L. Arbiza, J. Mohammed, and A. Siepel, "Inference of Natural Selection from Interspersed Genomic Elements Based on Polymorphism and Divergence", **30**, 1159 (2013).

[108]I. Molineris, E. Grassi, U. Ala, F. Di Cunto, and P. Provero, "Evolution of promoter affinity for transcription factors in the human lineage", Molecular Biology and Evolution **28**, 2173 (2011).

[109]C. Y. McLean, D. Bristor, M. Hiller, et al., "GREAT improves functional interpretation of cis-regulatory regions.", Nature biotechnology **28**, 495 (2010).

[110]M. Krupp, J. U. Marquardt, U. Sahin, et al., "RNA-Seq Atlas–a reference database for gene expression profiling in normal tissue by next-generation sequencing.", Bioinformatics (Oxford, England) **28**, 1184 (2012).

[111]A. Mathelier, X. Zhao, A. W. Zhang, et al., "JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles.", Nucleic acids research **42**, D142 (2014).

[112] A. Levy, S. Schwartz, and G. Ast, "Large-scale discovery of insertion hotspots and preferential integration sites of human transposed elements.", Nucleic acids research **38**, 1515 (2010).

[113] Y. Wolf, P. Novichkov, G. Karev, E. Koonin, and D. Lipman, "The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages", Proceedings of the National Academy of Sciences **106**, 7273 (2009).

[114] L. Zhang and W.-H. Li, "Mammalian Housekeeping Genes Evolve More Slowly than Tissue-Specific Genes", Molecular Biology and Evolution **21**, 236 (2004).

[115] I. E. Schor, J. F. Degner, D. Harnett, et al., "Promoter shape varies across populations and affects promoter evolution and expression noise.", Nature genetics **49**, 550 (2017).

[116] M. M. Albà and J. Castresana, "Inverse relationship between evolutionary rate and age of mammalian genes.", Molecular biology and evolution **22**, 598 (2005).

[117] R. Neme and D. Tautz, "Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution.", BMC genomics **14**, 117 (2013).

[118] A. J. Vilella, J. Severin, A. Ureta-Vidal, et al., "EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates.", Genome research **19**, 327 (2009).

[119] D. Schmidt, M. D. Wilson, B. Ballester, et al., "Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding.", Science (New York, N.Y.) **328**, 1036 (2010).

[120] F. Yue, Y. Cheng, A. Breschi, et al., "A comparative encyclopedia of DNA elements in the mouse genome.", Nature **515**, 355 (2014).

[121] A.-R. Carvunis, T. Wang, D. Skola, et al., "Evidence for a common evolutionary rate in metazoan transcriptional networks.", eLife **4** (2015) 10.7554/eLife.11615.

[122] M. Huber and M. Lohoff, "IRF4 at the crossroads of effector T-cell fate decision.", European journal of immunology **44**, 1886 (2014).

[123] A. G. Rolink, S. L. Nutt, F Melchers, and M Busslinger, "Long-term in vivo reconstitution of T-cell development by Pax5-deficient B-cell progenitors.", Nature **401**, 603 (1999).

[124] P Pfisterer, J Hess, and T Wirth, "Identification of target genes of the lymphoid-specific transcription factor Oct2.", Immunobiology **198**, 217 (1997).

[125] D. J. Mangelsdorf, K Umesono, S. A. Kliewer, et al., "A direct repeat in the cellular retinol-binding protein type II gene confers differential regulation by RXR and RAR.", Cell **66**, 555 (1991).

[126] H. Kölsch, D. Lütjohann, F. Jessen, et al., "RXRA gene variations influence Alzheimer's disease risk and cholesterol metabolism.", Journal of cellular and molecular medicine **13**, 589 (2009).

[127] P. Pajukanta, H. E. Lilja, J. S. Sinsheimer, et al., "Familial combined hyperlipidemia is associated with upstream transcription factor 1 (USF1)", Nature Genetics **36**, 371 (2004).

[128] F. M. J. Jacobs, D. Greenberg, N. Nguyen, et al., "An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons.", Nature **516**, 242 (2014).

[129]T Kimura, R Ivell, W Rust, et al., "Molecular cloning of a human MafF homologue, which specifically binds to the oxytocin receptor gene in term myometrium.", Biochemical and biophysical research communications **264**, 86 (1999).

