A comprehensive screening of X-linked and miRNA-driven signatures of positive selection and the role of *cis*-regulatory elements in the human genome

Pablo Villegas Mirón

TESI DOCTORAL UPF / 2021

DIRECTORS DE LA TESI

Dr. Hafid Laayouni El Alaoui

Dr. Jaume Bertranpetit i Busquets

DEPARTAMENT DE CIÈNCIES EXPERIMENTALS I DE LA SALUT

Universitat Pompeu Fabra Barcelona

A mis padres

Acknowledgements

Esta tesis es el resultado de cuatro años de mucho leer, escribir, picar teclados, mirar pantallas, de toneladas de tiempo sentado, de mucha ansiedad y estrés, de su derivados depresivos, de preguntas sin respuesta y de respuestas sin sentido, pero sobre todo es el resultado de experiencias junto a las personas con las que he vivido y trabajado, que no son pocas, y por ello estoy agradecido.

Por supuesto, quiero empezar agradeciendo a mis padres, sin cuyo apoyo y amor nada habría sido posible, es gracias a ellos por lo que constantemente siento lo afortunado que soy. A toda mi familia, por la que siento un inmenso amor, y a mis raíces extremeñas, incrustadas en lo más profundo de mi ser.

A pesar de estar lejos de casa, Barcelona ha acabado siendo mi hogar y en ella he hecho una familia enorme, compuesta por gente increíble que, aunque va y viene, siempre se quedarán en mi memoria. Me gustaría dar las gracias a la "peñita", a Toni, Isa, Sandra, Luis, Aitor, Julen, Andrea, Tania, Pablo y Gabriela, por haberme acompañado durante todos estos años y hacerme sentir como uno más. A Alex, Edu, Vadym, Lina, Mujica, Rocío, Tina, Grecia, y otros tantos que me han dado tanto durante estos años.

Me gustaría agradecer a Sandra Acosta, por toda su ayuda y apoyo en el desarrollo de una buena parte de mi tesis. También a Yolanda Espinosa Parrilla, gracias a la cual también he podido trabajar en otro de los capítulos de este proyecto. Ambas me han hecho trabajar duro, pero ha merecido muchísimo la pena. A mis supervisores Hafid y Jaume, por haberme dado la oportunidad de trabajar en el IBE y por todo su apoyo durante estos años.

This thesis has been possible thanks to the FPI PhD fellowship (FPI-BES-2016-077706) part of the "Unidad de Excelencia María de Maeztu" funded by MINECO (ref: MDM-2014-0370).

Abstract

The evolutionary history of the human genome has been shaped by the selection of multiple elements in response to different environmental pressures. The analysis of regulatory adaptations and the particularities of sexual chromosomes are key to understanding the different evolutionary outcomes of some of these processes. In this thesis, we describe how recent selection has shaped the X chromosome in human populations, focusing on several selection candidates, the special X-linked inheritance properties and the role of regulatory elements. We also propose the relevant implication of human enhancers in tissue-specific regulatory programs. The expression of genes implicated in tissue-specific functions is seen to be regulated mainly by enhancers located in introns, while ubiquitously expressed housekeeping genes are predominantly controlled by intergenic enhancers. The evolutionary role of human miRNAs are also analyzed with special emphasis on their global patterns of diversity and their implication in population-specific prevalence in some of the most common human disorders.

Resumen

La historia evolutiva del genoma humano ha sido configurada por la selección de numerosos elementos en respuesta a distintas presiones evolutivas. El análisis de adaptaciones regulatorias y de las particularidades de los cromosomas sexuales son clave para entender las distintas consecuencias de algunos de estos procesos. En esta tesis describimos cómo procesos de selección reciente han configurado el cromosoma X en distintas poblaciones humanas, centrándonos en varios candidatos bajo selección, sus propiedades hereditarias y el papel de elementos regulatorios. También describimos la relevancia de enhancers humanos en los programas regulatorios de tejidos específicos. La expresión de genes implicados en las funciones específicas de tejido aparece principalmente regulada por enhancers ubicados en intrones, por otro lado, genes implicados en el mantenimiento básico de la célula están predominantemente controlados por enhancers intergénicos. También analizamos el papel evolutivo de microRNAs humanos, con especial énfasis en patrones de diversidad globales y su implicación en prevalencias poblaciones de algunas de las enfermedades humanas más comunes.

Preface

The main motivation that pushed me to start this PhD was not the future prospects of my professional career, or even the possibility of publishing first line articles, but curiosity. Human evolution is not only about the analysis of biological and molecular patterns, but also about the understanding of our past and the acknowledgement of the future trajectory of human societies in their biological and cultural context, ranging from collaboration and conflicts to religion and diseases.

The work started with a general screening of positive selection using the last release of 1000 Genomes project, followed by a more in depth analysis of adaptive selection in the X chromosome. The door of new opportunities appears to dig into the world of microRNAs analysis and non-coding DNA evolution. The three different projects of this work share the common objective of shading light to different aspects of human genome evolution taking advantage of new data, new collaborations and facing new challenges. Putting all together looks like a complicated task when starting to draft the final thesis presentation. However, many shared questions, methodologies and interpretations emerge very easily. At the end, this work is a tiny dot on the vast landscape of human scientific adventure to understand genomic evolution and seeks answers about many questions related to biological evolution, and to Biology in general.

Given the different issues this work is addressing and to put the different presented papers into context, I will introduce each one separately and list previous works and findings that make possible and worth these new analyses. Parts of this introduction are about biological knowledge - necessary even though not enough - to go forward with evolutionary analysis, others are about methodological tools and previous published results.

To introduce the first paper on the detection of positive selection in the X chromosome, I will provide an overview of the hard and soft sweep model of selection and the methods to detect the signatures left by these processes. The special inheritance properties of the human X chromosome are then introduced in order to provide a context to the expected evolutionary patterns and differences in comparison with the autosomes. I also provide an overview of the different insights that researchers have obtained from the analysis of natural selection on the X chromosome during the last years.

The second paper is contextualized by firstly introducing a summarized overview of the different elements implicated in the regulation of gene expression in complex organisms. The differential regulation of tissue-specific and housekeeping genes is described in order to provide a starting point to understand the sophistication of regulatory programs implicated in tissue identity. Then I introduce the detection and annotation of regulatory elements as the first step in the analysis of the evolutionary implications of gene regulation, which are hereunder summarized.

The third paper constitutes an updated and comprehensive description of the human miRNA repertoire in terms of nucleotide diversity, selection signatures and human diseases. To introduce this part I provide an overall description of the origin, biogenesis, genomic properties and function of human miRNAs. Since the discovery of the first miRNA in 1993, the interest on the phenotypic consequences of miRNA genetic variation have produced an important amount of publications on the clinical and evolutionary relevance at population level, which I summarize in the last part of this introduction.

List of publications

1. Casillas, S., Mulet, R., **Villegas-Mirón**, P., Hervas, S., Sanz, E., Velasco, D., et al. (2018). PopHuman: The human population genomics browser. Nucleic Acids Res. 46, D1003–D1010. doi:10.1093/nar/gkx943.

2. **Villegas-Mirón**, P., Acosta, S., Nye, J., Bertranpetit, J. and Laayouni, H. Chromosome X-wide analysis of positive selection in human populations: from common and private signals to selection impact on inactivated genes and enhancers-like signatures. *Submitted for publication.*

3. Borsari, B.*, **Villegas-Mirón**, P.*, Perez-Lluch, S., Turpin, S., Laayouni, H., Segarra-Casas, A., Bertranpetit, J., Guigo, R., Acosta, S. Enrichment in intronic enhancers controlling the expression of genes involved in tissue-specific functions and homeostasis. *Submitted for publication.*

4. **Villegas-Mirón**, P., Gallego, A., Bertranpetit, J., Laayouni, H. and Espinosa-Parrilla, Y. Signatures of genetic variation in human microRNAs point to processes of positive selection related to population-specific disease risks. *Submitted for publication.*

Table of contents

I. INTRODUCTION

The detection of adaptive selection in human populations

The origin of modern humans

Modern humans originated in Africa more than 200.000 years ago, being the earliest remains of anatomically modern individuals located in Ethiopia and dated to about 150-190 thousand years ago (kya) (McDougall et al 2005). From there, the colonization of the globe has meant one of the greatest challenges in our history. The human diáspora across the globe started around 50-100 kya with the Out-of-Africa event (OOA) when groups of *Homo sapiens* abandoned Africa and started to spread across Eurasia, colonizing almost all corners of the globe (Figure 1) (Nielsen et al 2017). However, this hypothesis is still under controversy, since many theories debate about the number of waves and the consecutive movements that contributed to the human dispersal (Bergström et al 2021). In this travel humans encountered and admixed with other species of hominins. *Homo sapiens* interacted with *Neanderthals* and *Denisovans* during their global expansion. The current knowledge states that all non-african populations present around 2% of *Neanderthal* ancestry in their genomes, situating the timing of admixture around 60 kya (Prüfer et al 2014, Vernot et al 2015). On the other hand, *Denisovan* ancestry has been found only in some populations, like Melanesians in Oceania with an ancestry that ranges 3-6%, and southeast Asian populations, with 0.1-0.3% of genetic material of the ancestor (Reich et al 2010, Skoglund et al 2011).

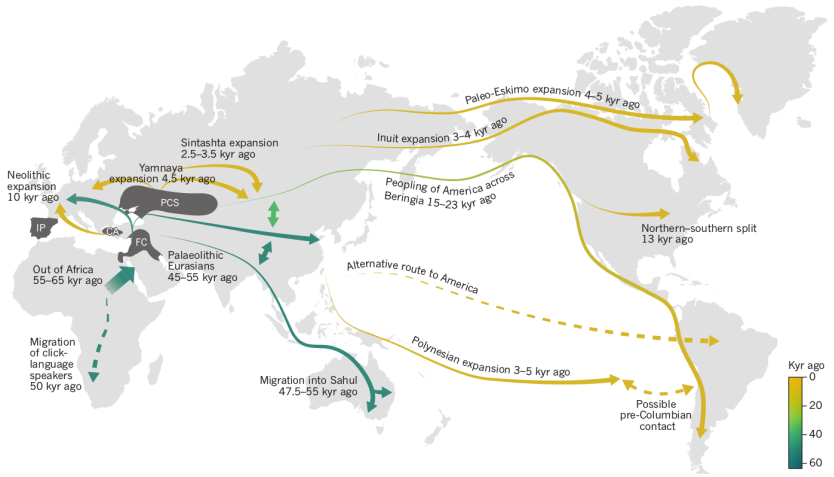


Figure 1. Major migrations across the globe during the human diaspora (Nielsen et al 2017).

The advent of sequencing technologies and its impact on the study of human variation

Since the publication of the first draft of the human genome (Lander et al 2001, Venter et al 2001), the development of *next generation sequencing technologies* (NGS) has led to the progressive decrease of the sequencing cost of individual genomes. Moreover, it facilitated the emergence of multiple new platforms that integrate different approaches and methods to increase the accuracy, speed and sequencing depth of genomes (Goodwin et al 2016). As a result, the implementation of these technologies has made it possible to open multiple lines of research devoted to the discovery of genetic variation and its role in human evolution and diseases.

Population genetic studies use several types of sequence variants, mainly those that differ in a single nucleotide, called single nucleotide polymorphisms (SNPs), and a more complex type of variants that involve longer portions of the genome termed as structural variants. The advent of sequencing projects like, the 1000 Genomes Project, has allowed the discovery of millions of single nucleotide polymorphism (SNPs) and structural variants that

became the source of thousands of published articles that shed light on the evolutionary history of *Homo sapiens* and the origins of multitude of genetic disorders (1000 Genomes Project Consortium et al 2015, Sudmant et al 2015). This project was developed in three different phases that involved an increment of the amount of data used and the improvement of the methods applied to analyse the human genetic variation. On its last release (phase III), the project made use of samples coming from 2504 individuals of 26 populations worldwide that were sequenced using both whole-genome sequencing (7.4X) and targeted exome sequencing (65.7X). This collection of samples encompasses multiple genetic backgrounds from Africa, East Asia, Europe, South Asia and the Americas, which provided a wider selection of different ancestries in comparison with other sequencing projects that present ethnic bias towards certain genetic backgrounds (Sirugo et al 2019). The amount of genetic variation described in this last phase also increased with respect to previous releases. Variant discovery made use of an integrated approach of 24 sequence analysis tools and machine learning methods that allowed to identify high quality variants from false positives, finally describing up to 80 million variants. They studied the genetic diversity across the different continental groups, providing a comprehensive description of how it has been shaped through the evolutionary history of these populations. The Out-of-Africa event (OOA) established the main differentiation among the current human populations in terms of genetic structure: non-Sub-Saharan African populations present a remarkable decrease of diversity due to the demographic reduction in the OOA bottleneck. Among other insights, they describe the genetic background of these populations and the changes of effective population sizes produced by their demographic dynamics (Figure 2). In essence, the 1000 Genomes Project is currently the most complete, unbiased and publicly available database of human genetic variation described so far.

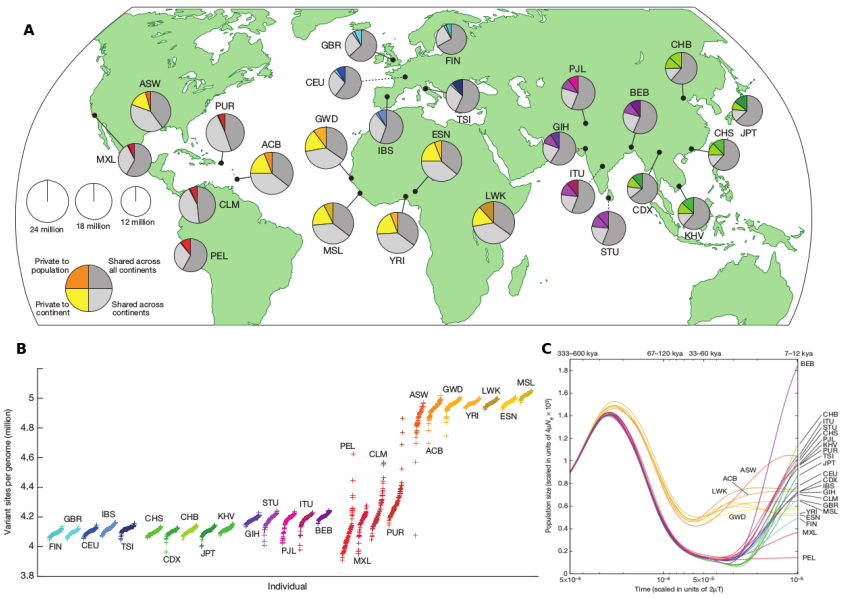


Figure 2. (A) Human populations used in the third phase of the 1000 Genomes Project. (B) Number of variant sites per individual genome. (C) Population size estimation across human populations (1000 Genomes Project Consortium et al. (2015)).

Adaptive selection in human populations

During the human diáspora, different factors contributed to shape the diversity of populations: demographic events, introgression from archaic hominins and processes of natural selection. The population movements during the human expansion led different groups of people to reach extreme climatic conditions, like the north of Siberia, and remote geographic locations, like the Tibetan or the Andian plateau. Among the major changes that experienced *Homo sapiens*, the transition from hunter-gatherer tribes to agriculturalist and pastoralist societies about 12.000 years ago became one of the cornerstones in the evolution of modern humans. This transition led *Homo sapiens* to a revolutionary development not only in biological terms, but also in the social and cultural dimension. The new conditions of this period of change led to massive increases of the population growth, which were accompanied by the emergence of

diseases, like infectious pathogens that took advantage of the population density to spread out. The dietary habits also suffered from massive changes. The cattle domestication allowed ancient populations to consume milk and its derivatives in adulthood. Also, the availability of resources and the introduction of new animal and plant species in the agricultural societies allowed to avoid periods of starvation.

The conquest of such territories and the introduction of these massive changes made humans confront environmental pressures that would shape the genetic configuration of our genome by the action of natural selection.

Charles Darwin introduced the concept of natural selection in his seminal work "*On the Origin of Species by Means of Natural Selection or the Preservation of Favored Races in the Struggle for Life*" in 1859. In a nutshell, environmental changes make species and populations face selection pressures that lead to the essence of the evolutionary game: "*the survival of the fittest*". However, this statement, coined by Herbert Spencer after reading Darwin's book, does not fit completely to modern biology. It requires the invocation of both survival and reproductive fitness of individuals to interpret the basic role of evolution: individuals that present the proper genes to survive and reproduce under specific environmental conditions would be the more adaptive and will transfer these genes to the following generation. In this sense, *positive selection*, also called adaptive selection, generates the increase in frequency of those mutations that proffer a beneficial effect or phenotype in the population. On the other hand, *negative selection* consists in the purge of mutations whose effect is deleterious from the population. This is the main force that drives the evolution of functional elements, since it is more likely that a mutation generates a damaging effect than a beneficial effect. *Balancing selection* is another mode that maintains polymorphisms at a certain allele frequencies, in this way it promotes the beneficial genetic diversity

of the affected locus. Therefore, the processes of adaptation are the genetic response of species and populations to face changes in their local environment. These processes ensure to increase the frequency of those heritable traits that make individuals present a higher fitness, while purging the traits that are detrimental to their reproduction and survival.

Numerous cases of genes under positive selection have been reported during the last decades. The most commonly detected genes under selection are those implicated in adaptations to extreme local conditions. The most well-known case of positive selection was first described by Bersaglieri et al 2004, and reports the selection of the lactase locus (*LCT*) in European populations associated with the lactase-persistence trait, which allows humans to consume milk in adulthood (Gerbault et al 2011). The transition to agricultural societies seems to be the main driver for this adaptation, a period when humans started to domesticate cattle and use milk as one of the main sources of carbohydrates. Long has been travelled since this discovery, and the current knowledge about this gene describes the selection signature also in other regions, like in west African populations (Tishkoff et al 2007). Recently, new insights about this signature implicate a miRNA located in the same locus and associated with metabolic traits, which suggests a possible relationship between this past signature and the genetic causes of current disorders like obesity and insulin resistance (Wang et al 2020).

Other genes reported under positive selection are involved in adaptations to high altitude (*HIF*, *EGLN1*, *EPAS1*), ultraviolet exposure (*SLC24A5*, *MC1R*) and resistance against pathogens (*G6PD*, *APOL1*), among many others (Fan et al 2016, Rees et al 2020) (Figure 3). However, adaptations to conditions in the past might imply maladaptations to the conditions of modern societies. The agricultural and industrial revolution that took place during the last 10.000 years has solved many of the problems that

hunter-gatherers would encounter. For example, ancient humans adapted to famine periods by increasing the energy efficiency and fat storage, adaptations that nowadays confront a very different situation when resources are almost unlimited. As a consequence, nowadays populations like Samoans present a high prevalence of metabolic-related disorders like obesity and type 2 diabetes (Minster et al 2016).

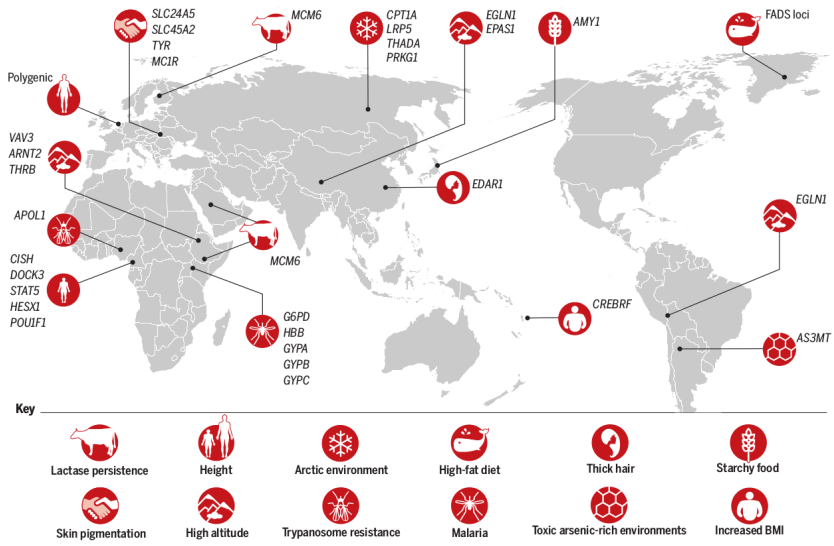


Figure 3. Examples of genes under positive selection found in different human populations that suggest processes of adaptation to local environments (Fan et al. 2016).

Signatures of positive selection, the sweep model

The signatures of positive selection are characterized by perturbations of the diversity in specific genomic regions that deviate from the usual patterns under neutrality. The most studied signature in selection scans is the so-called *hard sweep*. These signatures appear when a *de novo* mutation emerges in a region and the derived allele suddenly becomes adaptive in the local environment, becoming strongly selected and raising its frequency in the population in a short period of time. The linkage disequilibrium that this position might present with its vicinity

makes this allele to be inherited along with its surrounding sites, increasing the frequency of long unbroken haplotypes that escape from recombination. This is the well-known *hitchhiking effect* which, during the selection process, "sweeps" the variation that might be present in the genomic environment of the selected locus, leaving behind a region with no variation. This mark is seen in genomic regions as a pronounced valley of diversity where the peak of homozygosity is located on the selected allele and progressively decreases with distance (Figure 4).

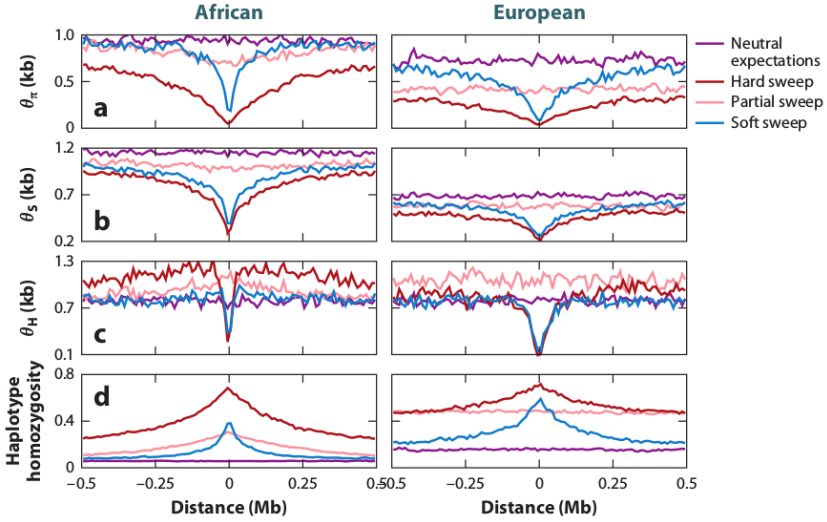


Figure 4. Metrics of genetic diversity (panels a, b and c) and haplotype homozygosity (panel d) showing the signatures of different types of selective sweeps in comparison with the neutral expectations (Fu and Akey 2013).

During the sweep phase a single haplotype is present at high frequency in the population, this signature is characteristic of this type of strong selection process and is called "partial" or "ongoing hard sweep". After the selected allele is fixed in the population the selective sweep is complete, where only a single haplotype dominates and the region gets into a period of relaxed selection in which, with time, starts to accumulate low frequency variants (Figure 5). A clear example of this kind of selection appears when a non-synonymous mutation affects a protein-coding gene and

changes the sequence to a more adaptive version of the protein. Now, the evident contrast of this type of signature with a neutral background makes it one of the easiest marks to detect on a genomic scan.

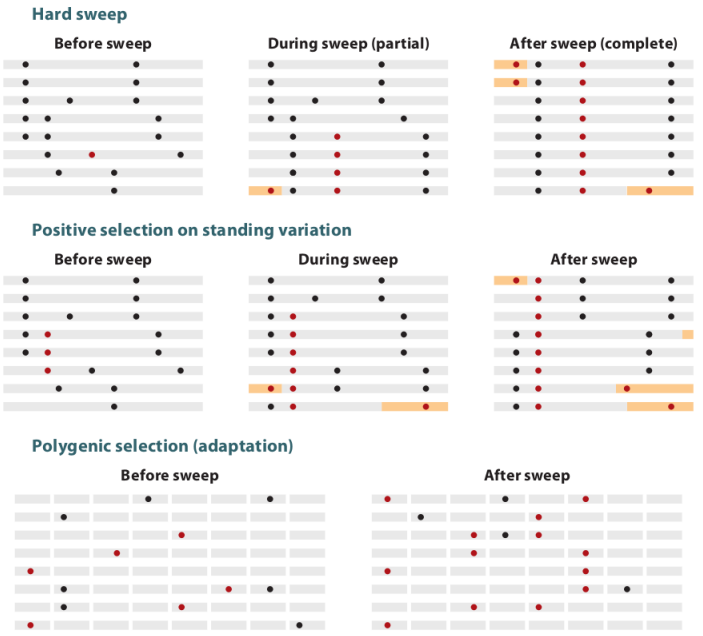


Figure 5. Different modes of positive selection based on the increase in frequency of adaptive variants in one or more selected loci (Fu and Akey 2013).

However, what would happen if the selected allele was already present in the population, at a certain frequency? In these cases, the selected allele is placed in different haplotypic backgrounds and the hitchhiking effect will generate another type of signature, characterized by the presence of different haplotypes at moderate frequency, a *soft sweep*. In this mode of selection, commonly referred to as selection on "standing variation", the wider collection of sweeping haplotypes generates a weaker reduction of haplotype diversity compared to hard sweeps, which leaves a narrower valley of diversity in the selected locus. The level of haplotype diversity depends on the frequency of the variant at the beginning of the selection process; at higher frequencies we encounter a wider

variety of haplotypes that will compound the signature. Other modes of selection might also generate soft sweps, like the selection of more than one *de novo* mutations in the same locus at the same time, where the selected alleles increase the frequency of the different sweeping haplotypes.

Selection might also act on multiple loci simultaneously. Polygenic traits are widely analysed in GWAS studies, where several loci are detected to be significantly associated with a phenotype. However, each of these variants individually account for a small fraction of the phenotypic variation in the population. In these cases, selection processes might systematically affect multiple loci present in standing variation and generate subtle frequency shifts. This frequency fine-tuning is the basis for *polygenic selection* processes. An intuitive example of this mode of selection occurs in quantitative phenotypes, like height, one of the first traits associated with polygenic selection on standing variation (Turchin et al 2012).

The patterns left by hard and soft sweeps are sometimes difficult to differentiate. Depending on the degree of "softness", which is correlated with the number of sweeping haplotypes in the region, a selection process on a *de novo* allele might resemble more a soft sweep than a hard sweep (Messer et al 2012). So where is the limit between these two types of signatures? At the end of the day, genomic scans are just dealing with perturbations of the haplotypic diversity in discrete regions of the genome characterized by the composition of these sweeping haplotypes. It depends on the selection coefficient of the new allele, its starting frequency, the moment when the sweep is detected, the disruption of the haplotypes by recombination, the presence of recurrent new adaptive alleles, among other factors, which determine the identity of the signal, which sometimes is impossible to disentangle. Also, the perturbations on diversity in a certain region might influence the surrounding sites generating signatures that mimic those seen in soft and partial hard sweeps. The *shoulder effect* is a phenomenon that

appears in the vicinity of hard sweeps (on its shoulders) that have completed the selection process. According to Schrider et al 2015, the regions on both sides of the selected locus might be affected by recombination in a way that leaves sweeping flanking haplotypes at intermediate frequencies, a signature similar to those left by soft sweeps. This signature might also be mistaken by the haplotypic pattern in a partial hard sweep, where the selected allele has not reached fixation yet.

Therefore, studies that do not present a previous hypothesis about the selection processes that might be undergoing in a certain region ("hypothesis free" studies), like genomic scans, are likely to encounter numerous cases of misclassified regions under positive selection, either by interpreting a signal of positive selection separated from the actual selected locus or by misidentify the mode of selection. All these factors must be taken into account when interpreting the putative signals of positive selection, and support the candidates under selection with evidence that helps to reject potential confounding or false positives.

Methods to detect genomic signatures of positive selection

Numerous methods and statistical tests have been developed during the last years to detect signals of positive selection in the genomes of natural populations (Vitti et al 2013, Rees et al 2020). The design of these statistical methods relies on the genomic properties of the selection signature, which are correlated with the age of the selection process (Figure 6).

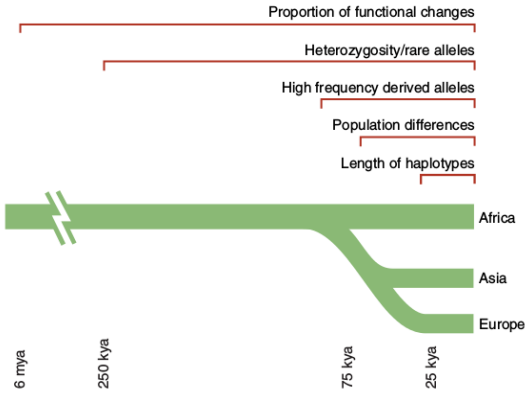


Figure 6. Time scales for signatures of selection that persist over varying time scales (Sabeti et al. 2006).

The comparison of sequences between different species can be used to look for nucleotide changes that alter the protein function, providing evidence of positive selection undertaken millions of years ago. For example, these signatures can be detected by comparing the rate of non-synonymous substitutions per site with the rate of synonymous substitutions per site between two lineages (*Ka/Ks*). The excess of non-synonymous changes would suggest that the gene has been affected by positive selection processes that favour novel protein structures. We can also look for signatures of positive selection within a species. Patterns of low genetic diversity left by the linked effect of selected alleles (hitchhiking effect) are associated with signatures left during the last 200.000 years. These signals can be detected by looking at alterations of the *site frequency spectrum* (SFS) compared to neutral expectations (Figure 7). In these regions the reduction of diversity is accompanied by the emergence, with time, of low frequency variants, and this will be reflected as an excess of rare alleles and fixed or nearly fixed derived alleles. Examples of tests sensible to these kinds of signatures are *Tajima's D* (Tajima F 1989) and *Fu and Li's D* (Fu and Li 1993). During the global human expansion, the geographical separation of different groups of people subjected these populations to different environmental pressures. The adaptation to these new conditions selected phenotypic relevant alleles that appear as highly

differentiated between populations. These differences in allele frequencies provide evidence of positive selection but also might reflect demographic events that took place in these populations separately. There are different methods designed to detect these population differences, like the *Population differentiation* scores (Fst) (Figure 7).

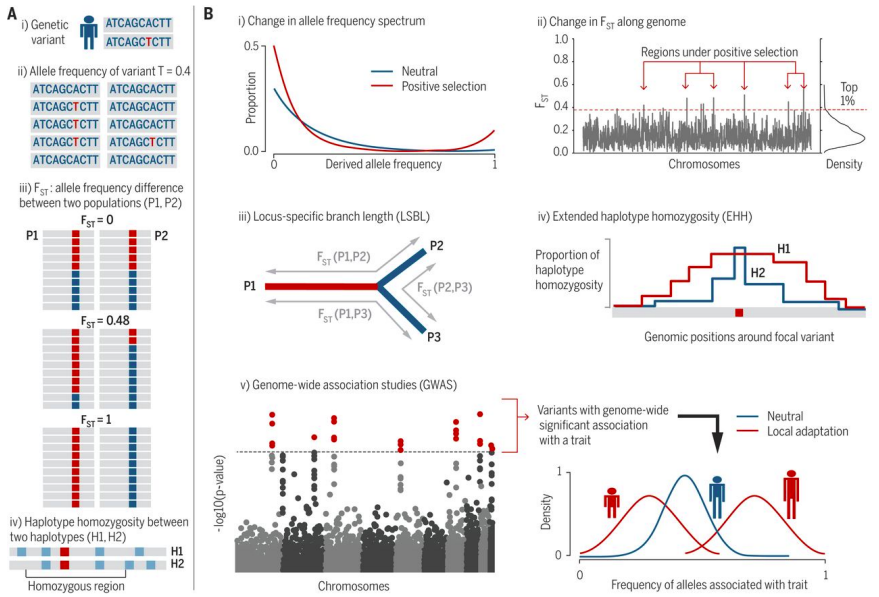


Figure 7. Methods and statistical tests to capture signatures of recent positive selection (Fan et al 2016).

One of the most used methods to detect recent selection (<30.000 years) is based on the detection of long unbroken haplotypes at high and moderate frequencies in population-specific genomes. This is a family of haplotype-based statistics that rely on phased genotypes, this is when the polymorphism alleles are localized in each of the two copies of the chromosome. The basis of this type of test is the calculation of the *extended haplotype homozygosity* (EHH). Introduced by Sabeti et al 2002, this test measures the decay, with distance, of the genetic identity in a collection of haplotypes from a variable position (core allele). This identity starts at 1 in the core allele and decreases with distance at further distances from this

position. In a scenario where an allele is under positive selection and rapidly rises in frequency, like in a hard sweep, the hitchhiking effect takes the linked sites along with the core allele and purges the diversity in the surrounding of the selected allele. In this scenario the decay of the haplotype homozygosity is slower than in the case of an allele that drifts under neutrality, which presents a more variable collection of haplotypes in the region. Now, when analysing positions that are under positive selection, the area under the EHH curve would be greater for the allele that has been selected in comparison with the non-selected or neutrally drifting. This is the foundation of the *integrated haplotype score* (iHS), introduced by Voight et al 2006, which is designed to detect signatures that resemble patterns left by ongoing hard sweeps. In this test, the calculations are made at each variable position and rely on the comparison of the area under the EHH curve for the set of haplotypes that harbour the ancestral and derived alleles as the core position. In a hard sweep, the EHH curve of the allele under positive selection would dominate over the others, indicating the presence of long unbroken sweeping haplotypes where this allele is placed (Figure 8).

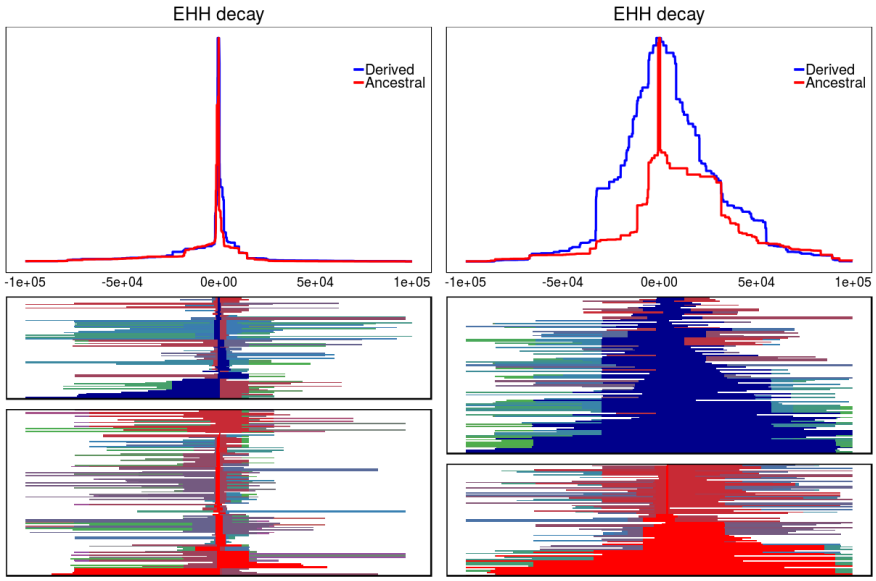


Figure 8. Extended Haplotype Homozygosity (EHH) decay of sets of haplotypes harbouring the ancestral (red) and derived (blue) alleles from a variable position (core SNP). The EHH decay is faster in neutrally evolving SNPs (left) and much slower in SNPs under positive selection (right). The bottom plots show the identity of the haplotypes in the population.

A variation of the iHS test is the *cross-population extended haplotype homozygosity* (XPEHH) described by Sabeti et al 2007, which relies on the comparison of regions between populations. While iHS is suited to detect hard sweeps that are in their way to fixation, XPEHH detects sweeps that are fixed or nearly fixed in a population A in comparison with drifting alleles in another population B. Other methods that rely on the extension of haplotypes in a population are designed to detect both hard and soft sweeps. Ferrer-Admetlla et al 2014 describes the *number of segregating sites by length* (nSL), a haplotype-based statistic similar to iHS with the difference that it does not rely on a recombination map to measure distances from a core allele. Instead, this statistic calculates the length of the haplotype homozygosity from a core allele by counting the number of segregating sites contained in the segment of homozygous haplotypes in the entire sample. In this way, nSL is more robust than iHS in terms of recombination and mutation rate variation. *H12* is another statistic suited to detect soft sweeps that was introduced by Garud et al 2015 and adapted by Torres et al 2018 (*iHH12*) to be more powerful to detect soft sweeps than iHS. This test relies on the assumption that in a soft sweep there are more than one haplotype at moderate frequencies that contribute to the haplotype homozygosity in the region, in contrast to a hard sweep where a single haplotype dominates over the others. Therefore, in order to detect with better power regions under the regime of soft sweeps, the statistic collapses the frequencies of the two most frequent haplotypes into a single class, obtaining a higher haplotype homozygosity score in those regions where there is more than one haplotype at moderate frequency.

Limitations of selection studies

Nowadays the study of local adaptation in humans remains challenging for many reasons. The availability of genomic data, biased towards specific population ancestries, is sometimes not representative for small ethnic groups and indigenous populations. This causes that most genetic studies favour the analysis of populations of certain origin, like Europeans (Sirugo et al 2019). This lack of representation in some human populations not only bias the interpretation of the results but also underestimate the relevance of disease-association studies and evolutionary processes that might shed light on specific events in the history of *Homo sapiens*.

Demography is another factor that can cause distortions of the neutral diversity in the genome. Populations might present an underlying structure based on non-genetic factors like geography, language, religion and social distribution. These might create barriers among individuals and alter the random mating expected in a population, and subsequently generate specific patterns of genetic variation that alter the initial demographic assumptions. Population movements and changes in population sizes can also generate distortions of the genetic variability. Migrations from one population to another can create genetic backgrounds product of different ancestries. This higher complexity in their genetic configuration can lead to situations where the variability within a population is increased while the genetic differentiation with other populations is decreased. Other examples of demographic events are population expansions and bottlenecks. These cases present the particularity that can create distortions of the genetic variability that might mimic the signatures left by processes of natural selection. Due to this increased complexity in the population dynamics, it is important to adequate the demographic models used to evaluate the empirical results in selection studies.

Also, the analysis of selection signatures, carried out by genomic scans, normally lack the phenotypic information that allow to link

the genomic signature with the environmental causes of adaptation. Some evident phenotypes, like Mendelian traits, are easier to associate with the selected allele. On the other hand, variation on gene expression or polygenic traits are less intuitive to associate with the locus under selection behind the hypothetical adaptation.

When performing "hypothesis free" studies, like genomic scans, there is a realistic risk to report a given proportion of false positives due to the influence of confounding factors, which hinders the discovery of true candidates behind actual adaptations.

The human X chromosome

The sequence of the human X chromosome was first published by Ross et al 2005 and recently completed by Miga et al 2020, where the authors used ultra-long nanopore reads to resolve gaps at the centromere and two segmental duplications. In mammals, the X and Y chromosomes derive from a pair of homologous ancestral chromosomes that diverged from each other during their evolution approximately 180 Myrs ago (Abbott et al 2017). During this process the ancestral Y chromosome degenerated and lost most of its content, presenting recombination only with two small regions at the tips of the X chromosome arms called *pseudoautosomal* regions (PAR1 and PAR2) (Figure 9). However, recent studies on genetic diversity in these X-Y recombining regions show that they might not present strict boundaries as considered to date (Cotter et al 2016).

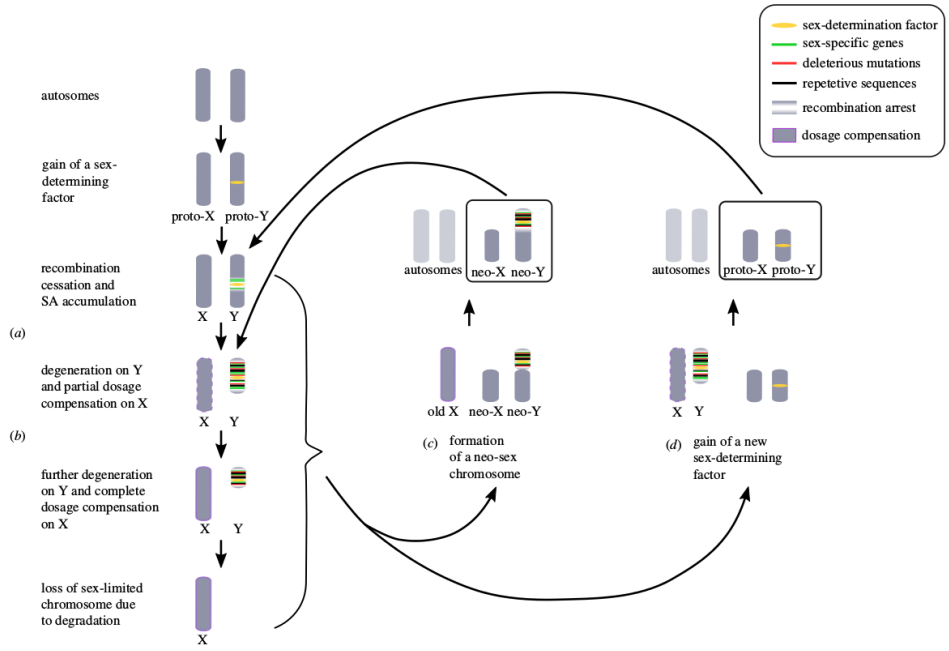


Figure 9. Schematic representation of the evolution of sex chromosomes, illustrated in a male heterogatmetic system (Abbott et al 2017).

Regarding the gene content, the human X chromosome is enriched in brain-expressed and cancer-testis genes (CT genes). Around 10% of X chromosome genes are CT, a class of genes found expressed in a great number of cancer types and in normal testis. In contrast with autosomal single-copy CT genes, the CT-X genes are organized in families and are thought to have originated from duplication events. The expansion of the CT-X genes seems to be in concordance with their rapid evolution (Stevenson et al 2007), with families of genes like *MAGE* and *SPANX* reported of being under strong positive selection (Kouprina et al 2004), which suggests the evolutionary advantages of this kind of genes in males. Also, it is of interest to note that the X chromosome presents the largest gene of the human genome, the *dystrophin* (*DMD*), spanning more than 2.2 Mb and located in the locus Xp21.1 is responsible of the Duchenne and Becker muscular dystrophy (Duan et al 2021).

The differences in the X chromosome gene dosage between the XX females and the XY males are compensated by the random inactivation of one of the chromosomes, first hypothesized by Lyon MF 1961. This mechanism, common to all mammals, takes place at the early stages of female embryonic development. The X chromosome inactivation is directed by the X inactivation center (XIC), a cluster where various non-coding genes control the transcriptional silencing of the inactivated chromosome (Xi). This process is triggered by the *XIST* gene, a long non-coding RNA that accumulates and coats *in cis* the future inactivated chromosome (Barr body), serving as a scaffold to other protein complexes that performs epigenetic modifications that finally cause the gene silencing (Figure 10) (Lu et al 2017).

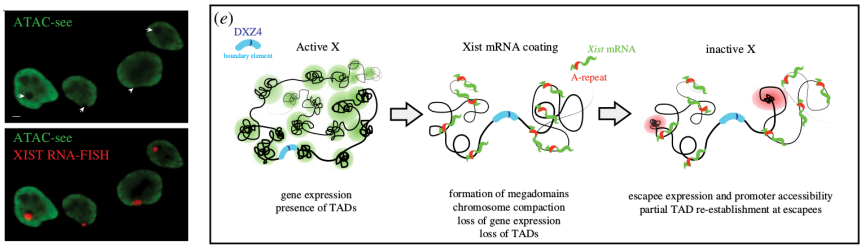


Figure 10. Detection of XIST with RNA-FISH technique in the inactivated X chromosome (Xi). Coating process of Xi during the formation of the Barr body and triggering of the transcriptional silencing (Fu et al 2017).

Female tissues are therefore quimeras where each cell harbours one of the X chromosomes inactivated. However, during the last years, several studies have demonstrated that this inactivation is not complete for all the genes in the Xi. Some genes *escape* at a certain degree from this inactivation, generating therefore sex-biased expression patterns that could lead to dimorphic traits and sex-specific diseases (Balaton et al 2016). The inactivation status of a gene can be established by using different strategies, showing that escape genes tend to cluster together and are enriched in the pseudoautosomal region 1 (PAR1) (Balaton et al 2015). One of the most recent studies on the X chromosome inactivation states that around 23% of genes escape from this process (Figure 11)

(Tukiainen et al 2017). Among other conclusions, the authors show that escape genes present different expression bias depending on the region they belong to. In the pseudoautosomal region 1 (PAR1) escape genes are mostly expressed toward males, while the genes located in the non-pseudoautosomal region (nPAR), these are the X-specific, present a female-biased expression. They also claim, by analysing different types of data, that the incomplete X inactivation is generally maintained and tightly controlled across tissues, but present numerous cases of genes whose expression is variable across populations and tissues, thus likely generating phenotypic diversity in humans. Whether or not escape genes are subjected to adaptive processes is still under discussion. Park et al 2010 reported signals of strong purifying selection (Ka/Ks ratio) in escape genes of primates. Among their conclusions, they suggest that this signal is mainly driven by escape genes that present an homologous gene in the Y chromosome, which evolve like autosomal genes and therefore are subjected to the dominance of the other allele. However, no other studies on adaptive selection have been performed on human escape genes, where events in the recent history of human populations might have left other kinds of signatures.

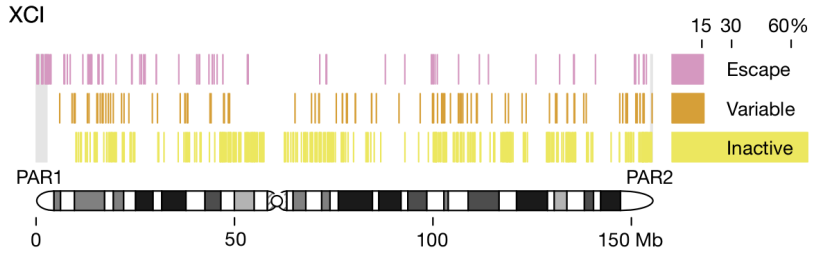


Figure 11. Inactivation status of the human X chromosome genes (Tukiainen et al 2017)

The formation of the XX/XY sexual system in mammals has led to differences in the inheritance patterns of the X-linked genes compared with the autosomes. The hemizygosity of males, this is the presence of only one copy of the X chromosome, imply that, in

the course of human evolution, the human X chromosome spends one third of the time in males and two thirds in females, which generates a decompensation of the mutation rate between the sexes. Spermatogenesis of males presents a higher rate of cell divisions, where DNA replication is a major source of mutations, than the oogenesis of females. However, the lower amount of time that the X spends in males drives a lower mutation rate in comparison with the autosomes. These differences yield to differences in the rate of molecular evolution since the neutral divergence is lower for the X-linked loci (Vicoso and Charlesworth 2006). The hemizygosity of males also leads to a higher exposure of mutations in the X chromosome. This is the reason why the number of X-linked diseases is disproportionate in comparison with the whole genome. Also, the phenotypic consequences of recessive mutations affect directly to their evolutionary fitness and are more affected by selection pressures in males than in females, where they are hidden by the presence of the other allele. Therefore, recessive or partially recessive beneficial mutations are more easily fixed, while recessive deleterious mutations are more efficiently purged from the X chromosome in comparison with autosomes.

This effect on the evolutionary dynamics of the X-linked loci is known as the *faster-X effect* and states that processes of positive and negative selection are more efficient in the X chromosome than in autosomes. In humans, the consequences of this effect is seen when comparing the divergence rate with chimpanzees, where the ratio of Ka/Ks is overly higher for the X-linked genes than in autosomes (Lu and Wu 2005). Furthermore, if we consider male-expressed genes like testis-specific, the effect is more pronounced due to the exclusive exposure of the mutation effect. Numerous studies have reported evidence of the faster-X effect in the human lineage. Veeramah et al 2014 used a MK-based framework that measures the proportion of fixed nonsynonymous substitutions to demonstrate that either positive or negative selection processes are significantly enhanced in the X when compared with autosomes. In Hammer et al

2010 the authors reported a correlation between the X/Autosome diversity ratio and the genetic distance measured from genes in human populations. This tendency exhibits a lower diversity near X-linked genes in comparison with autosomes, a tendency that increases with distance and reflects an enhanced selection effect that leaves a more pronounced signature of low variation around X-linked genes (Figure 12).

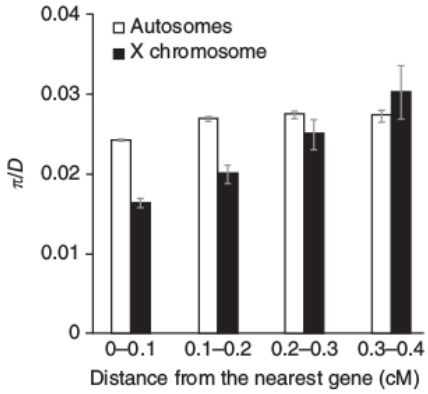


Figure 12. Nucleotide diversity as a function of genetic distance from genes (Hammer et al. 2010)

A similar approach was used by Arbiza et al 2014 to analyse the different forces that shape human diversity in the X and autosomes in a greater number of populations. In the study, the ratio of X/A diversity is seen proportional among the populations of the same continental group, however this ratio is found reduced in non-African populations compared to Africans. This reduction of diversity is not explained by selection forces (diversity as a function of genetic distance), since they are comparable between populations. Other factors like demography (Out-of-Africa event) are partially behind these patterns, however the authors conclude that to fully explain these discrepancies more specific factors must be considered, like male-dominating migrations or specific social changes that influence reproductive success.

In the study of selection signatures in the X chromosome, different authors have reported specific patterns that differ from those seen in the autosomes. Nam et al 2015 describes wide regions of low diversity in different great ape species, including humans, that evidence the presence of independent strong hard sweeps in specific locations of the X chromosome. These extreme signals of low diversity, in some cases spanning several megabases, are not seen in autosomes and are partially overlapping between species, which suggest a common factor that leads to these selection signatures. As the authors suggest, the enrichment of testis-expressed ampliconic genes in these regions might point to processes of meiotic drive (Figure 13A). In another paper (Dutheil et al 2015), the authors support these findings by identifying regions of low *incomplete lineage sorting* (ILS) in different great ape species that are compatible with recurrent selective sweeps. These regions are enriched in ampliconic genes that are suggested to be positively selected by the effect of meiotic drive. This, as the authors suggest, might be behind hybrid incompatibilities between diverging populations in the lineage of great apes, which could derive to speciation processes (Figure 13B). Another study on positive selection was comprehensively carried out in different human populations by Casto et al 2010. In this study the authors identify several regions with high population differentiation scores and associated with signatures of hard sweeps captured by haplotype-based statistics. A significant number of these signals fall within cancer-testis genes, supporting the idea of selection in spermatogenesis-related genes. Other regions were outliers of previously reported genes like the *dystrophin* (Figure 13C) (*DMD*). However, their main results focused on three outlier regions that reflect patterns of positive selection in genes like *EDAR2*, related to the well-known gene *EDAR*, found under positive selection in Asian populations (Bryk et al 2008).