[130]V. J. Lynch, R. D. Leclerc, G. May, and G. P. Wagner, "Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals.", Nature genetics **43**, 1154 (2011).

[131]W. Enard, M. Przeworski, S. E. Fisher, et al., "Molecular evolution of FOXP2, a gene involved in speech and language.", Nature **418**, 869 (2002).

[132]S. Chung, A. Leung, B.-S. Han, et al., "Wnt1-lmx1a Forms a Novel Autoregulatory Loop and Controls Midbrain Dopaminergic Differentiation Synergistically with the SHH-FoxA2 Pathway", Cell Stem Cell **5**, 646 (2009).

[133]E. Arenas, "Wnt signaling in midbrain dopaminergic neuron development and regenerative medicine for Parkinson's disease.", Journal of molecular cell biology **6**, 42 (2014).

[134]G. Jäger, A. Peltzer, K. Nieselt, et al., "inPHAP : Interactive visualization of genotype and phased haplotype data", BMC Bioinformatics **15**, 1 (2014).

[135]E. Paradis, "Pegas: An R package for population genetics with an integrated-modular approach", Bioinformatics **26**, 419 (2010).

[136]J. C. Barrett, B. Fry, J. Maller, and M. J. Daly, "Haploview: Analysis and visualization of LD and haplotype maps", Bioinformatics **21**, 263 (2005).

[137]L. Ségurel and C. Bon, "On the Evolution of Lactase Persistence in Humans.", Annual review of genomics and human genetics **18**, 297 (2017).

[138]C. J. E. Ingram, C. A. Mulcare, Y. Itan, M. G. Thomas, and D. M. Swallow, "Lactose digestion and the evolutionary genetics of lactase persistence.", Human genetics **124**, 579 (2009).

[139]N. S. Enattah, T. Sahi, E. Savilahti, et al., "Identification of a variant associated with adult-type hypolactasia", Nature genetics **30**, 233 (2002).

[140]T. Bersaglieri, P. C. Sabeti, N. Patterson, et al., "Genetic signatures of strong recent positive selection at the lactase gene.", American journal of human genetics **74**, 1111 (2004).

[141]Y. Itan, A. Powell, M. A. Beaumont, J. Burger, and M. G. Thomas, "The Origins of Lactase Persistence in Europe", **5**, 17 (2009).

[142]N. S. Enattah, T. G. K. Jensen, M. Nielsen, et al., "Independent Introduction of Two Lactase-Persistence Alleles into Human Populations Reflects Different History of Adaptation to Milk Culture", 57 (2008).

[143]N. S. Enattah, A. Trudeau, V. Pimenoff, et al., "Evidence of Still-Ongoing Convergence Evolution of the Lactase Persistence T - 1 3 9 1 0 Alleles in Humans", **81**, 615 (2007).

[144]S. Sankararaman, S. Mallick, M. Dannemann, et al., "The genomic landscape of Neanderthal ancestry in present-day humans", Nature **507**, 354 (2014).

[145]B. Vernot, S. Tucci, J. Kelso, et al., "Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals.", Science (New York, N.Y.) **352**, 235 (2016).

[146]C.-J. Hu, L.-Y. Wang, L. A. Chodosh, B. Keith, and M. C. Simon, "Differential roles of hypoxia-inducible factor 1alpha (HIF-1alpha) and HIF-2alpha in hypoxic gene regulation.", Molecular and cellular biology **23**, 9361 (2003).

[147]C. M. Beall, G. L. Cavalleri, L. Deng, et al., "Natural selection on EPAS1 (HIF2alpha) associated with low hemoglobin concentration in Tibetan highlanders.", Proceedings of the National Academy of Sciences of the United States of America **107**, 11459 (2010).

[148]S. Hackinger, T. Kraaijenbrink, Y. Xue, et al., "Wide distribution and altitude correlation of an archaic high-altitude-adaptive EPAS1 haplotype in the Himalayas.", Human genetics **135**, 393 (2016).

[149]W. McKinney, "pandas: a Foundational Python Library for Data Analysis and Statistics", Python for High Performance and Scientific Computing, 1 (2011).

[150]H. Li, B. Handsaker, A. Wysoker, et al., "The Sequence Alignment/Map format and SAMtools", Bioinformatics **25**, 2078 (2009).