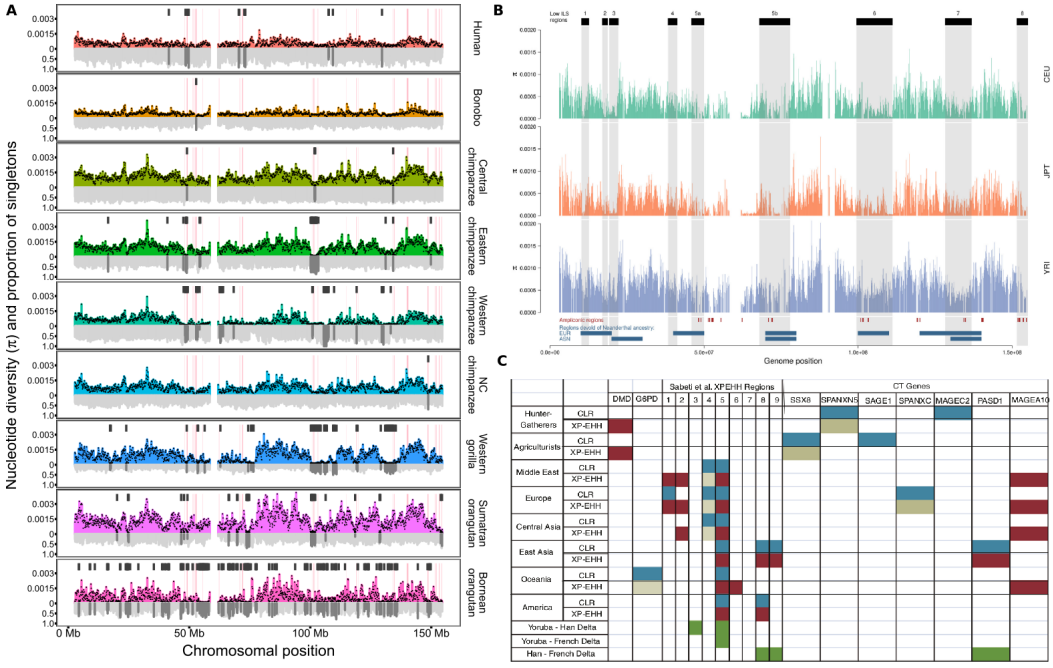


Figure 13. (A) Nucleotide diversity patterns of the X chromosome across different great ape species. Grey top squares indicate the presence of extreme hard sweep signatures (Nam et al. 2015). (B) Nucleotide diversity landscape in three human populations indicating the overlap between low incomplete lineage sorting regions and diversity deserts (Dutheil et al. 2015). Table of candidates under positive selection in the X chromosome and supporting selection signatures analysed in Casto et al. 2010.

The evolution of the regulatory genome

The description of the *lac operon* by Jacob and Monod in 1961 supposed the first step in the study of gene regulation with the discovery that genes can be regulated by other molecules. In the case of the lac operon, Gilbert and Müller-Hill argued later in 1966 that such regulatory elements are other proteins that bind to the lactose substrate and modulate the behaviour of the gene. These studies extended the concept of proteins as control agents of gene expression and not only enzymes, which set the field for the development of the transcription factor paradigm of gene regulation (Chen and Rajewsky 2007). The emergence of this concept was also

the seed for the idea that genes might present multiple combinatorial ways of being modulated, which increased the complexity of how genetic information is used and managed by the genome.

Soon, researchers started to think about the correlation of the genomic content of organisms and their so-called "phenotypic complexity" as a way to define the evolutionary sophistication of certain species. This correlation did not appear to fit properly with the expectations: how is it possible that organisms with similar amounts of genes present different morphological complexities?

The discovery of gene regulation mechanisms in the second half of the twentieth century brought to light the idea that, not only the gene diversification between species, but the multiple combinatorial patterns of such regulation is behind the potential diversity of organism development and evolution. Britten and Davidson were the first to propose in the 70's that phenotypic diversity between organisms might be mainly driven by changes in regulatory regions and not only in the protein-coding sequences of genes. This theory was extended by King and Wilson in 1975 to the differences between humans and non-human primates, suggesting that the question "*What makes us human*?" would be answered in the light of regulatory changes as the main fuel for evolution. In the early 2000's, a series of studies based on the comparison of gene sequences that control differential traits between species, like the presence of trichomes in Drosophila species or the morphological differences between marine and freshwater fishes, suggested that no coding changes are behind such differences, so *cis*-regulatory changes must be the most likely cause. These studies reinforced the idea that regulatory changes are the main drivers for morphological evolution. This new perspective of how evolution works at genomic level meant a dramatic crash among evolutionary biologists from both sides of the question, as it was illustrated in the controversy around these studies (Pennisi E. 2008). However, the idea that

regulatory changes are behind most evolutionary innovations started to displace the dominant concept of protein-coding changes as the main evolutionary fuel. In this context the *evo-devo* paradigm started to grow trying to respond to evolutionary questions from a developmental perspective.

The structure of gene regulation

The typical transcriptional unit of unicellular eukaryotes, like yeasts, depends on DNA sequences located at the 5' side of the *transcription start site* (TSS). The main region that triggers the transcription of a gene is the core promoter, which typically presents a sequence called *TATA* element, which serves as binding site for the *TBP* proteins (*TATA*-binding protein). However, the transcription activation only by the core promoter is weak and needs the presence of other close and distal regulatory sequences. In some cases, the transcription of the gene also requires the presence of distal regulatory regions. These are activating sequences that require the binding of other regulatory proteins (transcription factors) to promote the transcription of the gene (Levine and Tjian 2003).

In the case of metazoans, the anatomy of transcriptional units is more complex and responds to the multiple combinations of elements and regions in the tissue and cell-specific regulatory programs. In these genes the initiation of transcription by the RNA polymerase II (Pol II) depends on more sequence elements apart from the *TATA*-containing core promoter, like the initiator element (*INR*) and the downstream promoter element (*DPE*). The *cis*-regulatory elements (cREs) that regulate the gene transcription are more diverse than in the case of unicellular organisms. Enhancers are elements that contain short DNA motifs that bind transcription factors which, by recruiting co-activators and co-repressors, promote gene transcription. They can be found at the 5' and 3' regions, as well as in introns, which are normally

constituted by sequences of ~500 bp in length and harbour binding sites for several transcription factors. Other regulatory regions are called *tethering elements* and hold binding sites for factors that recruit distal enhancers to the core promoter. Insulators are another type of element that create regulatory barriers and prevent the cross-interaction between different regulatory domains and the inappropriate regulation of genes by neighbouring enhancers. This enhancer-promoter specificity is also achieved by the sequence elements located in the promoters, like the *TATA*-containing promoters or *DPE*-containing promoters, which are activated by different enhancers. This is how long-range regulation is achieved when activating the expression of distal transcriptional units. However, the decoupling of enhancer-target contacts leads to rerouted regulatory interactions and novel patterns of expression. The distribution of such complex collections of regulatory elements can span distances of hundreds of kilobases in mammals and are responsible for the control of the transcription of a single gene (Figure 14).

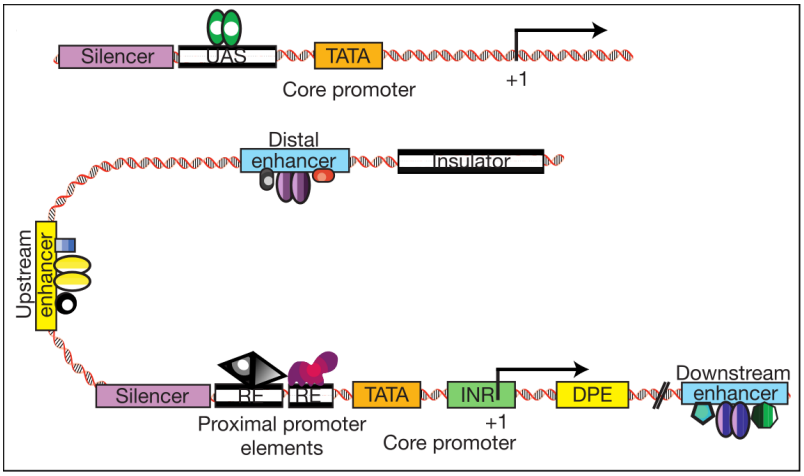


Figure 14. Anatomy of the canonical transcriptional units of unicellular eukaryotes and metazoans (Levine and Tjian 2003).

The role of enhancers in complex organisms

The first described enhancer was discovered in the *SV40* virus genome (Banerji et al 1981). Later on, in 1983, this kind of regulatory element was described in animals (Banerji et al 1983) and, from then, the study of their biochemical and functional properties expanded to other organisms, preparing the ground for the study of the evolutionary dimension of gene regulation. Active enhancers are found in regions devoid of nucleosomes, which allows for the access of transcription factors to DNA binding motifs. Also, the sides of these regions are typically characterized by the presence of the histone marks H3K4me1 and H3K27ac in their amino termini, product of post-translational modifications. Enhancers have demonstrated to be independent of the distance and orientation of their target genes, and also of the genomic context where they are located. Also, distal enhancers have shown to interact with their target promoters by looping, creating interactions spanning hundreds of kilobases. The regulatory role of enhancers in controlling the transcriptional levels of their target genes shows additive and redundant properties, this means that several enhancers, in combination with their transcription factors, are able to modulate the gene expression as a result of their combined activity (Shlyueva et al 2014) (Figure 15).

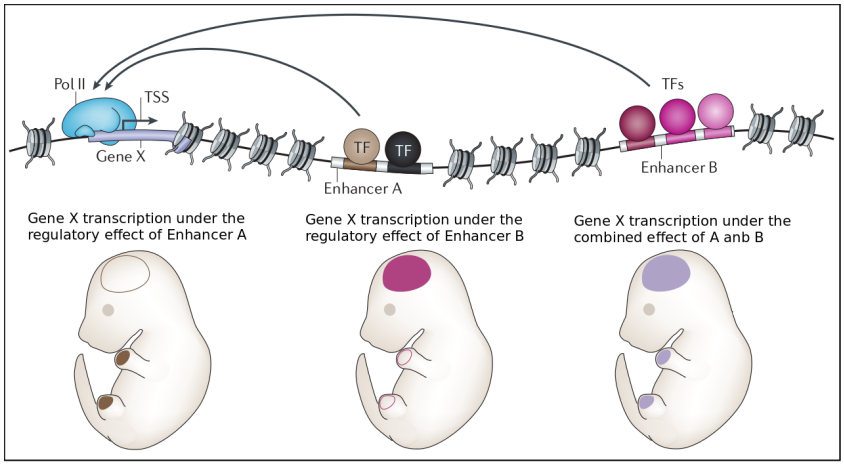


Figure 15. Example of the combined regulatory activity of two different enhancers on the transcription of the Gene X (Shlyueva et al. 2014).

The characterization of regulatory elements in the human genome

Massive efforts have been made to characterize in depth the regulatory signals exhibited by DNA elements in the human genome. One of the projects that have led this work is the *Encyclopedia of DNA elements* (ENCODE), which integrates in its last release (phase III) almost a million *cis*-regulatory element candidates in the human genome and more than 300.000 in mouse (ENCODE Project Consortium, 2020). During the development of this database, which started in 2003 with the Pilot Project of ENCODE, the collection of massive parallel sequencing techniques and biological samples used to identify regulatory elements have increased significantly, encompassing in the current release 503 biological cell and tissue types from more than 1369 biological samples. The integration of different types of assays has allowed the mapping of different genomic signatures involved in transcription factor occupancy, DNA accessibility, 3D chromatin interaction, among others.

One of the new features of the last ENCODE release is the development of the Annotation and Mapping of promoters for the analysis of Gene Expression (RAMPAGE). This new approach was born due to the necessity of giving response to the continuously expanding and highly diverse sea of RNA sequences in the transcriptomes. They seek to identify transcriptional start sites (TSS) of RNA sequences, measure the expression of promoter-specific RNA species and characterize the isoforms of such genes. Among other improvements and expanding datasets, ENCODE invested lots of efforts to integrate DNA accessibility and chromatin modification data to create a massive registry of candidate *cis*-regulatory elements (cREs) in the human genome. The different biochemical signatures exhibited by regulatory elements (enhancers, promoters and insulators) and their activation states were integrated, together with the annotation of TSSs,

according to a classification scheme, resulting in the annotation of 926,535 human cCREs (7.9% of the whole genome). According to the support of two different experimental evidence (high DNase signals that indicate accessible DNA and one of the ChIP-Seq signals (H3K4me3, H3K27ac, or CTCF)), this classification divided the cCREs in three main annotation groups: enhancer-like signatures (ELSs), promoter-like signatures (PLSs) and CTCF elements. The ELSs are characterized by the presence of high DNase and high H3K27ac signals and, depending on the proximity to the TSS, they present low relative H3K4me3 signal (proximal, closer than 2 kb from the TSS) or none at all (distal, further than 2 kb from the TSS). On the other hand, canonical PLSs present high DNase and high H3K3me3 within the 200 bp of an annotated TSS. Other signatures presenting the same peaks are thought to be non-canonical promoters or other kinds of regulatory elements. CTCF signatures together with DNase sites identify regions belonging to insulators or with looping functions where the protein CTCF participates (Figure 16). This registry of cREs is displayed in a browser-like webtool in the publicly available ENCODE application SCREEN, where they are integrated with other types of data, like transcript expression profiles, chromatin looping signals or transcription factor binding peaks, among others.

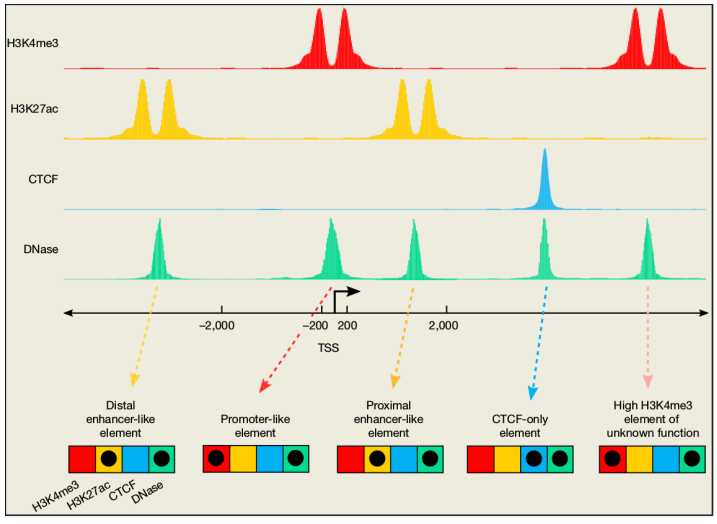


Figure 16. Cis-regulatory elements (cREs) classification scheme by ENCODE, the different combinations of DNase and ChIP-Seq signals provide the identity of the candidate elements (ENCODE Project Consortium et al 2020).

Tissue-specific and housekeeping gene regulation

The comprehensive annotation of *cis*-regulatory elements in the human genome, which greatly exceeds the number of protein coding and noncoding genes, shows the extensive landscape of regulatory possibilities that gene expression patterning exerts in cell and tissue contexts. From this extensive repertoire of regulatory sequences, each regulatory program selects a specific set of elements to carry out the modulation of gene transcription. However, the way these regulatory elements are chosen and the way they function to orchestrate their combined regulatory functions that confer the tissue identity and functionality is still under profound study.

Many lines of evidence indicate that the selection of tissue-specific sets of enhancers are directed by protein and signal-specific priming events in the cell. The model of *pioneer factors* describes the presence of DNA recognition motifs, associated with cell-specific enhancers, susceptible to be bounded by lineage-determining transcription factors (LDTFs) in the compacted chromatin (Zaret et al 2008, Heinz et al 2015). This binding triggers the opening of chromatin and initiates the activation of enhancers, which induces tissue-specific transcription programs. In essence, the action of these LDTFs precipitates the transition of enhancer elements from their closed (inactivated) state, when they are buried within the compacted chromatin, to a primed or poised state, where their DNA sequence is accessible, in which they start to be available to be bounded by other transcription factors. However, the action of these LDTFs by themselves sometimes is not sufficient to initiate regulatory programs of a certain cell type. Combined with these LDTFs, there is also the action of other kinds of factors dependent

on intra- and extracellular signalling events, like the members of nuclear receptor families. These signal-dependent transcription factors (SDTFs) may act as complementary factors for the LDTFs, binding in previously initiated enhancer regions by the latter (Samstein et al 2012, Heinz et al 2015). In this sense, the joint action of these two kinds of factors is hierarchical, in which the initiation of tissue-specific enhancer activation is firstly executed by the LDTFs. However, the action of SDTFs might also be independent of LDTFs, in the sense that they can trigger the *de novo* selection of enhancers. In summary, the combined interaction of these two kinds of factors and their binding to specific DNA sequence motifs is responsible for the selection of the enhancers that guide the execution of tissue-specific regulatory programs (Figure 17).

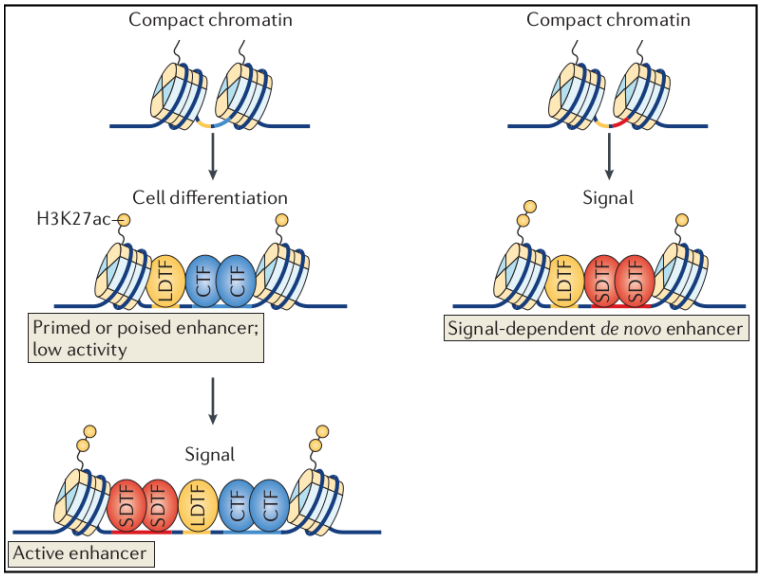


Figure 17. The action of lineage-determining transcription factors (LDTFs) and signal-dependent transcription factors (SDTFs) in the activation of tissue-specific enhancers (Heinz et al 2015).

The concept of *housekeeping gene* has been around for various decades already. Initially they were defined as genes that are devoted to basic functions, necessary for the maintenance and

survival of the cell. Therefore, they are expected to be expressed in all tissues under normal conditions, regardless of their specific identities, functions and external signals. Their genomic features are also different from those that are tissue-specific. For example, they are generally more compact, with shorter introns and exons, and might harbour different profiles of other elements, like transposons. The early detection of housekeeping genes allowed to compose lists that served as a guide to identify the basal functions of the cell, as well as internal controls for experimental assays that need a constantly and widely expressed gene as a reference. However, the early technology of microarrays presented significant limitations that were partly resolved by the development of high throughput sequencing technologies such as RNA-seq. In the post-genomic era, this technology allowed to quantitatively measure the expression of genes at a higher accuracy, regardless of their prior annotation, and in a more diverse collection of tissues. This permitted the discovery of a larger number of widely expressed loci, incrementing the catalogs of housekeeping genes, but also the identification of basal low expression levels throughout the genome. This means that the expression of genes in all tissues was not an accurate proxy to define a housekeeping gene, but it had also to consider the level to which it is expressed and a low expression variability across different tissues (Eisenberg et al 2014).

Although the regulation of both housekeeping and tissue-specific genes in the human genome is poorly understood, different studies have reported differential characteristics of this regulation. For example, in Zabidi et al 2014 the authors reported that enhancer-to-core-promoter specificities drive the differential regulation of housekeeping and developmental tissue-specific genes. They used self-transcribing active regulatory region sequencing (STARR-seq) constructs to identify the regulation of these two types of promoters across thousands of enhancers described in *Drosophila melanogaster*S2 and ovarian somatic cells. The comparison of the regulatory outcome in these promoters

yielded two sets of different enhancers that presented a low overlap between them, suggesting their specific interaction with the two promoters. These two groups of enhancers showed differences in their genomic distribution, housekeeping-specific enhancers (hkCP) were highly proximal to the TSS and located near genes enriched in basic cellular functions, while the tissue-specific enhancers (dCP) were mainly located in intronic regions and next to genes enriched in cell-type functions (Figure 18). The authors concluded that this differential regulation was due to sequence specificities of the core-promoters, whose regulation is mainly preferred by one of the two types of enhancers.

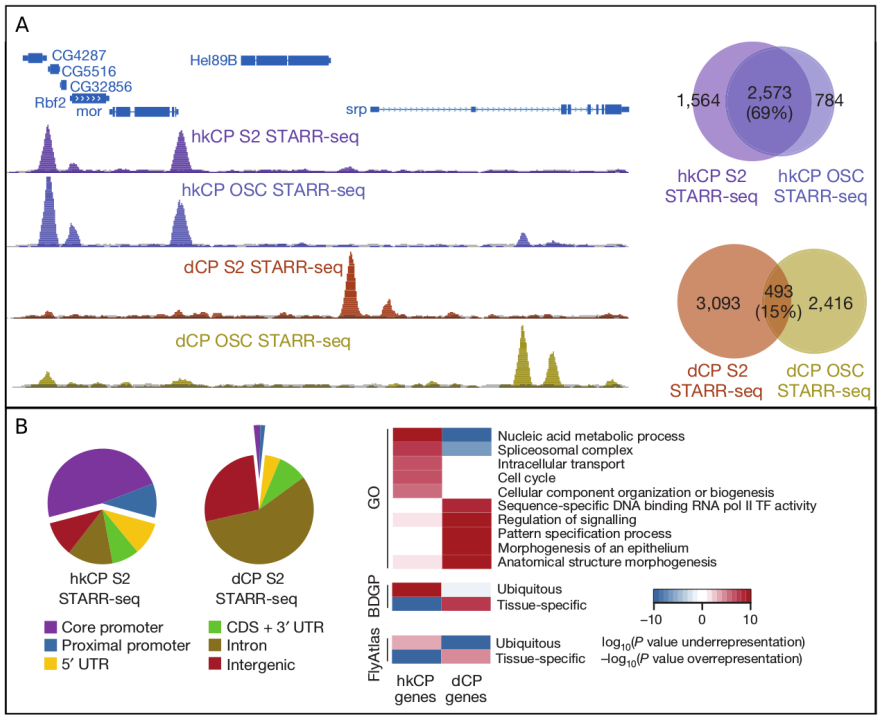


Figure 18. Identification of housekeeping and tissue-specific enhancers in different contexts (S2 and OSC). (A) High overlap between hkCP enhancers and low overlap between dCP enhancers in two different tissues. (B) Genomic location and GO enrichment of near genes to the two groups of enhancers (Zabidi et al. 2014)

The evolutionary role of regulatory elements in the human genome

The advent of new generation sequencing technologies has made possible the development of the regulatory field in increasing the potential discovery of regulatory changes between different species. However, this development has been accompanied by an inherent difficulty of characterizing regulatory variation in comparison with protein-coding changes. The latter are easier to analyse and interpret since, once known the function of the gene, the mutation can be linked more straightforwardly to the trait. On the other hand, DNA regulatory changes are more difficult to pinpoint and trace back to the trait or phenotypic advantage.

In the study of regulatory variation two different strategies can be applied. Researchers might first identify the genomic sequence changes present between, for example, humans and chimpanzees, and then evaluate their potential functional repercussions. This genotype-to-phenotype strategy is rooted in the comparative genomics field which confronts the problem of identifying the functionally relevant changes that might be involved in the human-specific innovations. In the landscape of human-specific changes we might face the distribution of the effects illustrated in Figure 19. Most of these variants involve changes that do not present any biological meaning due, for example, to neutral substitutions or alterations in elements like transcription factor binding sites (TFBS) or chromatin accessibility regions that are compensated by other regulatory mechanisms and, therefore, they do not generate an appreciable change. On the other side of the distribution, we find the minor fraction of these changes that generate large regulatory and expression changes and might be subjected to positive selection processes.

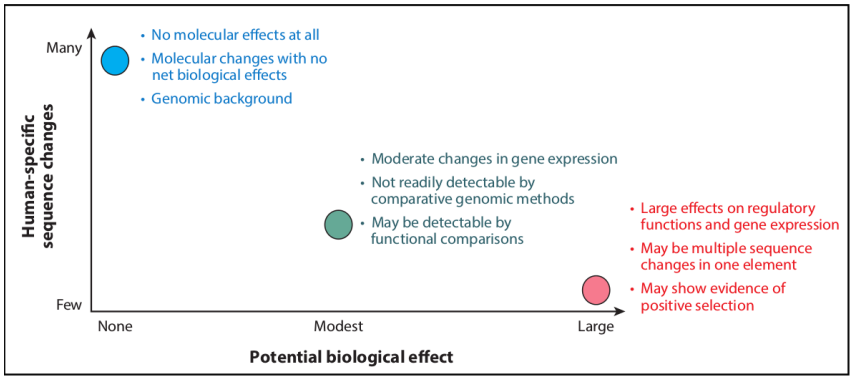


Figure 19. Hypothetical distribution of gene regulatory changes and their effect in the human genome (Reilly and Noonan 2016).

The first studies on human-specific regulatory changes were based on the comparison of multiple genomes and the identification of highly conserved regions in vertebrates with an accumulation of a remarkable number of changes in the human genome. These human accelerated regions (HARs) are mostly noncoding and are enriched near genes controlling developmental processes, like neurogenesis, and therefore affecting regulatory regions like enhancers that modulate the expression of such genes (Prabhakar et al 2006). Several studies were conducted to identify the regulatory role of these HARs, like in Kamm et al 2013, where the authors identified a large number of highly accelerated regulatory regions affecting the human gene *NPAS3*, involved in brain development, and therefore might give rise to human-specific cognitive traits. Although a large number of HARs have been identified to date, it is unlikely that they account for most of the human-specific traits. Moreover, it is still challenging to identify the specific traits in which they are involved without integrating other strategies, like experimental screenings.

Another strategy used in the analysis of regulatory variation among species is based on the identification of the phenotype and the consequent association with the genetic change (phenotype-to-genotype). This phenotype-directed strategy is based

on the development of multiple sequencing techniques that aim to measure in a quantitative way the gene expression and regulatory activity of specific cells and tissues. Gene expression methods, like the analysis of microarrays or the more versatile RNA-seq, provided the opportunity to characterize the expression profiles of different tissue types and compare them across species. The activity of regulatory regions can be measured with techniques like ChIP-Seq, which combines chromatin immunoprecipitation with high throughput sequencing in order to analyse histone modifications associated with enhancers and promoters, or DNase hypersensitivity techniques, which sought to identify chromatin accessibility regions as a proxy for transcription factor binding in regulatory regions. All these techniques aim to characterize the identity of different cell and tissue types, since the expression and regulatory profiles serve as fingerprints of the tissue-specific functionality, and compare them across different biological states, individuals and species. These are the foundations of the comparative functional genomics field.

The comparison of tissue-specific expression profiles among different primate species, including humans, provides an important source of insights into the evolutionary dynamics of gene regulation. The comparison of expression profiles conducted by diverse studies showed that gene expression patterns are overly conserved among phylogenetically related species. This evolutionary constraint is generally maintained across different tissues, although the divergence exhibited in some is greater than in others. The human brain has been the main focus of this kind of analysis. Although the primate brain presents stable gene expression levels, different studies have found particular deviations of such patterns that might reflect evolutionary lineage-specific innovations at regulatory level. In Brawand et al 2011, the authors used RNA-seq-based transcriptome profiling to characterize six organs across different mammalian species. They found different evolutionary rates at expression level, including specific differences in the X chromosome and primate brain, among others. The

differences in primate brain were based on expression shifts of functionally related genes, where the human lineage shows a specific increase in expression of genes involved in neural connectivity of the prefrontal cortex, suggesting the involvement of these regulatory changes in human cognitive evolution (Figure 20). Another source of transcriptional variation emerges from differential alternative splicing across species. In this matter, studies have reported that the transcription architecture of genes is generally maintained across different tissues of the same species, but highly variable among homologous tissues in different species (Young et al 2015).

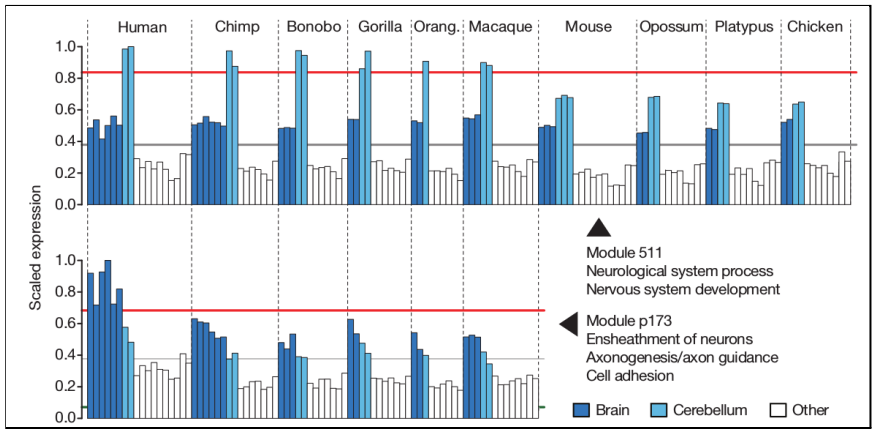


Figure 20. Lineage-specific expression shifts of modular brain-related genes (Brawand et al. 2011).

Comparative studies on epigenetic signatures, like transcription factor binding sites, histone modification levels and chromatin accessibility also serve to identify the differential regulatory programs that are behind the evolutionary diversification among different species. For example, differential studies on signatures like DNase hypersensitive sites locate lineage-specific epigenetic signatures near genes that are differentially transcribed between humans and other non-human primates. This regulatory landscape reveals signatures that are gained and lost in specific lineages and point to derived functions in each species (Gittelman et al 2015).

Comparative studies focused on early developmental stages try to identify the lineage-specific signatures responsible for the morphogenesis differences across different species. In the case of primates, these studies are a clear example on how the diversification of regulatory regions were determinant on the evolution of *Homo sapiens*. In Prescott et al 2015, the authors use epigenomic profiling to analyse the diversification of *cis*-regulatory elements in cranial neural cells from human and chimpanzee, providing a collection of enhancers that present species-specific activity and are candidates for the craniofacial diversification of higher primates. Among other results, the authors measured the genome-wide enrichment of H3K27ac marks, a typical signature of enhancer activity, in orthologous enhancers between human and chimp, reporting a remarkable number of species-biased elements. These elements were found flanking genes whose expression is also biased among primates and are involved in facial morphogenesis (Figure 21).

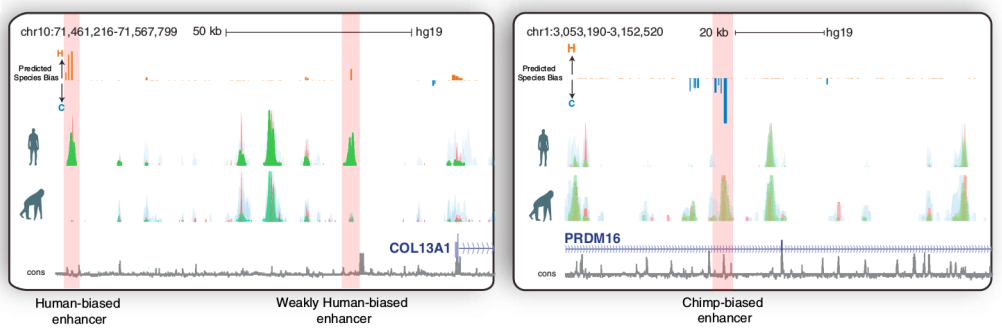


Figure 21. Examples of human and chimp-biased enhancers showing H3K27ac signals that indicate the activity of the region (Prescott et al 2015).

Small regulatory RNAs: miRNAs

The idea of RNA molecules as mere carriers of genetic information from DNA to proteins has long changed in the last sixty years. The

central dogma of biology initially proposed in 1958 by Francis Crick (Crick FH 1958) established a one directional flow of genetic information to their destiny in the function of proteins. It was not until the 50's that RNA biology started to gain more protagonism in the cellular mechanisms between genes and proteins. The roles of RNA molecules were upgraded not only to carriers of this information (mRNA) but also to infrastructural sequences (tRNA) essential for protein synthesis and structural compounds of ribosomes with the discovery of rRNAs. In the following years (Figure 22), the discovery of new classes of small RNA molecules implicated in the machinery of alternative splicing of protein-coding genes (snRNAs, snoRNAs) and the advent of the complexity of heterogeneous nuclear RNAs (hnRNA) led to the proposal by Britten and Davidson of the existence of RNA-based regulatory networks in complex organisms (Britten and Davidson 1969). In the 80's, the regulatory role of RNAs was complemented by the discovery of their catalytic properties, which situated their participation in cleavage reactions at post-transcriptional level, among others. The first indication of the existence of miRNAs was the discovery in 1993 of the loci *lin-4* and *let-7* (Lee et al 1993, Reinhart et al 2000), and their regulatory role in the development of the nematode *Caenorhabditis elegans*. At first sight these small RNAs appeared as mere curiosities in the landscape of regulatory sequences that began to be discovered in molecular biology, without the impression that they take part in a much wider, diverse and relevant group of regulatory players. It was not until 1998 that miRNAs were upgraded to the position they belong as regulatory sequences, in this year the complete picture of how miRNAs act was revealed with the discovery of the RNA interference pathway (RNAi) in plants and *C. elegans* (Fire et al 1998). This process was described as a silencing mechanism driven by double stranded RNAs (dsRNA) that are processed into short interfering RNAs (siRNAs) able to perform regulatory activities at transcriptional and post-transcriptional level. The detection of naturally endogenous dsRNAs as stem-loop sequences, together with the protein

machinery responsible for their processing (*Dicer* and *Drosha*) and function (*Argonaute* proteins), finally confirmed the hypothesis that miRNAs are a product of a maturation process that participate as interfering sequences in gene silencing regulatory processes.

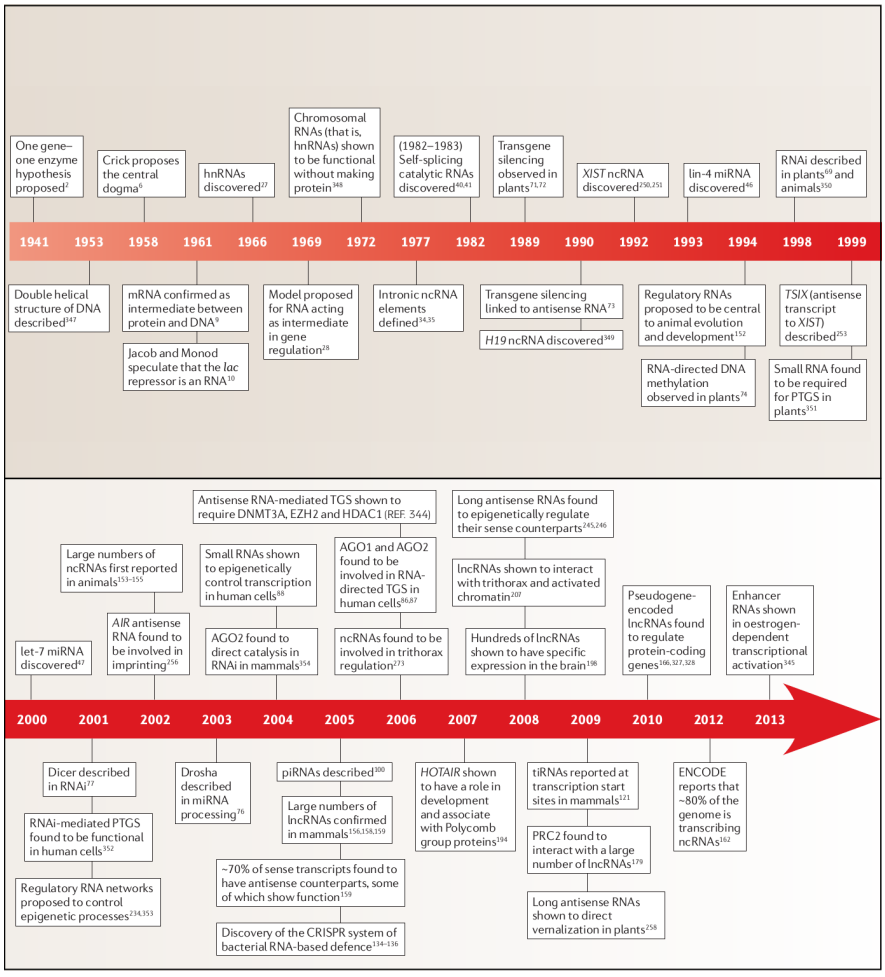


Figure 22. Timeline showing the main discoveries in the field of gene regulation and the acknowledgment of the regulatory roles of RNA sequences (Morris and Mattick 2014).

Biogenesis and function of miRNAs

The discovery of the *let-7* and *lin-4* loci by Ambros and colleagues lead to the description of miRNAs as short single-stranded RNA sequences of about 22 nucleotides that serve as a guide for a RNAi-related pathway of gene silencing at post-transcriptional level (Lee et al 1993, Reinhart et al 2000). Similar to the ancestral processing steps of the canonical RNAi pathway, miRNAs are first transcribed in the nucleus as a stem-loop primary miRNA sequence (pri-miRNA) (Figure 23A) that is recognised by the *Microprocessor* machinery, which involves the participation of the *Drosha* protein. This is an endonuclease with two *RNase III* domains that cleaves the primordial sequence at both single-stranded arm extremes and generates the subsequent stage of the sequence, a ~60 nucleotides stem-loop called "precursor miRNA" (pre-miRNA). The current knowledge of the internal canonical structure of the primordial sequence describes essential cleavage points for these processing proteins that are more or less robust to variations of the nucleotide sequence, like wobbles, mismatches and bulges (Figure 23A,B). This intermediate form is then transported to the cytoplasm by the *Exportin 5* and *RAN-GTP*, where it is further recognised and processed by another endonuclease called *Dicer*. This protein cuts the double-stranded part of the stem-loop at both extremes near the loop, liberating the miRNA duplex formed by the main mature miRNA sequence and its partner called "passenger" strand (Figure 23B). This double helix is then loaded into an *Argonaute* protein (AGO) forming the RNA-induced silencing complex (RISC) which separates both strands and expels the passenger miRNA when returning to its relaxed conformation. The loaded miRNA is located into a pocket of the protein serving as a template guide for the complementary binding of target mRNAs (Figure 23C). A particular region of the miRNA is involved in the recognition of the target site of mRNAs, this is the *seed* region, a six-nucleotide portion at the 5' extreme of the sequence (nucleotides 2-7) that establish a Watson-Crick pairing with the 3'UTR region of target mRNAs (Figure 23D) (Lewis et al 2003). An extended version of this seed involves the nucleotide at position 8 (nucleotides 2-8), which

increases the stability and establishes a stronger binding with the mRNA (Lewis et al 2005). This is the most effective binding and mediates most of the mRNA repression, however there are alternative regions in the miRNA that complements the miRNA:mRNA binding. Positions 13-16 correspond to a region that can mediate a supplementary binding and compensate for potential mismatches in the seed region that might awaken the miRNA:mRNA binding (Grimson et al 2007). These are non-canonical sites that, although not very effective, help to increase the stability of the miRNA:mRNA complex and reinforce the repression.

The annotation of human miRNAs

miRBase (Kozomara et al 2019) is the database of reference for annotated miRNAs. In this repository, the constantly increasing number of annotations already collect the miRNA repertories of 271 organisms, with a total of 38,589 miRNA hairpin entries in its last release (v.22.1). The main source of sequence data used to annotate miRNAs is author submission which, with the development of high-throughput sequencing, has increased in number and depth, permitting the annotation of miRNAs at low expression levels. Since 2010 the database has been collecting datasets from deep sequencing projects, which increased the miRNA profiling and the identification of *de bona fide* annotations with the application of quality criteria, like the presence of reads in both arms of the hairpin. However, the curation of the database implies large amounts of effort and time, and it is difficult to assess the accuracy of some of the annotations. In the last releases, *miRBase* has incorporated the annotation of “high confidence” miRNAs, which increased the accuracy of an important fraction of the database. In the case of human miRNAs, in its last release miRBase accounts for 1917 hairpin precursors, from which only 17 were categorized as “low confidence”.

miRNAs are organised in families according to the similarity of their seed sequences, which is directly correlated with their targeting preferences. The members of the same family normally originated from duplication events, therefore these families are evolutionary and functionally related, since they are involved in similar biological processes (Wang et al 2016). However, not all the members of a miRNA family originated from a single ancestral sequence, since different miRNAs might present identical seed regions product of convergent evolutionary processes. On the other hand, commonly originated miRNAs might be placed in different families when they present nucleotide differences among their seed regions which, therefore, generates a change in their targeting preferences.

Sometimes miRNAs might appear in tandem when their origin is from local duplication. These consecutive miRNAs are transcribed as polycistrons in the same pri-miRNA, which are further processed. These agglomerations of evolutionary related miRNAs are known as miRNA clusters and they are specifically considered in evolutionary and functional studies. In the human genome, different studies have identified numerous miRNA clusters according to different criteria (Guo et al 2014, Wang et al 2016). They are normally associated with related biological processes and are involved in specific pathological traits in human populations. In the human chromosome there are three main clustering hotspots that reunite more than 30% of all the cluster members. Chromosomes 14 and 19 present the largest clusters of the human genome with more than 40 members each, while the chromosome X holds smaller clusters but more widespread in location.

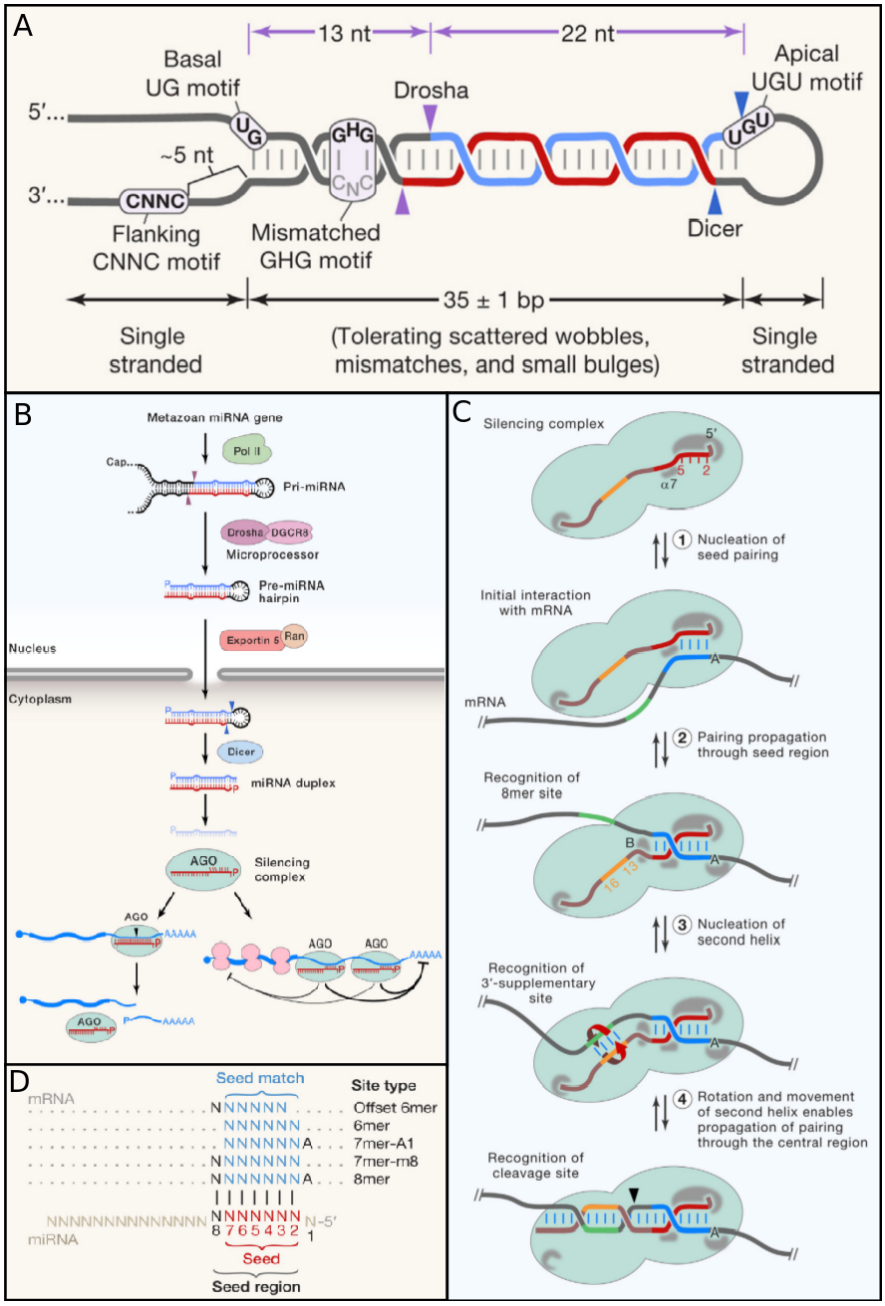


Figure 23. (A) Structural features and processing sites of a typical miRNA hairpin. (B) Biogenesis and function of a typical miRNA. (C) Target recognition and pairing propagation during the mRNA silencing process (D) Complementary binding of the seed region with the different versions of the canonical target sites in the mRNA 3'UTR. (Bartel DP 2018).

miRNA targeting

miRNAs reach a tremendous targeting breadth in the human genome. 3'UTR regions harbour an average of 400 conserved target sites per miRNA family, and each miRNA might present more than one target site in the same 3'UTR. It has been identified conserved target sites in 3'UTR regions of more than 60% of human genes, which together with the non-conserved target sites and their non-canonical binding regions, like the 5'UTR and the CDS of the mRNA, makes that virtually all the human transcriptome is under the regulatory influence of miRNAs (Bartel DP 2018).

The widespread presence of conserved target sites in the human transcriptome makes it rather difficult to study the potential spectrum of target genes that might present a single miRNA. The experimental validation of miRNA targets is a very limited way to decipher the regulatory scope of a miRNA due to its high economic cost and relative slow determination. The alternative to the experimental validation is the computational prediction of miRNA target sites in mRNA sequences. During the last years a wide variety of prediction softwares (Riffo-Campos et al 2016) have been developed to cover the necessity to analyse the potential target genes of miRNAs and determine the biological pathways they might be involved. The first approach that these methods apply to search for target genes is the identification of canonical sites of 7-8 nucleotides in the 3'UTR of mRNA that are complementary to the seed of a certain miRNA family. This is a basic nucleotide pattern search that reveals effective sites that a miRNA might use to generate a significant repression. However, this basic search also displays sites that are not effective and are reported as false positives. Therefore, other sequence properties must be taken into account to improve the accuracy of these predictions, like sequence conservation. The conservation of both the seed region and the target site makes it more likely that the complementary binding takes place in the cellular context and produces an effective

repression. However, even imposing this evolutionary constraint criteria, each miRNA family might still hold thousands of target sites probably populated by false positives. Other methods rely on the spatial restrictions in the formation of the miRNA:mRNA complex. For example, the *PITA* software (Kertesz et al 2007), instead of relying on the sequence conservation of the target site, it takes advantage of the role of the mRNA secondary structure and the analysis of the energetic cost that permits the formation of the miRNA:mRNA duplex. Therefore, the algorithm differentiates those targets that thermodynamically favour the miRNA binding to the target site and also takes into consideration those non-conserved sites that might be involved in lineage-specific regulatory programs. More sophisticated strategies are applied by programs like *TargetScanHuman* (TSH) (Agarwal et al 2015). In its last release (v7.2), TSH applied a total of 14 target and miRNA-specific features to build a quantitative model of targeting efficacy. On the miRNA side the model considers the different extended modes of seed pairing that might participate in the site binding (Figure 23D), taking into account the identity of the sites at positions 1 and 8. Moreover, the target site abundance (TA) for a certain seed sequence is determinant, since the lower is the number of TA the less "diluted" is the repressive effect on the mRNA expression levels. On the site side, the local AU content of the adjacent regions, the presence of the 3' supplementary pairing or the predicted structural accessibility, among other features, might increase the efficacy of the miRNA:mRNA binding. Broader context features are also taken into account, like the 3'UTR length or the presence of target sites in other parts of the open reading frame.

The phylogenetic distribution of human miRNAs

The analysis of sequence homology across animal and plant miRNAs led to the idea, time ago, that the miRNA system might have evolved independently in the two kingdoms. This origin might

be behind the differences in the processing and action between these two groups of miRNAs. This lack of homology is also seen when comparing other lineages, suggesting that this class of functional elements might have present convergent pathways during their evolutionary history. Several reports have estimated that the miRNA system has independently evolved at least nine times. However, other studies provide an alternative explanation to this lack of deep conservation, which is the high rate of turnover of plant miRNAs. The birth and death rates of these miRNAs are so high that they do not have the chance to be found in other lineages. In any case, the common or independent origin of animal and plant miRNAs is still under debate. Although remarkable differences are found at homology, structural and mechanistic level in the miRNA system, scattered examples of relative similarities between these two kingdoms provide reasons to reinforce the study of miRNAs in more species and fully understand their origin (Moran et al 2017).

As previously noted, miRNAs have been present as a regulatory system since the early periods of the metazoan (animal) evolution. The number of orthologs harboured in the genomes of representative lineages revealed a very low rate of miRNA secondary loss (Figure 24A). This is reflected in a continuous increase and significant bursts of newly emerged miRNAs at the base of bilaterians and vertebrate groups (Berezikov et al 2011). In vertebrates, the increase is more pronounced, with specific expansions in the mammalian lineage that are reflected in the current repertoire of human miRNAs. According to Iwama et al 2014, the human miRNA repertoire described in miRBase (release 18, November 2012) is the result of particular gene expansions that took place in localized episodes of their evolutionary history. The authors describe the presence of two peaks of accelerated miRNA rate origination that gave rise to more than 80% of this repertoire during mammalian evolution. These expansions are localized at the beginning of the placental mammals lineage, with the origination of ~28% of human miRNA genes, and mainly in the primate lineage,

with the emergence of more than fifty percent of the repertoire. Moreover, they report that 28% of these miRNAs are specific to the hominoid lineage, suggesting that the increment of miRNAs might be significantly linked to the evolutionary trajectory of humans (Figure 24B).