[151]R Hudson, "Ms a Program for Generating Samples Under Neutral Models", Bioinformatics **18**, 337 (2002).

[152]R. Durbin, "Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT)", Bioinformatics **30**, 1266 (2014).

[153]S. Gravel, B. M. Henn, R. N. Gutenkunst, et al., "Demographic history and rare allele sharing among human populations", **108** (2011) 10.1073/pnas.1019276108.

[154]R. E. Giles, N Shimizu, and F. H. Ruddle, "Assignment of a human genetic locus to chromosome 5 which corrects the heat sensitive lesion associated with reduced leucyl-tRNA synthetase activity in ts025Cl Chinese hamster cells.", Somatic cell genetics **6**, 667 (1980).

[155]J. P. Casey, P. McGettigan, N. Lynam-Lennon, et al., "Identification of a mutation in LARS as a novel cause of infantile hepatopathy.", Molecular genetics and metabolism **106**, 351 (2012).

[156]J. K. Lim, A. Lisco, D. H. McDermott, et al., "Genetic variation in OAS1 is a risk factor for initial infection with West Nile virus in man.", PLoS pathogens **5**, edited by C. M. Rice, e1000321 (2009).

[157]F. L. Mendez, J. C. Watkins, and M. F. Hammer, "Neandertal origin of genetic variation at the cluster of OAS immunity genes.", Molecular biology and evolution **30**, 798 (2013).

[158]T. G. Warner, L. M. Dambach, J. H. Shin, and J. S. O'Brien, "Separation and characterization of the acid lipase and neutral esterases from human liver.", American journal of human genetics **32**, 869 (1980).

[159]H Klima, K Ullrich, C Aslanidis, et al., "A splice junction mutation causes deletion of a 72-base exon from the mRNA for lysosomal acid lipase in a patient with cholesteryl ester storage disease.", The Journal of clinical investigation **92**, 2713 (1993).

[160]C Aslanidis, S Ries, P Fehringer, et al., "Genetic and biochemical evidence that CESD and Wolman disease are distinguished by residual lysosomal acid lipase activity.", Genomics **33**, 85 (1996).

[161]E. G. Lund, T. A. Kerr, J Sakai, W. P. Li, and D. W. Russell, "cDNA cloning of mouse and human cholesterol 25-hydroxylases, polytopic membrane proteins that synthesize a potent oxysterol regulator of lipid metabolism.", The Journal of biological chemistry **273**, 34316 (1998).

[162]N. Shibata, T. Kawarai, J. H. Lee, et al., "Association studies of cholesterol metabolism genes (CH25H, ABCA1 and CH24H) in Alzheimer's disease", Neuroscience Letters **391**, 142 (2006).

[163]S.-Y. Liu, R. Aliyari, K. Chikere, et al., "Interferon-Inducible Cholesterol-25-Hydroxylase Broadly Inhibits Viral Entry by Production of 25-Hydroxycholesterol", Immunity **38**, 92 (2013).

[164]V. Gburcik, W. P. Cawthorn, J. Nedergaard, J. A. Timmons, and B. Cannon, "An essential role for Tbx15 in the differentiation of brown and "brite" but not white adipocytes.", American journal of physiology. Endocrinology and metabolism **303**, E1053 (2012).

[165]D. Shungin, T. W. Winkler, D. C. Croteau-Chonka, et al., "New genetic loci link adipose and insulin biology to body fat distribution.", Nature **518**, 187 (2015).

[166]E. Lausch, P. Hermanns, H. F. Farin, et al., "TBX15 mutations cause craniofacial dysmorphism, hypoplasia of scapula and pelvis, and short stature in Cousin syndrome.", American journal of human genetics **83**, 649 (2008).

[167]M. Fumagalli, I. Moltke, N. Grarup, et al., "Greenlandic Inuit show genetic signatures of diet and climate adaptation", Science **349**, 1343 (2015).

[168]F. Racimo, D. Gokhman, M. Fumagalli, et al., "Archaic adaptive introgression in TBX15/WARS2", bioRxiv, 033928 (2015).

[169]E. Guichard, V. Peona, G. Malagoli-Tagliazucchi, et al., "Impact of non-LTR retrotransposons in the differentiation and evolution of Anatomically Modern Humans", BioRxiv, 1 (2017).