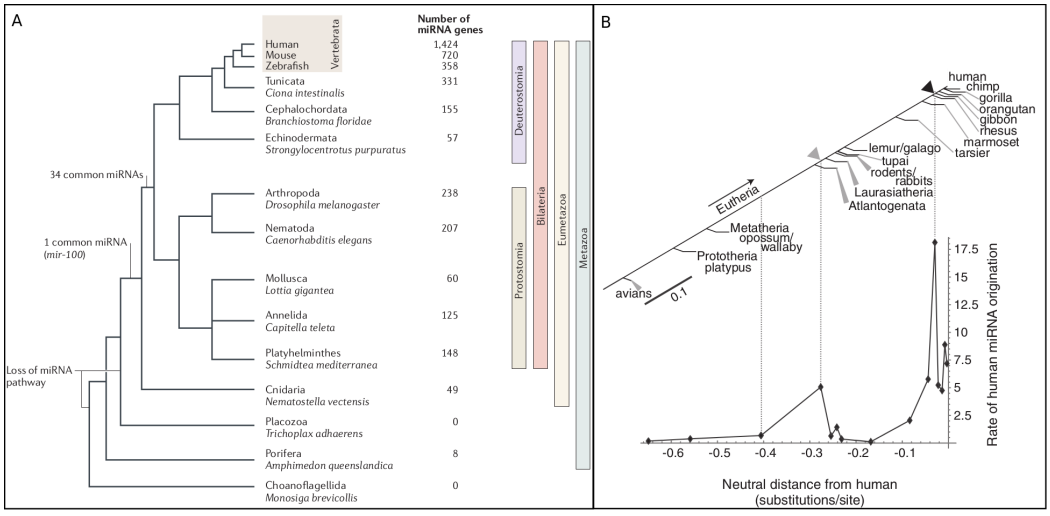


Figure 24. (A) Phylogenetic distribution of the number of miRNA genes showing different losses and expansions (Berezikov et al. 2011). (B) Two accelerated peaks of new miRNA expansions in the placental and primate lineages (Iwama et al. 2014).

The emergence of new miRNAs in the human genome

Multiple genomic sources of new miRNAs have been described in the human genome and other species (Berezikov et al. 2011). Annotation efforts relying on next generation sequencing technologies have been applied in the discovery of new miRNA sequences. Until now, hundreds of thousands of miRNA hairpin-like sequences have been identified in the human genome, however the confident annotation and evidence of the existence of actual functional genes is reduced to a very small fraction. The emergence of new miRNAs appears to be more likely than the origination of new protein-coding genes. The fact that a large

proportion of the human genome is transcribed, this is the *pervasive transcription* (Jacquier A 2009), leads to the hypothetical existence of multiple miRNA loci that are not conserved and might present a functional role in human regulatory programs.

The emergence of new miRNAs might be just due to the existence of a transcriptional locus whose RNA product is able to be folded into a hairpin-like structure. The processing and function of a miRNA is intimately associated with its secondary structure, therefore all RNA sequences potentially of being recognized by the miRNA processing machinery are susceptible to be finally integrated as a functional miRNA in the regulatory networks. In this sense, it is not surprising that the genome might be populated by non-conserved transitional forms of hairpin-like sequences that are either purged or gradually shaped by evolution until they become structures recognisable by the *Microprocessor* machinery. *Microprocessor* is the main point that restrains the processing of hairpin-like sequences, however the downstream steps in this pathway (*Dicer* processing, *RISC* loading) also constrains the selection of the correct sequence to be ultimately processed as a functional mature miRNA.

The main source of newly emerged miRNAs are gene duplication events. In the human genome multiple cases of highly homologous miRNA sequences are found to be evolutionarily related and forming miRNA clusters, however duplicated sequences can also be found in remote locations of the genome, like in different chromosomes. As previously mentioned in this introduction, these miRNA clusters are the result of an accumulation of locally duplicated miRNAs that are found to be functionally related due to their seed identity and targeting preferences. However, these paralogous sequences are also susceptible to undergo processes of neofunctionalization, when a duplicated sequence acquires novel functions product of the emergence of mutations in their sequence,

or subfunctionalization, when the duplicated sequence presents different aspects of the ancestral function (Figure 25).

Other sources of new miRNA sequences are those pre-existing transcriptional units that provide the raw material for the formation of these hairpin structures (Figure 25). In the human genome, around sixty percent of miRNAs are found within intronic regions, where the presence of the mRNA promoter unit allows the transcription of the miRNA itself. However, there are also a remarkable amount of intronic miRNAs that present their own promoter and their transcription is independent of the host gene. In this sense, it is worth noting the different evolutionary properties of the miRNAs whose transcription is ligated to their host genes. For example, in França et al 2016 the authors describe the effect of host gene ages on the expression patterns of intragenic miRNAs and their evolutionary fate in the long term. In summary, miRNAs hosted by old genes tend to present a broader expression breadth than intergenic miRNAs. This would present evolutionary advantages for young intragenic miRNAs, which would reach a higher number of tissues, being in this way more efficiently selected to be purged or incorporated in regulatory networks.

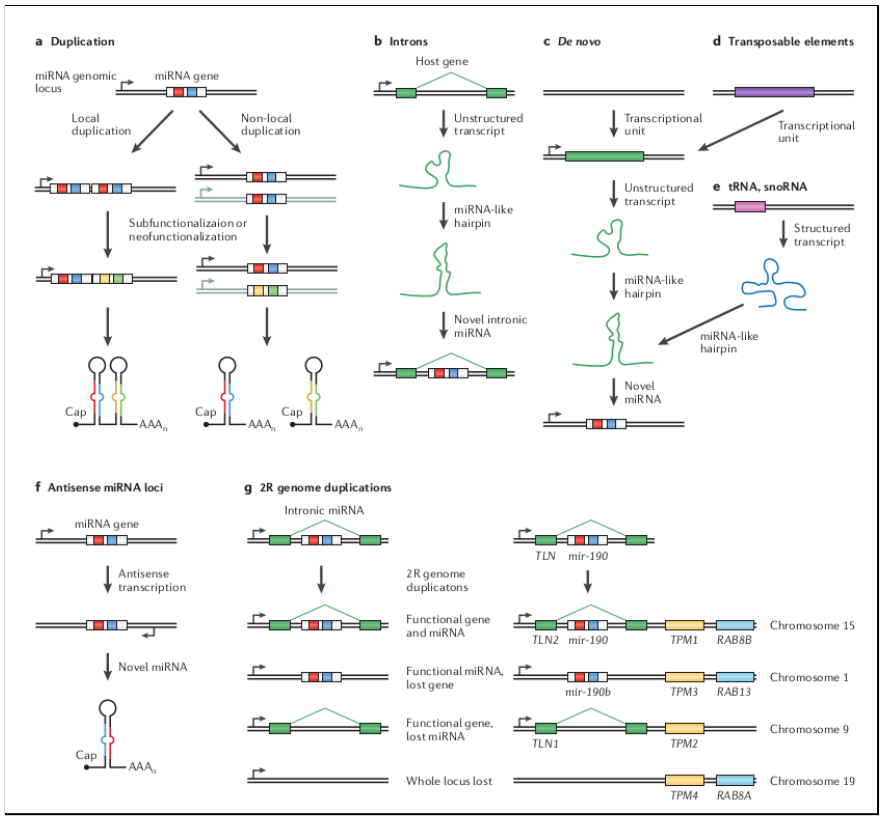


Figure 25. Schematic representation of the different genomic sources of new miRNAs (Berezikov et al 2011).

Genomic regions like those harbouring other RNA genes (e.g. snoRNAs, lncRNAs) and transposable elements (TEs) are also susceptible to be the source of new miRNAs. In the last case, TEs are considered a source of genomic innovation due to their ability to replicate and jump to other parts of the genome, carrying functional regions, like regulatory elements, that could potentially be incorporated into regulatory networks. Different evolutionary episodes of miRNA expansions have been coupled to the parallel emergence of transposable elements (Figure 25). This is the case of the miRNA expansion in the primate lineage. Normally, the criteria of computational predictions uses conservation as a feature to confidently annotate potential new miRNAs and also exclude highly repetitive regions like TEs to reduce the number of false positives.

Considering this, Piriyapongsa et al 2007 argued that numerous cases of TE-derived miRNAs might have been missed and used their own computational approach to report numerous cases of these kinds of miRNAs in the human genome. Petri et al 2019 use an experimental approach to identify miRNA-target interactions and examine their involvement in the regulatory networks of the human brain. They report that many of these miRNAs are derived from L2 transposable elements and present complementary sequences, also derived from this family of TEs, in the 3'UTR regions of many protein coding genes. Several studies have reported evidence on the role of the primate-specific Alu repeats in the origin of new miRNAs and their target sites. For example, in Spengler et al 2014 the authors provide evidence of conserved and functional target sites in Alu sequences located in mRNA 3'UTR regions that are highly susceptible to be recognised by some human miRNAs. Also, in Gu et al 2009 they identified several cases of miRNAs whose transcription is regulated by the presence of Alu repeats. Although a lot of experimental efforts must be applied to confidently annotate human miRNAs, it is clear the relationship between repetitive elements and the function of miRNAs in regulatory networks.

As previously noted, a large part of the mammalian genomes is transcribed (pervasive transcription), and therefore any transcriptional unit might be a potential source of *de novo* emergence of novel miRNAs from unstructured transcripts. Under these assumptions, it is clear that some of these newly emerged hairpin-like structures can generate distorsions in regulatory networks that might derive to deleterious effects. Chen and Rajewsky 2007 describe a model of transcriptional control on those newly emerged miRNAs and how they can evolve in this context. This model postulates that young lineage-specific miRNAs are expected to be expressed at low levels in a low number of tissues in order to reduce the chances of accidental targeting with abnormal effects. With time, deleterious interactions are eliminated and positively selected miRNAs are incorporated in networks of higher

order, incrementing their expression and acquiring a broader range of targets in a larger number of tissues.

The effect of genetic variation in human miRNAs

In the second half of the 2000's the first studies on genetic variation in human miRNAs revealed an overall signature of conservation in this class of genes and the potential consequences of naturally occurring variants in their hairpin sequence (Iwai and Naraba 2005). In Saunders et al 2007 the authors describe the occurrence of sixty-five common variants in 49 precursor miRNAs. They compared the hairpin-like sequence variation with their flanking neutral regions and found a significant decrease of the SNP density in the precursor sequence. They also found the seed as the region with the lowest SNP density, indicating a strong evolutionary constraint in this part of the miRNA (Figure 26A-C). Along with previous studies (Chen and Rajewsky 2006) they also show how miRNA target sites present a dearth of genetic variants in comparison with adjacent non-targeted regions. This was presented as an indication of the evolutionary constraints and, therefore, the functional role of not only the seed, but also their complementary sites at the 3'UTR regions of target genes. The functional consequences of these variants were hypothesized along with these discoveries. Georges et al 2007 exposed different arguments on the effect of the inherited variation affecting the different levels of the miRNA system. The presence of variants either in the seed region and the target site of the miRNA might alter the miRNA:mRNA interaction, originating changes in the targeting profiles and, therefore, expression variation of their targets. Also, genetic variants affecting the hairpin sequence might alter the processing of the pri-miRNA or pre-miRNA, generating changes in their expression profiles. Similar effects are expected when variants occur in the coding sequence of the processing and silencing protein machinery of the miRNA system.

It was clear since the beginning of these studies that genetic variation affecting the different dimensions of the miRNA system would be a massive source of phenotypic variation and diseases not only in humans, but also in other species (Georges et al 2006). In addition, the analysis of the effect of genetic variation was accompanied by analysis of these variants in terms of adaptive selection. The alteration of the miRNA:mRNA interaction at any level might generate regulatory disorders but also changes that might be adaptive under certain conditions. In Saunders et al 2007 they already provide suggestive evidence of two miRNA-related SNPs that might undergo processes of recent positive selection by reporting long unbroken haplotypes in their loci. In Quach et al 2009 the authors provide supportive evidence of miRNAs subjected to natural selection forces. They reinforce the idea of miRNA hairpins as highly conserved sequences in the human genome and how their structural modules present different levels of evolutionary constraints that point to the functional relevance of these specific regions (Figure 26D,E). They also make use of different sequence-based neutrality tests to evaluate the participation of human miRNAs in processes of positive selection. Among the 47 potential candidates that deviate from neutrality expectations, they report several cases of miRNAs that present negative values of Tajima's D, indicative of an excess of rare alleles, and an enrichment of high frequency derived variants captured by the Fay and Wu's H test (Figure 26F). These evidence indicate that these candidates might undergo processes of positive selection, however they note that these patterns might be also affected by demographic events.

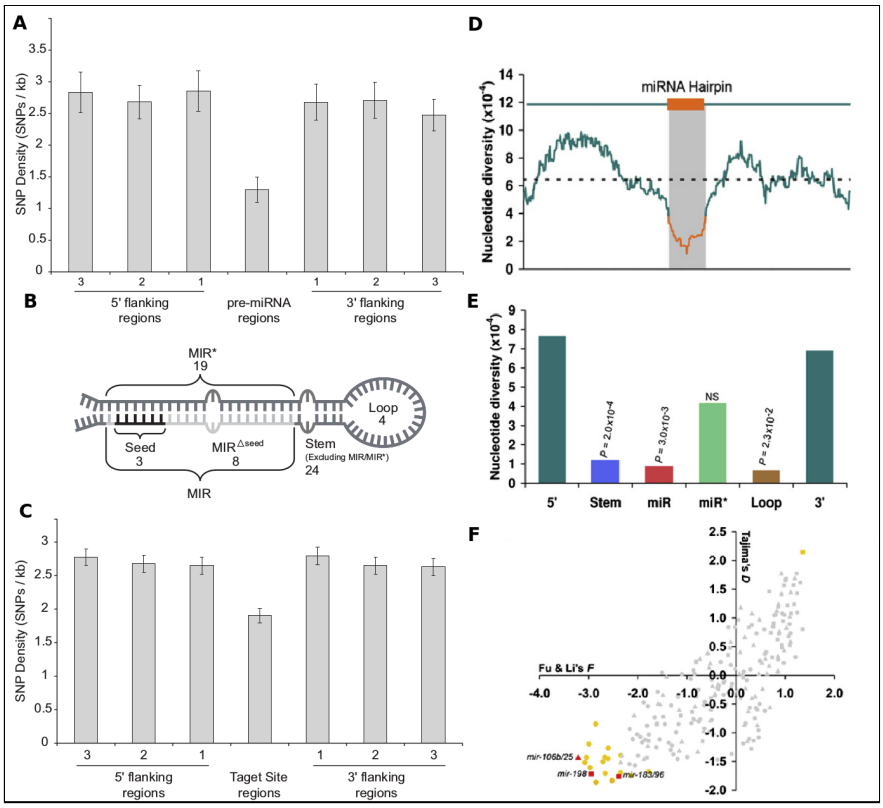


Figure 26. Genetic variation signatures in miRNA hairpins, functional regions and target sites. (A) SNP density (SNPs/kb) in the precursor miRNA sequence and flanking regions. (B) Schematic representation of a miRNA hairpin. (C) SNP density (SNPs/kb) in the miRNA target sites and flanking regions (Saunders et al. 2007). (D) Overall signature of sequence conservation of miRNA hairpins. (E) Levels of nucleotide diversity in the different miRNA regions. (F) Neutrality tests (Tajima's D and Fu & Li's F) calculated in human miRNAs (Quach et al. 2009).

In the following years the study of genetic variation in different human populations led to the description of multiple miRNA candidates behind population-specific phenotypes associated with miRNA-harbouring variants. These population disparities might be the result of demographic events, but also of processes of adaptive evolution. Some studies evaluate population differentiated miRNA candidates that are behind prevalence in diseases like cancer, one of the most common disorders associated with miRNAs. In Rawlings-Goss et al 2014 the authors report some cases of miRNAs with high population differentiation values (F_{ST}) that present

different cancer susceptibilities. For example, they show that the T-allele of rs12355840 (hsa-mir-202) is associated with an increase of the miRNA expression and suggested to be protective against breast cancer mortality. They found that this SNP is highly differentiated between African and non-African populations, having the individuals of African ancestry a lower T-allele frequency compared to Asians and Europeans. This lower frequency is associated with a lower miRNA expression and a weaker repression of cancer-related genes, which increases the progression of breast and ovarian cancers in African women (Figure 27). In another work, Torruella-Loran et al 2016 found that, despite the low SNP density within the mature and seed regions of human miRNAs, the high degree of differentiation among populations in the seed indicates processes of local adaptation that might create differences in their targeting profiles. In particular, they describe population-specific functional differences between three common miRNA SNPs associated with cancer (hsa-miR-146a-3p, hsa-miR-196a-2, hsa-miR-499). The presence of SNPs in the mature sequences of these miRNAs affect the regulation of their target genes in a dosage- and allele-dependent manner, which is suggested to result in genetic susceptibilities to different cancers.

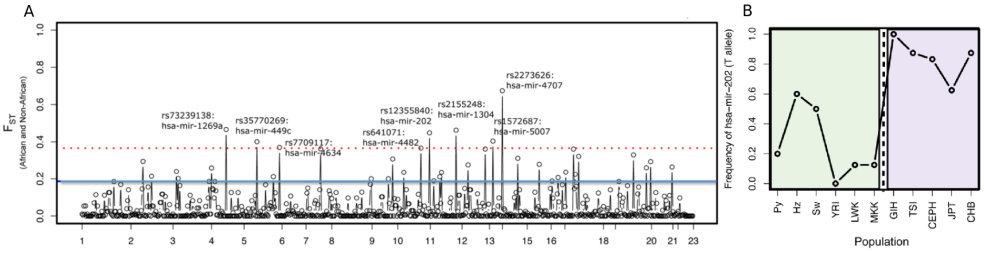


Figure 27. Population differentiation analysis of human miRNAs. (A) Pairwise Fst calculations of human miRNAs between African and non-African populations. Potential candidates under positive selection outstand as genomic outliers. (B) Differences in frequency of the T allele of a highly differentiated miRNA (miR-202) across populations (Rawlings-Goss et al 2014).

Together with the analysis of miRNA candidates, other authors have put a focus on the potential adaptive processes of their target sites.

For example, in Li et al 2012 the authors describe the differential regulation of *TYRP1,* involved in skin pigmentation, driven by miR-155 in human populations. In this analysis, they experimentally validated the action of the target site SNPs rs683 and rs910 in mediating the *TYRP1* preferential regulation in African and Asian populations (YRI, CHB and JPT), which hold an almost fixed derived allele, in contrast with European populations, where it is segregated only at ~30%. Positive selection signatures at the *TYRP1* target sites of Africans and Asians support the idea that adaptive selection processes mediate a stronger repression of *TYRP1* (Figure 28). In another work, Pandey et al 2016 report cases of miRNAs that perform an Alu-mediated regulation of genes involved in stress response. These genes present Alu repeats that operate as target sites and hold signatures of positive selection in specific human populations. The authors conclude that Alu repeats might confer additional mechanisms of transcriptional modulation that increase the regulatory plasticity of miRNA networks to be adaptive under environmental changes.

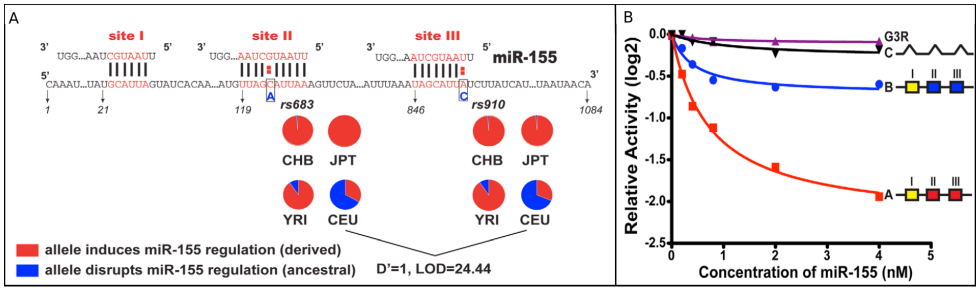


Figure 28. Example of the effect of SNPs in the complementary binding between miRNAs (e. miR-155) and the target mRNA. (A) Predicted target sites mediated by the presence of the SNPs rs683 and rs910, which present different allele frequencies among Africans (YRI) and Europeans (CEU). (B) Comparison between the miR-155-mediated suppression of the ancestral (blue), derived (red) and deleted variants of the target gene. The presence of population-specific variants generates a differential gene suppression (Li et al. 2012).

II. OBJECTIVES

Objectives

The main objective of this thesis is the analysis and interpretation of signatures of genetic variation driven by processes of positive selection in different human populations. To study these selective forces we have focused on protein-coding genes, regulatory RNA sequences (miRNAs) and regulatory elements. In addition, we have studied the role of human enhancers in tissue-specific regulatory programs and how their genomic location is determinant in the regulation of genes involved in tissue-specific functions.

1. In our first work, we aimed to analyse in depth the potential signatures of positive selection of the human X chromosome in populations from the three main geographical groups: Sub-Saharan Africa, Europe and Asia. In this analysis we accounted for the special inheritance properties of the X chromosome and focused on selection signatures of genic and non-genic regions, hypothesizing the role of regulatory elements in selection processes.

2. In our second work, we studied the activation and targeting patterns of enhancer-like signatures (ELSs) from the ENCODE database in different human tissues. Our main goal was to understand the role of the genomic location of these ELSs in the differential regulation of tissue-specific and housekeeping genes and how these patterns change throughout development.

3. The last work presented here is focused on the analysis of signatures of genetic variation in the human miRNA repertoire. The main objective of this analysis was to understand the contribution of highly variable miRNAs in positive selection processes and their relationship with human diseases.

III. RESULTS

Chromosome X-wide analysis of positive selection in human populations: from common and private signals to selection impact on inactivated genes and enhancers-like signatures

Pablo Villegas-Mirón, Sandra Acosta, Jessica Nye, Jaume Bertranpetit and Hafid Laayouni

Submitted for publication

Preprint citation reference:

Villegas-Mirón P, Acosta S, Nye J, Bertranpetit J, Laayouni H. 2021. Chromosome X-wide analysis of positive selection in human populations: from common and private signals to selection impact on inactivated genes and enhancers-like signatures. bioRxiv doi: BIORXIV/2021/445399

Chromosome X-wide analysis of positive selection in human populations: from common and private signals to selection impact on inactivated genes and enhancers-like signatures

Pablo Villegas-Mirón[1], Sandra Acosta[3], Jessica Nye[1], Jaume Bertranpetit[1] and Hafid Laayouni[1, 2*]

[1]Institut de Biologia Evolutiva (UPF-CSIC), Universitat Pompeu Fabra, Barcelona, Catalonia, Spain.
[2]Bioinformatics Studies, ESCI-UPF, Pg. Pujades 1, 08003 Barcelona, Spain.

[3]Dpt. Pathology and Experimental Therapeutics, Medical School, University of Barcelona, Feixa Llarga, 08907, L'Hospitalet de Llobregat, Barcelona, Spain

*** Correspondence:** Hafid Laayouni (hafid.laayouni@upf.edu)

Keywords: Homo sapiens; hard and soft sweeps; neural development; reproduction; enhancers; escape genes.

Abstract

The ability of detecting adaptive (positive) selection in the genome has opened the possibility of understanding the genetic bases of population-specific adaptations genome-wide. Here we present the analysis of recent selective sweeps specifically in the X chromosome in different human populations from the third phase of the 1000 Genomes Project using three different haplotype-based statistics. We describe numerous instances of genes under recent positive selection that fit the regimes of hard and soft sweeps, showing a higher amount of detectable sweeps in sub-Saharan Africans than in non-Africans (Europe and East Asia). A global enrichment is seen in neural-related processes while numerous genes related to fertility appear among the top candidates, reflecting the importance of reproduction in human evolution. Commonalities with previously reported genes under positive selection are found, while particularly strong new signals are reported in specific populations or shared across different continental groups. We report an enrichment of signals in genes that escape X chromosome inactivation, which may contribute to the differentiation between sexes. We also provide evidence of a widespread presence of soft-sweep-like signatures across the chromosome and a global enrichment of highly scoring regions that overlap potential regulatory elements. Among these, enhancers-like signatures seem to present putative signals of positive selection that might be in concordance with selection in their target genes. Also, particularly strong signals appear in regulatory regions that show differential activities, which might point to population-specific regulatory adaptations.

INTRODUCTION

The evolution of *Homo sapiens* has been strongly shaped by positive selection in the last 100,000 years, by adaptations to specific environments, diets, and cognitive challenges that modern human populations encountered as they expanded across the globe. Surviving such challenges has left remarkable footprints of selection in the human genome, like in the lactase (*LCT*) locus in European populations (Bersaglieri et al., 2004; Wang et al., 2020), genes involved in skin pigmentation like *MC1R* (John et al., 2003) or genes implicated in resistance to severe malaria infection like *CD40L* and *G6PD* (Sabeti et al., 2002). Studying the evolutionary processes that resulted from these adaptations can uncover which path our ancestors travelled along to give rise to extant adaptations of present human populations.

The development of new methods to study recent selection in natural populations (Fan et al., 2016; Field et al., 2016; Pavlidis et al., 2017) has settled genomic selection scans as one of the main approaches to study the genetic origin behind such adaptations (Mathieson et al., 2015; Casillas et al., 2018; Lopez et al., 2019; Walsh et al., 2020). However, most of these scans have focused on coding regions as the main target of selection and have attached greater importance to the study of processes driven by *de novo* mutations, that leave strong and more evident selection signatures (classical hard sweeps). Although gene regulation is considered to be the primary driver of phenotypic changes in the evolution of *Homo sapiens* (King and Wilson, 1975), these strategies might have overlooked standing variation in regulatory regions as the main targets of rapid adaptations, which seem to be more likely selection targets and are marked by more subtle signatures, like soft sweeps (Fu and Akey, 2013; Scheinfeldt and Tishkoff, 2013; Messer and Petrov, 2013).

Selection on standing variation seems to be a more common mode of selection and soft sweeps a more widespread signature in human genomes (Hernandez et al., 2011; Schrider and Kern, 2017). Multiple modes of selection can originate a soft sweep signature: on standing variation, de novo mutation on multiple haplotypes and recurrent origination of adaptive alleles (Schrider et al., 2015; Hermisson and Pennings, 2017). However, sometimes patterns of variation might exhibit different degrees of "softness" and, together with confounding factors like demography or recombination, display sweep-like signatures where the picture is not clear enough so as to define a region under a specific selection regime (Messer and Petrov, 2013). Therefore, sometimes it is difficult to differentiate signatures due to hard or soft sweeps, and often linked regions under selection may present properties of both types of signals (Schrider et al., 2015).

The X chromosome, although it's been studied in terms of recent positive selection in human populations (Casto et al., 2010; Veeramah et al., 2014; Johnson and Voight, 2018), remains to be addressed more completely, including selection on regulatory regions and a focused analysis of signatures of selection on standing variation. The X and Y chromosomes have different inheritance models than the autosomes as well as different effective population sizes, making the outcome of selection pressures inconsistent to the rest of the genome. In order to study the X chromosome, these different properties have to be taken into account and, to analyse the selection events that took place, chromosome-specific demographic models and region-specific recombination maps must be incorporated to approximate a more realistic scenario.

The unique properties of the X chromosome, as compared with autosomes, have been extensively studied (Vicoso and Charlesworth, 2006; Mank et al., 2016; Meisel and Connallon, 2013). Dosage compensation of the X chromosome, the process that allows XY males and XX females to cope with different gene copy

numbers on the X, might lead to sex-specific patterns of selection. This process involves the random transcriptional silencing of one of the X chromosomes in females. However, this inactivation is not complete for all the genes. Evidence suggests that around 23% of the X-linked genes "escape" inactivation and express both chromosomal copies (Balaton et al., 2015; Tukiainen et al., 2017), leading to a sex-biased expression of these genes, which might be responsible for dimorphic traits, and potentially, adaptations associated with phenotypic diversity. Despite the few studies of selection on these genes, some evidence indicates that these regions have been under purifying selection (Park et al., 2010). Thus, it is of interest to see whether positive selection has operated in these regions and the relative importance that inactivation may have on the process of natural selection.

The faster-X hypothesis (Meisel and Connallon, 2013) postulates that selection occurs faster in genes on the X than in autosomes due to the hemizygosity of males, this is supported by recent evidence that found increased selection levels in the sexual chromosome (Veeramah et al., 2014). Moreover, different effects of mutations in males versus females have been well-established (Vicoso and Charlesworth, 2006). The difference in the replication rate between female and male germ lines favours this hypothesis. The higher probability of suffering consequences due to deleterious and adaptive mutations most likely has led to a different selection process. Altogether, these factors may lead to specific patterns which reflect the sex-biased evolution in humans.

In this study, we conduct a selection scan on the X chromosome of 15 human populations from three different continental groups (Sub-Saharan Africa, Europe and Asia). We sought to identify signatures of recent positive selection by considering hard and soft sweeps. potentially affecting both coding and non-coding regions. With this we aim to disentangle how positive selection has shaped the diversity patterns in the X chromosome across the globe.

MATERIALS AND METHODS

Genetic Data

Phased VCF files from the third phase of the 1000 Genomes Project were downloaded from the project database (Auton et al., 2015). These data are whole-genome (mean depth of 7.4X) and targeted exome sequences (mean depth of 65.7X) with a total of 2,504 individuals across 26 different populations, covering three continental groups. Due to the methodological complexity, only the non-admixed populations of each geographical group were analysed. In Africa: Esan (Nigeria, ESN), Gambian (Wester Divisions in the Gambia, GWD), Luhya (Webuye, Kenya, LWK), Mende (Sierra Leone, MSL), Yoruba (Ibadan, Nigeria, YRI); Europe: Utah residents with northern and western European ancestry (CEU), Finnish (Finland, FIN), British (England and Scotland, GBR), Iberians (Spain, IBS), Toscani (Italy, TSI); and Asia: Chinese Dai, (Xishuangbanna, China, CDX), Han Chinese (Beijing, China, CHB), Southern Han Chinese (China, CHS), Japanese (Tokyo, Japan, JPT), Kinh (Ho Chi Minh City, Vietnam, KHV). We applied filters to remove duplicated variants found in the X chromosome. These errors were reported to the 1000 Genomes Project (www.1000genomes.org).

The X chromosome consists of both pseudoautosomal regions (PAR) and non-pseudoautosomal regions (nPAR). Since the PAR behaves differently and does not follow the same inheritance rules than the rest of the X chromosome, we removed these regions keeping only bi-allelic variants within the position range of the nPAR region (~2.7-155.0 Mb) (Flaquer et al., 2008).

We reformatted the VCF file so that the ancestral allele was the reference and the derived allele was the alternative. The human ancestral alleles determined by their state in chimpanzee were

downloaded from the 1000 Genomes Project mapped to human reference GRCh37. We removed any SNP whose ancestral status was unknown, resulting in a total of 2.852.479 SNPs from 1511 individuals (504 Africans, 503 Europeans, and 504 Asians).

We downloaded a population-combined genetic map of the nPAR region (http://mathgen.stats.ox.ac.uk). This map was based on the first phase of The 1000 Genomes Project (GRCh37). In order to use the map for phase three data, we performed a linear interpolation of the missing values using the command *approx* from the statistical programming language R (R Core Team, 2020).

Neutral simulations

We used the *msms* software (Ewing and Hermisson, 2010) to simulate neutral scenarios. For the X chromosome we implemented a three-population demographic neutral model adapted from Henn et al. (2015) for the continental populations Africa (AFR), Europe (EUR), and Asia (ASI) with a mutation rate of $1.25x10^{-8}$ mutations per base per generation (Henn et al., 2015), a generation time of 30 years, a recombination rate of $1.3x10^{-8}$ per nucleotide, and a Watterson estimator θ ($4N_e\mu$) of 328.79. We chose a three-population model due to the high similarity within continents, with a mean sample size of AFR: 152, EUR: 153, and ASI: 149. Since the effective population size of the X is ¾ the size of the autosomes, we accounted for this by modifying the population sizes, resulting in N_e for AFR: 23220, EUR: 2479, and ASI: 907. We simulated multiple regions of 600 kb in order to reproduce the total length of the X chromosome, by using the following parameters:

```
msms -N 10538.25 -ms 454 254 -t 316.1475 -r 328.7934 600000
-I 3 152 153 149 0 -n 1 2.204 -n 2 3.2542 -n 3 7.4055 -g 2
56.61 -g 3 96 -ma x 0.3542 0.1462 0.3542 x 1.3562 0.1462
1.3562 x -ej 0.0464 3 2 -en 0.0464 2 0.2939 -em 0.0464 1 2
4.9314 -em 0.0464 2 1 4.9314 -ej 0.14022 2 1 -en 0.364 1 1
-oTPi 30000 25000 -tt -oAFS
```

In order to contrast the results obtained for the X chromosome, we analysed the complete set of autosomes in the human genome. The same procedure to detect positive selection as for the X was followed. To do so we performed the appropriate autosomal neutral simulations and used the percentile 99^{th} as extreme distribution cut-off to compare the regions under positive selection. Also, the Refseq gene annotation from the UCSC database table browser (Karolchik et al., 2004) (downloaded June 2020) was considered.

Scan for signals of selection

Advances in the statistics used to detect selective sweeps allow for the analysis of linkage disequilibrium decay (Pybus et al., 2015, Biswas and Akey, 2006; Vallender, 2004; Sabeti et al., 2006; Garud et al., 2015). These methods rely on detecting decreased variation surrounded by a region with high linkage disequilibrium (LD). The LD increases and the variation decreases as the frequency of the selected allele rises in the population. Once the selected allele is fixed, selection will relax, allowing for variation to recover through new mutations and recombination. The extended haplotype homozygosity (EHH) computes the probability that, at a given distance from a core region, two randomly chosen chromosomes carry homozygous SNPs for the entire interval. In this analysis we made use of three different haplotype-based statistics that rely on the EHH computation at a tested SNP, taking into account the ancestral and derived allele state.

The integrated haplotype score (iHS) is the integral (Voight et al., 2006) of EHH and is designed to detect incomplete hard sweeps. These are signatures of recent, ongoing selection that are characterized for presenting long blocks of homozygosity found in haplotypes with a high frequency of derived alleles. We have used two methods to detect signatures that resemble soft sweeps. The integrated haplotype homozygosity pooled (iHH12) (Torres et al.,

2018) is an adaptation of the H12 statistic by Pickrell et al. (2018) and is able to detect signatures of both hard and soft sweeps, and the number of segregating sites by length (nSL) (Ferrer-Admetlla et al., 2014), a modification of iHS with a higher robustness to recombination rate variation and with an increased power to detect soft sweeps. These are the footprints left by selection processes that target variants at intermediate frequencies. On the contrary to the hard sweeps, that involve the fixation of a single *de novo* mutation due to a specific environmental change, soft sweeps might be generated by the selection of an allele that was drifting neutrally at the moment of the change. Also, these footprints might appear when different alleles are selected simultaneously at the same locus. Therefore, the footprints left by this kind of process are not as evident as the signatures left by the hard sweeps, since the diversity reduction left by the sweep is lower. These tests for recent positive selection are standardized (mean 0, variance 1) by the distribution of observed scores over a range of SNPs with similar derived allele frequencies. Here, we use the three tests, iHS, nSL and iHH12, to detect selective sweeps in the X chromosome.

The candidate signals for selection may point to putative targets of recent selection which are of particular interest in the study of human evolution and may help to understand complex phenotypes of medical relevance. The calculations of iHS, nSL and iHH12 were computed with the software *selscan* (Szpiech and Hernandez, 2014), an application that implements different haplotype-based statistics in a multithreaded framework. We allow for a maximum gap of 20kb and keep only SNPs with a minor allele frequency (MAF) higher than 5%. These parameters reduce the number of false positives due to the presence of gaps in the data, however special care must be taken when interpreting these results since false positive rate could increase with other confounding factors. The same procedure was applied on the simulated data in order to compare the empirical distributions with a neutral score background. The standardization was performed by the *norm*

function within the *selscan* package for each population and test separately. The calculations of the Tajima's D scores in Figure 2 were performed by using the software package *VCFtools* (0.1.14) with a non-overlapping 10kb sized window-based approach (Danecek et al., 2011).

The program *selscan* considers ancestral and derived alleles separately. iHS and nSL report positive values when the derived allele is selected while negative values indicate the ancestral allele is favoured (Szpiech and Hernandez, 2014). iHH12 makes no distinction between the two allele states. Since a sweep may also be produced by the hitchhiking of ancestral alleles with the selected variant, absolute values were considered. The per-SNP scores were summarized by using a position-based sliding window approach of size 20kb with a 20% overlap (4kb). Windows with 20 SNPs or fewer were removed. The mean scores were calculated in each test in order to interpret the presence or absence of a selective sweep. To search for candidate windows under positive selection, we compared the distributions of the summary observed values to the simulations and considered 99^{th} and 99.9^{th} percentiles in the simulated distribution as critical values to have evidence against neutrality. Empirical summary values over these thresholds were considered as putative signals of positive selection. No p-values were associated with the significance of these windows.

The haplotype structure in regions under putative positive selection was determined with the program *fastPHASE* (Scheet and Stephens, 2006). This software applies a Hidden Markov Model (HMM) on haplotype data to obtain the frequencies of a certain SNP to be in a haplotype cluster according to the similarity between them, such that the region is divided into a mosaic of clusters per population that reflects the patterns of haplotypic variation.

In order to assess commonalities and differences across populations, we identified the regions under selection that are in the extreme tail

of more than one population. Since a region under positive selection can be captured by more than one test depending on the variable degree of "softness" in its locus, the shared sweeping regions were constructed by using the candidate windows reported in the extreme 99th percentile across the three selection tests. Sweeping regions that overlap across more than two populations of the same continental group were considered shared in that group.

Gene Ontology

We downloaded the Refseq gene annotations from the UCSC database table browser (Karolchik et al., 2004) in June 2020 to annotate the X chromosome. This annotation describes all the transcripts including 5' and 3' untranslated regions (UTR), coding, and non-coding genes. We merged these annotations with our empirical data using *Bedtools intersect* (Quinlan and Hall, 2010). We intersected our candidate windows under selection with the annotated genomic regions to obtain a list of genes under putative positive selection. Finally, an Overrepresentation Enrichment Analysis (OEA) was performed on the most extreme top 100 genes for each population with the online tool *WebGestalt GSAT* (Gene Set Analysis Toolkit). The multiple testing was adjusted using the Benjamini-Hochberg correction, accepting ontology terms with a global false discovery rate (FDR) ≤ 0.05 as significant.

In order to focus on putative regions with the highest selection scores, we selected the top windows that fall into the 99.9th percentile. The SNPs contained in these windows were annotated using the *ANNOVAR* program (Wang et al., 2010), which aggregates the UCSC annotations: GWAS Catalog, CADD scores, GERP++ scores, Conserved transcription factor binding sites (TFBS) in the human/mouse/rat alignment, segmental duplications, and clusters of TFBS based on ChIp-seq data. In order to identify the most interesting SNPs inside each region, we considered SNPs with an individual selection value within the 1% extreme tail of the

distribution (|iHS| and |nSL| ~ 2.5 in all populations, and iHH12 ~ (Africa: 4.1, Europe: 3.8 and Asia: 3.6)) and a PHRED-scaled CADD score ≥ 10, which represents the whole genome 1% most deleterious SNPs according to Kircher et al. (2014). Also, as a way to prioritize SNPs located in regulatory regions, we explored the potential effects of SNPs from both 99th and 99.9th top windows within functional regions by using RegulomeDB (Boyle et al., 2012). This database uses ENCODE data sets to annotate variants that are likely to belong to a functional region and thus suggest possible hypotheses to the SNPs within the selection signal. This database presents a classification scheme that scores the variants according to the support they have of functional elements. The functional categories decrease with the relevance of each variant. In this line, the category 1 corresponds with those variants that present an eQTL and support from other ENCODE data, while the category 6 only presents a hit in a single motif.

"Escape" genes selection analysis

The putative selection signals were used to explore potential signatures of recent positive selection in genes with X chromosome inactivation (XCI) status. Several studies have cataloged XCI gene status in order to categorize genes that escape inactivation (Balaton et al., 2015; Carrel and Willard, 2005; Cotton et al., 2013). For this analysis, the inactivation status was considered using the catalog by Tukiainen et al. (2017), which includes a consensus of XCI statuses from previous studies (Carrel and Willard, 2005; Cotton et al., 2013) and extends it by creating a landscape of human XCI across different tissues (GTEx project, v6p release) and individuals. The integrated statuses of these studies fall into three categories: *escape* (if "escape" and "variable"), *variable* (if "escape" and "inactive"), and *inactive* (if "variable" and "inactive"). Contingency tables were constructed based on selection (Selected/Not selected) and XCI (Escape/Inactive) statuses. The independence among these categories was tested with Fisher's exact test method.

Regulatory regions under positive selection

The HACER database (Wang et al., 2019) was used to annotate intergenic windows in order to study potential signals of positive selection in enhancer-like regions. HACER annotates a total of 1,676,284 active enhancers (whole genome) detected by different methods (GRO-seq, PRO-seq and CAGE) in numerous cell lines and supported by different databases (VISTA, ENCODE Enhancer-like Regions, The Ensembl Regulatory Build and chromatin state segmentation by ChromHMM) which, integrated with variation data, provides a useful resource to hypothesise about the origin of non-genic signals of natural selection. In order to reduce the noise and provide a higher confidence to our intergenic signals, we have used the enhancers that at least are supported by the annotation of one database which, in the X chromosome, leave a total of 23790 active enhancers. In HACER, a given region can be annotated as an active enhancer in different cell lines, targeting the same closest gene but presenting slightly different coordinates. In order to deal with the different cell-type-specific annotations we created a "consensus" dataset of enhancers by using genomic windows. We collapsed the multiple cell-type annotations to unique enhancer coordinates when there are different overlapping enhancer regions, active in different cell lines, targeting the same gene and overlapping continuous windows. In this way we ended up with a final dataset of 1322 consensus enhancers that we used to annotate our intergenic signals. When extracting the top hits under positive selection (99.9^{th} percentile) we only took into account those enhancers that are supported by 3 or more databases in the HACER annotation, in this way we only considered high confidence enhancers that might present signals of positive selection.

Luciferase analysis

Enhancer peaks from the top candidates were selected upon the ENCODE signals. Ancestral (A) and derived (D) haplotypes were amplified by PCR from male (*KDM6A*: NA07357 (A), NA12003 (D); *SH2D1A*: NA18501 (D)) and female (*SH2D1A*: NA18502 (A); *HUWE1*: NA18502 (A), NA18861 (D)) individuals, after checking for homozygosity, using the following primers and the KAPA high-fidelity Taq polymerase:

KDM6A (F): 5'-CATCAGAGCTCCTCTAGGCATGGGAGGGAGT-3'
KDM6A (R): 5'-TCATCTCGAGCCAGTAAGAACCTACTAGGGATCA-3'
HUWE1 (F): 5'-CATCATCTCGAGGACCAGCCACTGGGTGTAGT-3'
HUWE1 (R): 5'-TCATAAGCTTTAGGGTCCATGGTCTTCTGG-3'
SH2D1A (F): 5'-CATCATCTCGAGACAAATGTTATTGATTCCCTC-3'
SH2D1A (R): 5'-TCATAAGCTTCGACCTAAAAGAGTATA-3'

Cloning into the PGL4.10 luciferase clone was performed by using XhoI, HindIII or SacI restriction enzymes. Renilla vector was used to normalize the values as a control of transfection. Transfection into 293T cells was performed by using Lipofectamine 3000 (Thermo Fisher, L3000001), using 100 ng of luciferase and 1ng of Renilla control vector and maintained for 48 hours in OptiMEM. Cells were harvested and luciferase activity was measured using the Dual-GLO kit (Promega, E2920). Luciferase/renilla ratio calculated in 4 replicates and 2 independent experiments.

RESULTS

We inferred recent positive selection in human X chromosomes using genomic data from 1,511 individuals of 15 populations. We conducted selection scans by applying the haplotype-based statistics iHS, iHH12 and nSL, which were designed to detect signatures of hard and soft sweeps (see Methods for details) and can be used as complementary selection tools. To assess whether a region has evolved under recent positive selection we performed coalescent simulations with *msms* (Ewing and Hermisson, 2010) to build the expected distributions under neutrality, considering human demography and the particular ascertainment bias of our data. We

observed a good fit of our neutral model by comparing the observed site frequency spectrum (SFS) of the fifteen populations with their neutral simulations (Supplementary Figure 1). Small deviations in singletons are observed in some populations, but with a tight fit of alleles segregating at intermediate and high frequencies.

Regions under putative positive selection

The per-SNP metric scores might reflect the presence of particularly homozygous regions, which could indicate the location of a selective sweep in the genome. In order to detect these signatures, the selection scores were averaged separately across sliding overlapping windows (see Methods; Supplementary Figure 2), which in most populations show distributions with a larger tail as compared with the simulations (Supplementary Figure 2A). We considered two cut-offs based on the simulated data (99^{th} and 99.9^{th}) in order to extract the putatively selected windows in the empirical distributions (Supplementary Table 1).

Putative selective sweeps in regions under positive selection might present different degrees of "softness". As noted by different authors, hard and soft sweeps are sometimes difficult to differentiate (Messer and Petrov, 2013; Schrider et al., 2015), and regions under selection might be captured by methods designed to detect both selection processes at the same time. In order to study the signature similarity in the regions under selection, we assessed the degree of overlap between the signals reported by the three metrics. Under the 99^{th} percentile in the global population, the general trend shows that iHH12 presents a similar proportion of commonly targeted regions as with iHS and nSL (~60%), while iHS targets fewer common regions as with nSL (~36%). This could be expected since iHH12 and nSL are sensitive to both hard and soft sweeps (Ferrer-Admetlla et al., 2014; Torres et al., 2018), and iHS depends on recombination rate, which might differentiate these signals from nSL signals. However, the signal overlap proves that some regions might present

mixed properties of hard and soft sweeps, which could be due to the mode of selection, the degree of softness or a linked selection effect (Schrider et al., 2015).

We observed a larger proportion of signals that fall outside the simulated distribution in the African populations in the three selection tests, in comparison with non-Africans. These results are in line with previous reports which show that the number of detectable selective sweeps by haplotype-based statistics is correlated with the effective population size (Johnson and Voight, 2018; Voight et al., 2006) (Supplementary Table 1). When comparing both hard and soft selection processes we observed that soft-sweep-like signals reported by nSL and iHH12 are more abundant and widespread along the X chromosome, as was previously reported at genomic level (Messer and Petrov, 2013; Schrider et al., 2017).

The analysis reveals that high statistical values are clustered in specific spots of the X chromosome, indicating the presence of putative selective sweeps in these regions (Figure 1) (Voight et al., 2006). The distribution of signals of selective sweeps along the X chromosome is more similar between non-African than with African populations in both selection processes, indicating a common clustering of extreme signals among the different out-of-Africa populations. This was noted by Pickrell et al. (2009) and might reflect the common origin of the out-of-Africa populations and must have been acquired since leaving the African continent.

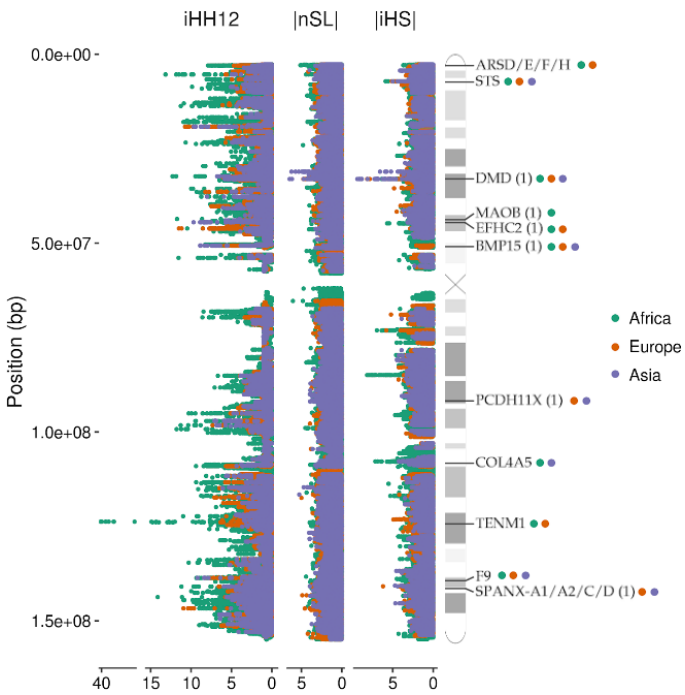


Figure 1. Manhattan plots of the X chromosome showing the distributions of the three selection tests used in the analysis. Some examples of genes found under selection in continental groups (99th; coloured circles) are shown in the ideogram. Candidates found in previous studies are indicated with (1).

Comparison with autosomes

The unique inheritance rules of the X chromosome might generate different selection patterns as compared with the rest of the genome. In order to contrast the X chromosome signatures, we assessed selection on the autosomes of three populations of reference (Yoruba, YRI; Utah residents with northern and western European ancestry, CEU; Han Chinese, CHB) and compared the score distributions in the three haplotype-based statistics (iHS, nSL, iHH12). We see similar patterns of selective sweeps across the different populations as in the X: a higher number of outlier regions fall into the extreme tails of the autosomes in Africans (YRI) than Europeans (CEU) or Asians (CHB) (Supplementary Table 2). As seen in the X, a higher number of windows under selection are captured by the statistics nSL and iHH12 in comparison with iHS

across the autosomes, probably due to the higher presence of soft-sweep-like signatures across the genome. One large difference that stands out, is that the X chromosome exhibits a consistent increase in the nSL extreme tail score distributions in non-African populations (Supplementary Figure 3). We evaluated the nSL scores in the top distribution quartile and decile, and found significant differences between the X chromosome and the pooled scores of autosomes (CEU, Kruskal Wallis: 36.04, p = 1.93e-09; CHB, Kruskal Wallis: 93.62, p = 3.81e-22). These higher selection values might be a reflection of the effect due to the haploid state in males and the smaller effective population size of the X (Veeramah et al., 2014; Johnson and Voight, 2018). However, it is difficult to associate these differences with a higher selection efficiency due to the faster-X effect, since the top 1% shows inconsistent distributions across the genome due to the presence of extreme outliers. This result might indicate that the faster-X effect is not properly captured with these selection statistics and other causes might be behind the differences seen in the distribution extreme tails.

Gene ontology in the candidate regions

Generally, the closest gene to the estimated sweep is considered the best candidate for the target of selection. Putative selected regions were annotated as genic (protein-coding and non-protein coding; Supplementary File 1) where at least 1 bp of the window overlaps with Refseq gene coordinates. We do note that some caution is required when interpreting these results, as the strongest and widest signals are likely to span more than the target of selection.

To determine which processes are likely under selective forces, we performed a functional enrichment analysis with *Webgestalt* (Liao et al., 2019) on the top 100 putatively selected genes across all populations for the two selection regimes. There is a ubiquitous enrichment in neural-related terms in the three continental groups (Supplementary Table 3). In the two selection processes we report

numerous synaptic and dendrite-related terms (e. “postsynaptic membrane” (GO:0045211), “dendrite” (GO:0030425)) with genes like *DMD*, *IL1RAPL1* and *GABRA3*, among others. Neuron-surface specific genes are also highly represented among the enriched terms with kinases like *CASK*, channels (*TRPC5*) and neuroligins (*NLGN4X*, *NLGN3*), which present their own term in numerous populations (“neurexin family protein binding” (GO:0042043)). However, for the African populations (Supplementary Table 3A) “sulfuric ester hydrolase activity” (GO:0004065) and “endoplasmic reticulum lumen” (GO:0004065) are consistently enriched non-neurological terms represented by members of the arylsulfatase family (*ARS*) and steroid sulfatase (*STS*) gene (Holmes, 2017). These genes, which are involved in hormone metabolism and are associated with X-linked diseases like chondrodysplasia punctata (Franco et al., 1995) and ichthyosis (Basler et al., 1992), present a strong signal of selection (99.9^{th}) in African populations. We also observe genes consistently selected in continental groups which do not correspond with any enriched term, including reproduction-related genes, like *SPANX-A1/A2/C/D* and *SPANX-OT1* in non-African populations. These genes belong to the spermatogenesis-related gene family *SPANX-A/D*. This is a highly paralogous hominin-specific group of genes which are expressed post-meiotically in testis and some cancer types (Westbrook et al., 2006) whose members were previously reported as positively selected (Casto et al., 2010; Kouprina et al., 2004) and related to male fertility (Urizar-Arenaza et al., 2020). We observe signals of positive selection on the *BMP15* gene, related to ovarian insufficiency in women and subjected to positive selection in Hominidae clade (Ahmad et al., 2017). Other spermatogenesis-related genes (*SAGE1*, *SEPT6*, *CDK16*) and genes involved in human fertility (*ADGRG2, DIAPH2, FAM122C*) also appear in the highest scoring regions (99.9^{th}) of our scans (Supplementary File 3).

Shared sweeps in human populations

Previous reports have shown that signatures of positive selection are often shared between different human populations (Johnson and Voight, 2018). Common evolutionary trajectories might generate similar selective pressures which leave shared signatures of positive selection. These common patterns might reveal important traits that were crucial in the adaptation of ancestral populations. To that end, we assessed the degree of sharing of the candidate regions under putative positive selection. We considered the 99th percentile candidates in the three selection tests and identified those regions whose genomic coordinates overlap across multiple populations. We found that 41% of the selective sweeps are unique to a specific population, 38% are shared between populations of the same continental group and 20% are shared across different continents. These results are in line with previously reported selection patterns (Johnson and Voight, 2018): common sweep events are more frequent between closely related populations, and cross-continental sweeps are rarer and more likely to result from common selective pressures and older processes of positive selection.

Among the cross-continental selected regions we found that one of the most commonly shared falls within the *DMD* (dystrophin) gene. This is the largest gene in the human genome and is involved in the stabilization of the sarcolemma and synaptic transmission. We found multiple signatures of hard and soft sweeps across the 15 included populations, which together span a region which reaches up to ~2Mb (Supplementary Figure 4A). The variable length of this sweeping region might indicate that multiple selection events took place in the three continental groups, which generated different patterns that suit the two selection processes. Positive selection signals were previously reported in several components of the dystrophin protein complex (*DPC*) (Williamson et al., 2007) in non-African populations and in *DMD* in Africans (Casto et al., 2010). Our *DMD* results are complementary to these previous

studies and validate evidence for adaptations in neurological and muscle-related phenotypes in other populations.

Another globally shared region overlaps the *F9* gene, which encodes the coagulation factor protein FIX and is involved in Hemophilia B. In this case, the *F9* region harbours windows under positive selection in the 99.9th percentile reported by iHH12, which reflects a sweeping region that spans up to ~50kb (Supplementary Figure 4B). A previous study reported coagulation factors underwent positive selection in different clades (Rallapalli et al., 2014), which might be a consequence of selective pressures due to the direct relationship with the immune system and host-pathogen interactions. Although the FIX factor has not been identified as related to any selective pressure to date, it might be under recent positive selection in human populations due to its role in the coagulation system as the first line of defence against pathogens.

TENM1 gene

The most extreme signals in the analysis are reported by iHH12, reaching in some cases values between 10 and 15 in African populations (>99.97%). Patterns of soft and incomplete hard sweeps might be a side effect of linked regions targeted by complete hard sweeps, referred to as the "soft sweep shoulder" (Schrider et al., 2015). A possible example of this is seen in the *TENM1* gene, which is the highest scoring region in the chromosome with an iHH12 signal composed of two high peaks (Figures. 2A,B). This gene is involved in neural development and is specifically determinant for the synapse organization of the olfactory system. In African populations this region exhibits a peak value of iHH12 > 40, while in non-African populations is hardly captured by iHH12 due to an excess of low minor allele frequency variants (MAF < 0.05), which are filtered out by *selscan*. iHS and nSL outlier windows are also found within this region, suggesting the presence of haplotype patterns which fit with both soft and hard sweep

signatures. In order to elucidate the haplotype structure of this region, we inferred clusters of similar haplotypes with *fastPHASE* (Scheet and Stephens, 2006) on representative populations of the three continental groups (CEU, CHB and YRI). Figure 2C shows different long haplotypes at high frequency with the main presence of two highly homozygous clusters overlapping the iHH12 peaks, either in African or non-African populations. This pattern is expected in regions that underwent selection processes and left long, unbroken haplotypes where no recombination events occurred. The two main clusters span ~300kb of the *TENM1* gene and their location suggests that an ancient strong selection event took place in this region before the population split in the out-of-Africa event. For confirmation, we calculated the Tajima's D statistic, which was designed to detect ancient complete sweeps (Pybus et al., 2015), in all the populations. Figure 2B depicts the spanning region which presents an ancient complete hard sweep with windows that reach a Tajima's D $\leq$ -2 (1% extreme). This suggests that, despite not observing iHH12 signals in non-African populations, the underlying haplotype pattern reflects a signature of positive selection that includes the global population. No clear phenotype could be associated with this signal, however recent evidence indicates mutations in *TENM1* are linked with congenital general anosmia (Alkelai et al., 2016), suggesting the potential for olfactory adaptations. Previous studies have shown the importance of the olfactory system in the evolution of *Homo sapiens* (Hoover, 2010), olfactory receptors were subjected to non-neutral selection (Hoover, 2015) which accounts for population-specific phenotypic variability (Trimmer et al., 2019). This evidence suggests that olfactory receptors, and the associated neural system, might be subjected to important adaptive processes in human evolutionary history.

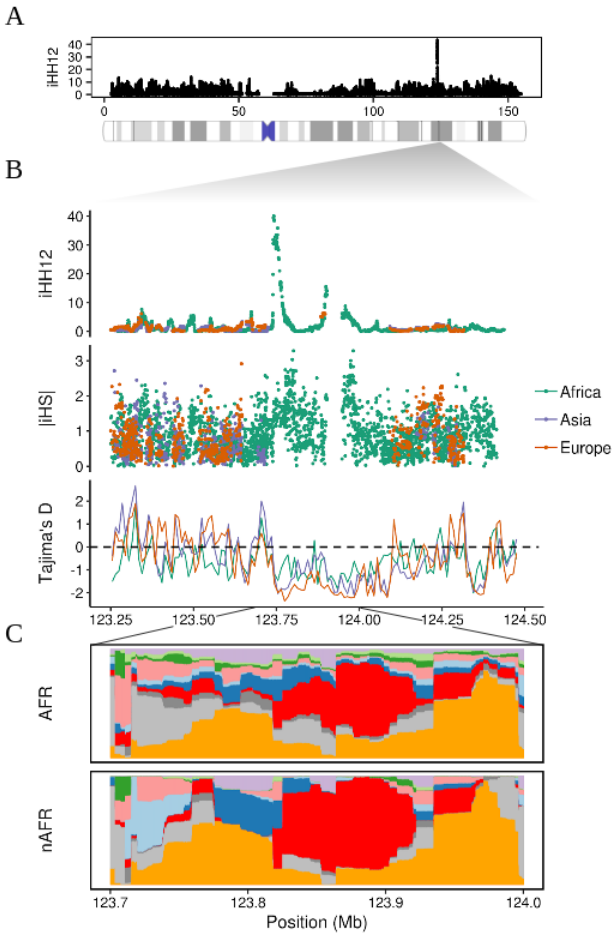


Figure 2. Putative positive selection signal on the *TENM1* gene. **(A)** Whole chromosome iHH12 scores in the global sample. **(B)** Manhattan plot showing the iHH12, iHS and Tajima's D scores on the *TENM1* gene region. **(C)** Clusters of highly similar haplotypes (in red and orange) estimated by *fastPHASE* were found in African (AFR) and non-African populations (nAFR). The different colouring represents changes in the haplotypic composition through the region, where each row represents a haplotype and each column a SNP.

Selection of X-inactivation escape genes

The incomplete inactivation of some genes, during the process of gene dosage compensation in females, might expose these escapees to sex-especific adaptive processes due its biased expression. We wanted to investigate whether patterns of positive selection could be detected amongst the genes that escape from the X chromosome

inactivation. We obtained the X chromosome inactivation (XCI) status in the combined set of populations from Tukiainen et al. (2017). We considered 59 genes as “escape” and 381 genes as “inactive”, keeping only genes with the strongest support. We constructed contingency tables based on these categories performing Fisher's exact test of independence between selection and XCI status for different extreme tail thresholds of the selection tests. We found that genes that escape from the X-inactivation had a higher probability of being targeted by positive selection according to two of the tests. This trend is significant for iHS, is marginally significant for iHH12 and does not reach significance for nSL (Supplementary Tables 4A,B). Notably, escape genes under positive selection had similar proportions from iHS and iHH12 (19% and 20%, respectively; Supplementary Table 4A), however only reached 11% for nSL. This may suggest that escape genes are more likely to be targeted by selection processes that leave signatures with a degree of “softness” closer to hard sweeps rather than soft sweeps.

Supplementary Table 4C lists the genes under selection that escape inactivation. On this list, we found enrichment in sulfuric ester hydrolase activity (GO:0008484), due to the sulfatase group of genes. Among these top candidates, we found four members of the *ARS* family. Three of these members participate in bone and cartilage matrix composition during development (*ARSE*, *ARSD*, *ARSF*). These genes are associated with the X-linked Chondrodysplasia Punctata, a syndrome that affects almost exclusively females, and is characterized by abnormal embryo development, including skeletal malformations, skin abnormalities and cataracts (Franco et al., 1995).

The *STS* gene, also escaping inactivation, presents another highly shared sweeping region among populations (iHS 99.9th percentile in African populations and 99th percentile in Europeans and Asians). It is associated with the X-linked Ichthyosis, a syndrome caused by a placental steroid hormone deficiency and is characterized by skin

and eye abnormalities (Basler et al., 1992). This gene was reported to be one of the top female-biased genes differentially escaping inactivation in Yoruba (YRI) (Johnston et al., 2008). As hypothesized by Tukiainen et al. (2017) most of the escape genes reported as under selection show female-biased expression, suggesting these genes might be involved in some adaptive trait in females.

Functional non-coding regions under positive selection

Previous studies have reported numerous signatures of positive selection with an unknown coding genic cause. This might be accounted by a high false positive rate in genomic scans but also by the presence of signatures in non-genic regions, suggesting that many true signals are located in non-coding, potentially regulatory elements (Fraser, 2013; Enard et al., 2014).

In order to identify the strongest and most interesting candidates of positive selection on the X chromosome, we evaluated the signals in the 99.9th percentile and attempted to pinpoint the target of selection within each signal by annotating SNPs with *ANNOVAR* (Wang et al., 2010). A large portion of single nucleotide polymorphisms (SNPs) over the 1% per-SNP score extreme tail are intergenic, in addition, a large fraction fall within intronic regions for all statistics (iHS: 0.29, iHH12: 0.32, nSL: 0.2), with few in exons or untranslated regions. Combined Annotation Dependent Depletion (CADD) scores (Kircher et al., 2014) were used to identify functional variants according to their deleteriousness (see Methods). After filtering by functionality (CADD $\geq$ 10), the majority of the variants were excluded, however, the SNP composition remained higher in intergenic regions (Supplementary Table 5), with an average prevalence in signals reported by iHH12 and nSL in non-African populations (Africa: ~0.62, Europe: ~0.72, Asia: ~0.9). These results suggest that there is an excess in signals driven by

intergenic SNPs that fall in non-annotated and potentially regulatory regions.

Several intergenic regions are under positive selection in the different continental groups (Supplementary Table 1). In order to assess the functional impact of these signals, we explored the overlap of the extreme SNPs within the 99.9th percentile windows with RegulomeDB (Boyle et al., 2012) annotated elements. The combined signals across all populations had higher proportions of SNPs within an ENCODE element (iHS: 19.1%, iHH12: 26.3%, nSL: 13%) compared to the whole chromosome (5.5%). This enrichment is more prevalent for iHH12 signals, which may be due to its power to detect both hard and soft sweep signatures. This finding shows, as expected, intergenic regions under putative positive selection are enriched in functional elements and likely points to selection of regulatory processes.

Intergenic signals cluster around genic regions, suggesting a regulatory function influencing surrounding genes. Under the 99th percentile, we found instances of genic windows that overlap genic and intergenic SNPs, this is more prevalent in iHH12 and nSL statistics (iHS: 2%, iHH12: 5.7%, nSL: 4.4%) across all populations. Since regulatory elements are expected to be found in the extremes and within coding regions, we used the RegulomeDB annotation to associate the signal of putative selection with any potential regulatory function. In these overlapping regions we found that the odds of intergenic SNPs overlapping a functional element is higher than genic SNPs (Supplementary Table 6A) according to iHH12 and nSL, moreover when considering extreme SNPs (99.9th) these values reflected a much higher dominance of functional intergenic SNPs in these tests (Supplementary Table 6B). These findings indicate that the overlapping genic windows under selection are more enriched in regulatory elements in their intergenic portion, something that points to the presence of sweeps in regulatory elements.

This evidence suggests, as previously noted, amino acid changes may play a less important role in recent adaptation and that regulatory changes may drive a more important part of adaptation events in recent human evolution (Fraser, 2013; Enard et al., 2014; Grossman et al., 2013).

Enhancer-like signatures under positive selection

In order to analyse in more detail the regulatory roles of the regions under putative positive selection, we intersected the intergenic windows in the extreme tails with the enhancer coordinates described in the Human Active Enhancer to interpret Regulatory variants (HACER) database (Wang et al., 2019). Supplementary Table 7 shows the overlapping/non-overlapping windows with enhancer regions (in any cell line) in the 99th percentile extreme tail. As the table shows, the intergenic regions under positive selection are more probable (odds ratio values) to present overlapping enhancers in the case of iHH12.

In several cases these enhancers were located close to genes also reported as positively selected in the analysis. We wanted to determine if this pattern is a by-product of the selection in adjacent regions by genetic linkage (hitchhiking effect), or due to independent selection processes on both elements, the enhancers and their target genes. In order to deal with the different cell-type-specific enhancers described in HACER we created a consensus enhancer dataset (see Methods) with unique coordinates. We pooled all the populations and selection tests in order to maximize the statistical power of our analysis. A Chi-squared test shows the dependency between the selection of the enhancers and their target genes (p-value = 0.0021). However, despite the dependency between these two variables we observe a higher probability of both elements, the enhancer and its closest gene, as being under positive selection (YY category) and not being under

positive selection (NN category) than expected by chance (Supplementary Tables 8A,B). We compared the mean distances between the selected/non-selected enhancers and their selected/non-selected closest genes. These distances do not seem to support the physical genetic linkage as a possible explanation of this association. It must be taken into account that the reported distances are sometimes too large (~2.5Mb) to be the reason for selection by hitchhiking of both elements. Therefore, the YY set of enhancers and target genes must be regions that are jointly swept by hitchhiking (most of them) combined with few regions that are selected by independent processes. This suggests that selective pressures might affect some genes and their regulatory elements in a coordinated way, modifying not only their coding sequence but also their expression level.

Next, we wanted to study the potential origin of some of the most extreme intergenic signals and the regulatory effect of the sweeping haplotypes in the different populations. We focused on the highest scoring candidate enhancers (99.9^{th}) and their closest genes (Supplementary Table 9). Among these candidates, we found at X:73,135,561-73,145,161 an iHS African-shared extreme signal that overlaps an enhancer (Supplementary Figure 6) located in the XIC region (X-inactivation center) and whose closest gene is *JPX*. This region is active in five different cell lines according to HACER (H1, HUVEC, HCT116, AC16, REH) and is supported by three databases (Ensembl Regulatory Build, ENCODE Enhancer-like Regions and ChromHMM). The gene *JPX* (~23kb away) is an activator of the lncRNA *XIST*, which is involved in the X chromosome inactivation. Among the potential causal variants of this signal, the SNP rs112977454 reported as expression quantitative trait loci (eQTL) by the Genotype-Tissue Expression (GTEx) project, is the most likely candidate. In addition, this eQTL has a CADD score of 9.018, close to the 1% pathogenicity threshold (CADD = 10) used by Kircher et al. (2014), and an average derived allele frequency (DAF) of 17% in African populations, while is

absent from the rest of populations. This eQTL is also found overlapping a transcription factor binding site (TFBS) in the HUVEC cell line, which targets the *JPX* gene through the transcription factors *FOS*, *GATA2*, *JUN* and *POLR2A*. No specific phenotype is associated with this variant; however, these results suggest that its segregation in African populations might influence the transcription factor binding and affect the regulation of the *JPX* gene.

Functional analysis of enhancers under positive selection

In order to explore the potential regulatory effect behind the selection processes in the candidate enhancers (Supplementary Table 9), we compared the regulatory activity of the putative haplotype under selection with that of its ancestral sequence. To perform this task, we analyzed the changes in the expression of the reporter gene luciferase under regulation of the two ancestral and derived haplotypes in some of these enhancers. This method allows us to test all the potential causal variants independently on the possibility of testing a passenger variant (not causal) of the sweep. We tested the enhancer regions targeting the genes *HUWE1*, *KDM6A* and *SH2D1A* (Figures 3A,B,C), which also harbor signals of positive selection in their sequences. These genes are implicated in intellectual dissability (*HUWE1*) (Giles and Grill, 2020) and the Duncan disease (*SH2D1A*) (Sumegi et al., 2002), and, in the case of *KDM6A*, this gene is reported as X-inactivation escapee by Tukiainen et al. (2017), which makes it susceptible to participate in sex-specific processes (Dunford et al., 2017, Itoh et al., 2019). In all these cases, the enhancer region overlaps with more than one potential causal SNPs, located almost all of them in the 99th percentile of the selected populations. Ancestral and derived haplotypes of the candidate enhancers were obtained from males of the relevant population under selection and subsequently cloned in a luciferase-reporter vector. Upon transfection in 293T cells, significantly differential luciferase activity amongst the ancestral

and derived haplotypes for *HUWE1* and *KDM6A* enhancers was observed, showing a clear distinction of the regulatory activity between these two haplotypes (Figure 3D). Yet this analysis did not show differential activity between the ancestral and derived form of the *SH2D1A* enhancer. Although no specific phenotypes were able to be assigned to the selection of these regions, our data suggest that positive selection has contributed to the adaptation of different human populations by differentially regulating the expression of certain genes. Further studies will be needed to understand the phenotypic consequences of such adaptations.

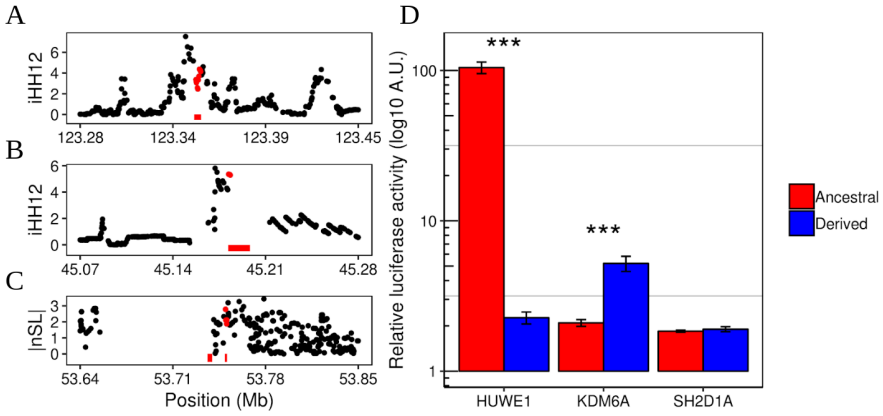


Figure 3. Candidate enhancers under putative positive selection. Manhattan plots show the selection scores overlapping the enhancer coordinates (bottom red bars) targeting *SH2D1A* **(A)**, *KDM6A* **(B)** and *HUWE1* **(C)** genes in YRI, CEU and YRI populations, respectively. Although *HUWE1* appears under positive selection in Gambians (GWD) (Table S9) YRI individuals were used in the luciferase assay instead, since the signal is also present in this population at 99th percentile. Red dots correspond to enhancer overlapping SNPs. **(D)** Relative luciferase activity comparisons between the ancestral and derived haplotypes in each of the candidate enhancers. Significant differential activities are seen in *HUWE1* (p-value = 5.75e10^{-8}) and *KDM6A* (p-value = 0.004) enhancers.

DISCUSSION

In this analysis, we report a comprehensive analysis of recent positive selection in the X chromosome of 15 non-admixed sub-Saharan African, European and East Asian populations. We have focused on the spectrum of signatures captured by the selection statistics iHS, iHH12 and nSL, which are based on the

detection of extended long haplotypes at moderately high and intermediate frequencies (hard and soft sweeps). These three statistics present different approaches and statistical power to detect the different modes of selective sweeps. However, in some cases, the similarity between the haplotypic patterns behind hard and soft sweep signals might lead to the simultaneous detection of the same selected region by these three methods. Results indicate Sub-Saharan African populations have a higher proportion of windows that fall outside the extreme simulation thresholds in comparison with Europeans and Asians. This is directly related to the effect of haplotype-based statistics, in which the number of detectable windows under selection is correlated with the effective population size (Johnson and Voight, 2018; Voight et al., 2006). In contrast with iHS, a higher amount of soft sweep-like signatures is presumably captured by nSL and iHH12 statistics. This was previously noted by authors who claimed that regions targeted by hard sweeps are much less common than soft sweeps (Messer and Petrov, 2013; Schrider and Kern, 2017). Subtle changes of frequency in multiple loci might be also behind numerous quantitative adaptations that would require a more profound and comprehensive analysis than the one conferred by the "sweep" vision (Höllinger et al., 2019). Therefore, it is more likely that genomes, and the human X chromosome in this case, are populated by a greater number of signatures with different degrees of "softness" that are misclassified or overlooked by most selection statistics.

The faster-X effect is believed to act on the X chromosome when the hemizygous state leads to a complete penetrance of mutations, allowing for a quicker and stronger adaptive process. Differences between autosomes and the X chromosome are seen for the nSL statistic in non-African populations, which might suggest some kind of effect that generates the skewed distributions. However, these differences could not easily be associated with the faster-X effect, due to the inconsistencies found in the top 1%. However, as

previously noted by Arbiza et al. (2014), natural selection seems to be a more powerful force in the sexual chromosome than in autosomes, which might explain differences in X/Autosome diversity in human populations. Particular selection events and sex-biased processes might leave specific pronounced signatures in the X chromosome, like we report in this paper. Nevertheless, despite accounting for demography and different mutation rates in our simulations, selection is not the only factor that could be invoked to explain the differences in haplotype diversity.

We report signals of recent positive selection in particular regions of the X chromosome. The difficulty of identifying clear signals from particular selection processes relies on the mixed properties of most signatures. In our scan most of the observed signals are captured by more than one statistic. One of the most remarkable cases of selection in our analysis is the *TENM1* gene. This gene harbours a region of ~300kb (Figure 2) with selection signals that indicate the presence of a haplotype pattern indicating an old and strong event of positive selection before the human populations split. Moreover, the haplotype clusters inferred by *fastPHASE* show a clear predominance of two types of sequences that might derive from a whole unique sweeping haplotype that could be broken by recombination in this hotspot region. Although the role of *TENM1* selection might be linked to recent changes of the olfactory system, the origin of the haplotype patterns seen in our analysis could have more general implications for neural development. Genic regions under putative positive selection seem to be dominated by genes involved in neural development enriched processes. This is widely reported by the three tests used and appear globally distributed in the three continental groups. These findings fit the general picture of previous evolutionary studies which describe the role of neural genes in human recent history (Wei et al., 2019).

Commonalities with previous studies reinforce evidence of X-linked selection in human populations. Despite differences in the

approach, we found complementary results. Indeed, the great diversity between populations in our study has confirmed previously described signals, like selection in *DMD* or reproduction-related genes like the *SPANX* family, and expanded the findings in new populations and genes. It is of interest to remark on the case of the *SPANX* members and other reproduction-related genes reported above. It was previously mentioned the potential importance of fertility-related genes in recent human history (Ramm et al., 2014; Hart et al., 2018). The *SPANX* members, like other cancer-testis (CT) genes (*MAGE* family in the 99th percentile), are known to be under rapid evolution and appear to be subjected to positive selection affecting their coding sequences (Kouprina et al., 2004). Previous reports found members of the spermatogenesis-related family *SPATA* to be under recent positive selection and suggest that testis-enriched genes are the target of population-specific selection (Schrider and Kern, 2017; Schaschl et al., 2020). Other studies report specific ampliconic gene-enriched regions in humans and other primates targeted by strong selective sweeps, where meiotic drive and sperm competition seem to be a potential explanation (Dutheil et al., 2015; Nam et al., 2015). Although an important number of previously reported genes under selection have been captured in our scan, it is important to note that a high false discovery rate is expected from this "hypothesis free" approach. Nonetheless, despite the likely presence of false positives, our findings are in line with previous evidence and supports the importance of reproduction and male fertility in recent human evolutionary history.

The gene dosage compensation of the X chromosome occurs in females by the random inactivation of one of the copies during the early stages of embryogenesis. However, this process of transcriptional silencing is not complete for all the genes. Evidence suggests that around 23% of the X-linked genes "escape" inactivation and express both chromosomal copies. Most of these genes are located in the pseudoautosomal region 1 (PAR1) and only

a small fraction is distributed along the non-pseudoautosomal region (nPAR) (Balaton et al., 2015; Tukiainen et al., 2017) analysed in this study. Overall, our analysis shows an enrichment of genes under selection which escape X-inactivation mainly driven by hard-sweep-like signatures. These genes were previously described as being under purifying selection (Park et al., 2010), however, no evidence for positive selection has been reported until now. Although one could argue that background selection might be behind such a pattern, a recent study has shown that this kind of selection is not expected to mimic the signatures left by selective sweeps (Schrider, 2020). Therefore, these X-linked escape genes are expression-biased between sexes and might be responsible for sexual dimorphic traits, likely producing phenotypic diversity which has been adaptive in females during human evolution. However, more specific analyses on escape genes are needed in order to establish a phenotypic cause for such potential adaptation.

A large fraction of regions under selection have no annotations. We report significant evidence of intergenic regions with high selection scores in the three selection tests, reflecting the presence of signatures that fit the two selection processes we consider in this analysis. Enrichment in the regulatory elements annotated by RegulomeDB is seen globally in the two selection processes, with a higher prevalence in regions exhibiting soft sweep-like signatures (iHH12 and nSL signals). Sometimes genic regions might be affected by the selection of the surrounding intergenic regions that harbour regulatory elements. In our analysis a fraction of selected windows classified as genic have intergenic portions that exhibit a dominance of highly scored SNPs that overlap a functional non-genic element reported by RegulomeDB.

A recent analysis of selection in enhancers revealed that approximately 5.90% of the enhancers studied in different tissues present signatures compatible with recent positive selection events (Moon et al., 2019). Other cases of selection in enhancers have

shown how a SNP subjected to positive selection is able to modify the regulatory activity of the region in a population specific manner (Nakayama et al., 2017). Having this in mind we used the HACER database to study in more detail the potential role of selection in active human enhancers. We show several cases of reported enhancers under selection whose closest gene (also considered target gene) is under putative positive selection in our analysis. This result might reflect a linkage effect between these two elements; however, we suggest that in some cases this is an indication of concurrent selection of the gene and the regulatory region. We report specific cases of putative positive selection signals in enhancers that might drive population-specific regulatory changes. African populations had a highly scoring hard sweep-like signature in an enhancer located in the XIC region. Among the top SNPs we find rs112977454 (99.96^{th} percentile) as an eQTL highly segregated in Africans which might affect the binding of transcription factors that regulate the expression of the lncRNA *JPX*. This gene is a key participant in the X chromosome inactivation as it promotes the expression of *XIST* (Tian et al., 2010), which finally silences the transcription by coating the chromosome into the Barr body. This is an interesting candidate since it might affect the expression patterns of genes that escape from the X-inactivation and thus play a role in the potential adaptations of dimorphic traits hypothesized before.

In order to reveal the potential regulatory effect of our enhancers under selection, we performed luciferase-based assays on three of our top candidates that met the requirements to be cloned. *HUWE1* and *KDM6A* enhancers exhibit a significant difference in the luciferase activity between the two most differentiated haplotypes. This effect clearly suggests a differential regulation of these genes which might fit with the idea of population-specific selection processes. The case of *KDM6A* is rather remarkable since it has been associated with female-specific traits where its ability to escape from the X-inactivation plays a significant role. The biallelic expression of this gene seems to confer a protective effect in

females in a wide range of cancer types, where males are more exposed due to their hemizygous state (Dunford et al., 2017). The same overexpression of *KDM6A* appears to be involved with sex differences in autoimmune disease susceptibility, contributing to a higher incidence of multiple sclerosis in females (Itoh et al., 2019). Although we were not able to make a direct association between our selection signals and these phenotypes, the evident effect of selection in these enhancers and the potential role of adaptations in escape genes suggest that selection might be behind secondary processes that affect women and men in different ways. As for other genomic scans, the power to detect regions under positive selection in our analysis might leave behind more complete patterns that explain in a more comprehensive way the potential adaptations presented here. This, together with the inherent difficulty of identifying the precise target of natural selection, make this type of analysis a challenging aspect in the study of evolution.

Contribution to the field

We conducted a comprehensive analysis of positive selection in human X chromosomes of 15 different human populations, describing remarkable signals in genes involved in neural development and reproduction, as well as extending evidence in previously known gene selection candidates. We also report positive selection in genes that escape X-inactivation and might be behind sex-especific adaptive traits. Regulatory elements appear to be significantly enriched in regions with high selection scores, which provide evidence of the importance of gene regulation in driving adaptation processes. Our work provides new evidence on how positive selection has shaped the diversity of the human X chromosome leading to potentially adaptive changes in recent human history, however more profound and comprehensive analysis and further functional studies are needed in order to understand the phenotypic consequences behind such adaptations.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Authors' contributions

JB, HL conceived the study. PV-M, SA and HL analysed and interpreted the data. PV-M wrote the manuscript. JN, HL, SA, JB revised the manuscript. All authors approved the final manuscript.

Funding

This study has been possible thanks to grant PID2019-110933GB-I00/AEI/10.13039/501100011033 awarded by the Agencia Estatal de Investigación (AEI), Ministerio de Ciencia, Innovación y Universidades (MCIU, Spain) and with the support of Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya (GRC 2017 SGR 702). Part of the "Unidad de Excelencia María de Maeztu", funded by the AEI (CEX2018-000792-M). P.V-M is supported by an FPI PhD fellowship (FPI-BES-2016-077706) part of the "Unidad de Excelencia María de Maeztu" funded by the MINECO (ref: MDM-2014-0370).

List of abbreviations

CADD: Combined Annotation Dependent Depletion
DAF: Derived Allele Frequency
EHH: Extended Haplotype Homozygosity
eQTL: Expression Quantitative Trait Loci
GO: Gene Ontology
HACER: Human Active Enhancer to interpret Regulatory variants database

HMM: Hidden Markov Model
iHH12: Integrated Haplotype Homozygosity pooled test
iHS: Integrated Haplotype Score test
Kb: Kilobases
LD: Linkage disequilibrium
MAF: Minor Allele Frequency
Mb: Megabases
nPAR: Non-pseudoautosomal region
nSL: Number of segregating sites by length test
OEA: Overrepresentation Enrichment Analysis
PAR: Pseudoautosomal region
SFS: Site Frequency Spectrum
SNP: Single Nucleotide Polymorphism
TFBS: Transcription Factor Binding Site
XCI: X Chromosome Inactivation
XIC: X-inactivation center.

Acknowledgments

The authors would like to thank the contribution of Andrea Martí Sarrias in the performance of the luciferase assays of enhancers under positive selection.

References

Ahmad, H. I., Liu, G., Jiang, X., Edallew, S. G., Wassie, T., Tesema, B., et al. (2017). Maximum-likelihood approaches reveal signatures of positive selection in BMP15 and GDF9 genes modulating ovarian function in mammalian female fertility. Ecol. Evol. 7, 8895–8902. doi:10.1002/ece3.3336.

Alkelai, A., Olender, T., Haffner-Krausz, R., Tsoory, M. M., Boyko, V., Tatarskyy, P., et al. (2016). A role for TENM1 mutations in congenital general anosmia. Clin. Genet. 90, 211–219. doi:10.1111/cge.12782.

Arbiza, L., Gottipati, S., Siepel, A., and Keinan, A. (2014). Contrasting X-Linked and Autosomal Diversity across 14 Human Populations. Am. J. Hum. Genet. 94, 827–844. doi:10.1016/j.ajhg.2014.04.011.

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Bentley, D. R., Chakravarti, A., et al. (2015). A global reference for human genetic variation. Nature 526, 68–74. doi:10.1038/nature15393.

Balaton, B. P., Cotton, A. M., and Brown, C. J. (2015). Derivation of consensus inactivation status for X-linked genes from genome-wide studies. Biol. Sex Differ. 6, 35. doi:10.1186/s13293-015-0053-7.

Basler, E., Grompe, M., Parenti, G., Yates, J., and Ballabio, A. (1992). Identification of point mutations in the steroid sulfatase gene of three patients with X-linked ichthyosis. Am. J. Hum. Genet. 50, 483–491.

Bersaglieri, T., Sabeti, P. C., Patterson, N., Vanderploeg, T., Schaffner, S. F., Drake, J. A., et al. (2004). Genetic signatures of strong recent positive selection at the lactase gene. Am. J. Hum. Genet. 74, 1111–1120. doi:10.1086/421051.

Biswas, S., and Akey, J. M. (2006). Genomic insights into positive selection. Trends Genet. 22, 437–446. doi:10.1016/j.tig.2006.06.005.

Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., et al. (2012). Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 22, 1790–7. doi:10.1101/gr.137323.112.

Carrel, L., and Willard, H. F. (2005). X-inactivation profile reveals extensive variability in X-linked gene expression in females. Nature 434, 400–404. doi:10.1038/nature03479.

Casillas, S., Mulet, R., Villegas-Mirón, P., Hervas, S., Sanz, E., Velasco, D., et al. (2018). PopHuman: The human population genomics browser. Nucleic Acids Res. 46, D1003–D1010. doi:10.1093/nar/gkx943.

Casto, A. M., Li, J. Z., Absher, D., Myers, R., Ramachandran, S., and Feldman, M. W. (2010). Characterization of X-linked SNP genotypic variation in globally distributed human populations. Genome Biol 11, R10. doi:10.1186/gb-2010-11-1-r10.

Cotton, A. M., Ge, B., Light, N., Adoue, V., Pastinen, T., and Brown, C. J. (2013). Analysis of expressed SNPs identifies variable extents of expression from the human inactive X chromosome. Genome Biol. 14, R122. doi:10.1186/gb-2013-14-11-r122.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. Bioinformatics 27, 2156–2158. doi:10.1093/bioinformatics/btr330.

Dunford, A., Weinstock, D. M., Savova, V., Schumacher, S. E., Cleary, J. P., Yoda, A., et al. (2017). Tumor-suppressor genes that escape from X-inactivation contribute to cancer sex bias. Nat. Genet. 49, 10–16. doi:10.1038/ng.3726.

Dutheil, J. Y., Munch, K., Nam, K., Mailund, T., and Schierup, M. H. (2015). Strong Selective Sweeps on the X Chromosome in the Human-Chimpanzee Ancestor Explain Its Low Divergence. PLoS Genet. 11, e1005451. doi:10.1371/journal.pgen.1005451.

Enard, D., Messer, P. W., and Petrov, D. A. (2014). Genome-wide signals of positive selection in human evolution. Genome Res. 24, 885–895. doi:10.1101/gr.164822.113.

Ewing, G., and Hermisson, J. (2010). MSMS: A coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics 26, 2064–2065. doi:10.1093/bioinformatics/btq322.

Fan, S., Hansen, M. E. B., Lo, Y., and Tishkoff, S. A. (2016). Going global by adapting local: A review of recent human adaptation. Science (80-.). 354, 54–59. doi:10.1126/science.aaf5098.

Ferrer-Admetlla, A., Liang, M., Korneliussen, T., and Nielsen, R. (2014). On detecting incomplete soft or hard selective sweeps using haplotype structure. Mol. Biol. Evol. 31, 1275–1291. doi:10.1093/molbev/msu077.

Field, Y., Boyle, E. A., Telis, N., Gao, Z., Gaulton, K. J., Golan, D., et al. (2016). Detection of human adaptation during the past 2000 years. Science (80-.). 354, 760–764. doi:10.1126/science.aag0776.

Flaquer, A., Rappold, G. A., Wienker, T. F., and Fischer, C. (2008). The human pseudoautosomal regions: A review for genetic epidemiologists. Eur. J. Hum. Genet. 16, 771–779. doi:10.1038/ejhg.2008.63.

Franco, B., Meroni, G., Parenti, G., Levilliers, J., Bernard, L., Gebbia, M., et al. (1995). A cluster of sulfatase genes on Xp22.3: Mutations in chondrodysplasia punctata (CDPX) and implications for warfarin embryopathy. Cell 81, 15–25. doi:10.1016/0092-8674(95)90367-4.

Fraser, H. B. (2013). Gene expression drives local adaptation in humans. Genome Res. 23, 1089–1096. doi:10.1101/gr.152710.112.

Fu, W., and Akey, J. M. (2013). Selection and adaptation in the human genome. Annu. Rev. Genomics Hum. Genet. 14, 467–489. doi:10.1146/annurev-genom-091212-153509.

Garud, N. R., Messer, P. W., Buzbas, E. O., and Petrov, D. A. (2015). Recent Selective Sweeps in North American Drosophila melanogaster Show Signatures of Soft Sweeps. PLoS Genet. 11, 1–32. doi:10.1371/journal.pgen.1005004.

Giles, A. C., and Grill, B. (2020). Roles of the HUWE1 ubiquitin ligase in nervous system development, function and disease. Neural Dev. 15, 6. doi:10.1186/s13064-020-00143-9.

Grossman, S. R., Andersen, K. G., Shlyakhter, I., Tabrizi, S., Winnicki, S., Yen, A., et al. (2013). Identifying recent adaptations in large-scale genomic data. Cell 152, 703–713. doi:10.1016/j.cell.2013.01.035.

Hart, M. W., Stover, D. A., Guerra, V., Mozaffari, S. V., Ober, C., Mugal, C. F., et al. (2018). Positive selection on human gamete-recognition genes. PeerJ 6, e4259. doi:10.7717/peerj.4259.

Henn, B. M., Botigué, L. R., Bustamante, C. D., Clark, A. G., and Gravel, S. (2015). Estimating the mutation load in human genomes. Nat. Rev. Genet. 16, 333–343. doi:10.1038/nrg3931.

Hermisson, J., and Pennings, P. S. (2017). Soft sweeps and beyond: understanding the patterns and probabilities of selection footprints under rapid adaptation. Methods Ecol. Evol. 8, 700–716. doi:10.1111/2041-210X.12808.

Hernandez, R. D., Kelley, J. L., Elyashiv, E., Melton, S. C., Auton, A., McVean, G., et al. (2011). Classic selective sweeps were rare in

recent human evolution. Science (80-.). 331, 920–924. doi:10.1126/science.1198878.

Höllinger, I., Pennings, P. S., and Hermisson, J. (2019). Polygenic adaptation: From sweeps to subtle frequency shifts. PLOS Genet. 15, e1008035. doi:10.1371/journal.pgen.1008035.

Holmes, R. S. (2017). Comparative and evolutionary studies of mammalian arylsulfatase and sterylsulfatase genes and proteins encoded on the X-chromosome. Comput. Biol. Chem. 68, 71–77. doi:10.1016/j.compbiolchem.2017.02.009.

Hoover, K. C. (2010). Smell with inspiration: The evolutionary significance of olfaction. Am. J. Phys. Anthropol. 143, 63–74. doi:10.1002/ajpa.21441.

Hoover, K. C., Gokcumen, O., Qureshy, Z., Bruguera, E., Savangsuksa, A., Cobb, M., et al. (2015). Global survey of variation in a human olfactory receptor gene reveals signatures of non-neutral evolution. Chem. Senses 40, 481–488. doi:10.1093/chemse/bjv030.

Itoh, Y., Golden, L. C., Itoh, N., Matsukawa, M. A., Ren, E., Tse, V., et al. (2019). The X-linked histone demethylase Kdm6a in CD4+ T lymphocytes modulates autoimmunity. J. Clin. Invest. 129, 3852–3863. doi:10.1172/JCI126250.

John, P. R., Makova, K., Li, W. H., Jenkins, T., and Ramsay, M. (2003). DNA polymorphism and selection at the melanocortin-1 receptor gene in normally pigmented Southern African individuals. in Annals of the New York Academy of Sciences (New York Academy of Sciences), 299–306. doi:10.1111/j.1749-6632.2003.tb03193.x.

Johnson, K. E., and Voight, B. F. (2018). Patterns of shared signatures of recent positive selection across human populations. Nat. Ecol. Evol. 2, 713–720. doi:10.1038/s41559-018-0478-6.

Johnston, C. M., Lovell, F. L., Leongamornlert, D. A., Stranger, B. E., Dermitzakis, E. T., and Ross, M. T. (2008). Large-scale population study of human cell lines indicates that dosage compensation is virtually complete. PLoS Genet. 4, 0088–0098. doi:10.1371/journal.pgen.0040009.

Karolchik, D., Hinricks, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., et al. (2004). The UCSC table browser data retrieval tool. Nucleic Acids Res. 32. doi:10.1093/nar/gkh103.

King, M. C., and Wilson, A. C. (1975). Evolution at two levels in humans and chimpanzees. Science (80-.). 188, 107–116. doi:10.1126/science.1090005.

Kircher, M., Witten, D. M., Jain, P., O'roak, B. J., Cooper, G. M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. 46, 310–315. doi:10.1038/ng.2892.

Kouprina, N., Mullokandov, M., Rogozin, I. B., Collins, N. K., Solomon, G., Otstot, J., et al. (2004). The SPANX gene family of cancer/testis-specific antigens: Rapid evolution and amplification in African great apes and hominids. Proc. Natl. Acad. Sci. U. S. A. 101, 3077–3082. doi:10.1073/pnas.0308532100.

Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z., and Zhang, B. (2019). WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. Nucleic Acids Res. 47, W199–W205. doi:10.1093/nar/gkz401.

Lopez, M., Choin, J., Sikora, M., Siddle, K., Harmant, C., Costa, H. A., et al. (2019). Genomic Evidence for Local Adaptation of Hunter-Gatherers to the African Rainforest. Curr. Biol. 29, 2926-2935.e4. doi:10.1016/j.cub.2019.07.013.

Mank, J. E., Vicoso, B., Berlin, S., and Charlesworth, B. (2010). Effective population size and the Faster-X effect: Empirical results and their interpretation. Evolution (N. Y). 64, 663–674. doi:10.1111/j.1558-5646.2009.00853.x.

Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S. A., et al. (2015). Genome-wide patterns of selection in 230 ancient Eurasians. Nature 528, 499–503. doi:10.1038/nature16152.

Meisel, R. P., and Connallon, T. (2013). The faster-X effect: Integrating theory and data. Trends Genet. 29, 537–544. doi:10.1016/j.tig.2013.05.009.

Messer, P. W., and Petrov, D. A. (2013). Population genomics of rapid adaptation by soft selective sweeps. Trends Ecol. Evol. 28, 659–669. doi:10.1016/j.tree.2013.08.003.

Moon, J. M., Capra, J. A., Abbot, P., and Rokas, A. (2019). Signatures of recent positive selection in enhancers across 41 human tissues. G3 Genes, Genomes, Genet. 9, 2761–2774. doi:10.1534/g3.119.400186.

Nakayama, K., Ohashi, J., Watanabe, K., Munkhtulga, L., and Iwamoto, S. (2017). Evidence for very recent positive selection in mongolians. Mol. Biol. Evol. 34, 1936–1946. doi:10.1093/molbev/msx138.

Nam, K., Munch, K., Hobolth, A., Dutheil, J. Y., Veeramah, K. R., Woerner, A. E., et al. (2015). Extreme selective sweeps

independently targeted the X chromosomes of the great apes. Proc. Natl. Acad. Sci. U. S. A. 112, 6413–6418. doi:10.1073/pnas.1419306112.

Park, C., Carrel, L., and Makova, K. D. (2010). Strong purifying selection at genes escaping X chromosome inactivation. Mol. Biol. Evol. 27, 2446–2450. doi:10.1093/molbev/msq143.

Pavlidis, P., and Alachiotis, N. (2017). A survey of methods and tools to detect recent and strong positive selection. J. Biol. Res. 24, 7. doi:10.1186/s40709-017-0064-0.

Pickrell, J. K., Coop, G., Novembre, J., Kudaravalli, S., Li, J. Z., Absher, D., et al. (2009). Signals of recent positive selection in a worldwide sample of human populations. Genome Res. 19, 826–837. doi:10.1101/gr.087577.108.

Pybus, M., Luisi, P., Dall'Olio, G. M., Uzkudun, M., Laayouni, H., Bertranpetit, J., et al. (2015). Hierarchical boosting: A machine-learning framework to detect and classify hard selective sweeps in human populations. Bioinformatics 31, 3946–3952. doi:10.1093/bioinformatics/btv493.

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842. doi:10.1093/bioinformatics/btq033.

R Core Team (2019). R: A language and environment for statistical computing. R Found. Stat. Comput. Available at: https://www.r-project.org/.

Rallapalli, P. M., Orengo, C. A., Studer, R. A., and Perkins, S. J. (2014). Positive selection during the evolution of the blood coagulation factors in the context of their disease-causing

mutations. Mol. Biol. Evol. 31, 3040–3056. doi:10.1093/molbev/msu248.

Ramm, S. A., Schärer, L., Ehmcke, J., and Wistuba, J. (2014). Sperm competition and the evolution of spermatogenesis. Mol. Hum. Reprod. 20, 1169–1179. doi:10.1093/molehr/gau070.

Sabeti, P. C., Schaffner, S. F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., et al. (2006). Positive natural selection in the human lineage. Science (80-.). 312, 1614–1620. doi:10.1126/science.1124309.

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., et al. (2002). Detecting recent positive selection in the human genome from haplotype structure. Nature 419, 832–837. doi:10.1038/nature01140.

Schaschl, H., and Wallner, B. (2020). Population-specific, recent positive directional selection suggests adaptation of human male reproductive genes to different environmental conditions. BMC Evol. Biol. 20, 27. doi:10.1186/s12862-019-1575-0.

Scheet, P., and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. Am. J. Hum. Genet. 78, 629–644. doi:10.1086/502802.

Scheinfeldt, L. B., and Tishkoff, S. A. (2013). Recent human adaptation: Genomic approaches, interpretation and insights. Nat. Rev. Genet. 14, 692–702. doi:10.1038/nrg3604.

Schrider, D. R. (2020). Background selection does not mimic the patterns of genetic diversity produced by selective sweeps. Genetics 216, 499–519. doi:10.1534/genetics.120.303469.

Schrider, D. R., and Kern, A. D. (2017). Soft sweeps are the dominant mode of adaptation in the human genome. Mol. Biol. Evol. 34, 1863–1877. doi:10.1093/molbev/msx154.

Schrider, D. R., Mendes, F. K., Hahn, M. W., and Kern, A. D. (2015). Soft shoulders ahead: Spurious signatures of soft and partial selective sweeps result from linked hard sweeps. Genetics 200, 267–284. doi:10.1534/genetics.115.174912.

Sumegi, J., Seemayer, T. A., Huang, D., Davis, J. R., Morra, M., Gross, T. G., et al. (2002). A spectrum of mutations in SH2D1A that causes x-linked lymphoproliferative disease and other Epstein-Barr virus-associated illnesses. Leuk. Lymphoma 43, 1189–1201. doi:10.1080/10428190290026240.

Szpiech, Z. A., and Hernandez, R. D. (2014). Selscan: An efficient multithreaded program to perform EHH-based scans for positive selection. Mol. Biol. Evol. 31, 2824–2827. doi:10.1093/molbev/msu211.

Tian, D., Sun, S., and Lee, J. T. (2010). The long noncoding RNA, Jpx, Is a molecular switch for X chromosome inactivation. Cell 143, 390–403. doi:10.1016/j.cell.2010.09.049.

Torres, R., Szpiech, Z. A., and Hernandez, R. D. (2018). Human demographic history has amplified the effects of background selection across the genome. PLOS Genet. 14, e1007387. doi:10.1371/journal.pgen.1007387.

Trimmer, C., Keller, A., Murphy, N. R., Snyder, L. L., Willer, J. R., Nagai, M. H., et al. (2019). Genetic variation across the human olfactory receptor repertoire alters odor perception. Proc. Natl. Acad. Sci. U. S. A. 116, 9475–9480. doi:10.1073/pnas.1804106115.

Tukiainen, T., Villani, A. C., Yen, A., Rivas, M. A., Marshall, J. L., Satija, R., et al. (2017). Landscape of X chromosome inactivation across human tissues. Nature 550, 244–248. doi:10.1038/nature24265.

Urizar-Arenaza, I., Osinalde, N., Akimov, V., Puglia, M., Muñoa-Hoyos, I., Gianzo, M., et al. (2020). SPANX-A/D protein subfamily plays a key role in nuclear organisation, metabolism and flagellar motility of human spermatozoa. Sci. Rep. 10, 5625. doi:10.1038/s41598-020-62389-x.

Vallender, E. J. (2004). Positive selection on the human genome. Hum. Mol. Genet. 13, R245–R254. doi:10.1093/hmg/ddh253.

Veeramah, K. R., Gutenkunst, R. N., Woerner, A. E., Watkins, J. C., and Hammer, M. F. (2014). Evidence for increased levels of positive and negative selection on the X chromosome versus autosomes in humans. Mol. Biol. Evol. 31, 2267–2282. doi:10.1093/molbev/msu166.

Vicoso, B., and Charlesworth, B. (2006). Evolution on the X chromosome: Unusual patterns and processes. Nat. Rev. Genet. 7, 645–653. doi:10.1038/nrg1914.

Voight, B. F., Kudaravalli, S., Wen, X., and Pritchard, J. K. (2006). A map of recent positive selection in the human genome. PLoS Biol. 4, 0446–0458. doi:10.1371/journal.pbio.0040072.

Walsh, S., Pagani, L., Xue, Y., Laayouni, H., Tyler-Smith, C., and Bertranpetit, J. (2020). Positive selection in admixed populations from Ethiopia. BMC Genet. 21, 108. doi:10.1186/s12863-020-00908-5.

Wang, J., Dai, X., Berry, L. D., Cogan, J. D., Liu, Q., and Shyr, Y. (2019). HACER: An atlas of human active enhancers to interpret

regulatory variants. Nucleic Acids Res. 47, D106–D112. doi:10.1093/nar/gky864.

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 38, e164–e164. doi:10.1093/nar/gkq603.

Wang, L., Sinnott-Armstrong, N., Wagschal, A., Wark, A. R., Camporez, J. P., Perry, R. J., et al. (2020). A MicroRNA Linking Human Positive Selection and Metabolic Disorders. Cell 183, 684-701.e14. doi:10.1016/j.cell.2020.09.017.

Wei, Y., de Lange, S. C., Scholtens, L. H., Watanabe, K., Ardesch, D. J., Jansen, P. R., et al. (2019). Genetic mapping and evolutionary analysis of human-expanded cognitive networks. Nat. Commun. 10, 4839. doi:10.1038/s41467-019-12764-8.

Westbrook, V. A., Schoppee, P. D., Vanage, G. R., Klotz, K. L., Diekman, A. B., Flickinger, C. J., et al. (2006). Hominoid-specific SPANXA/D genes demonstrate differential expression in individuals and protein localization to a distinct nuclear envelope domain during spermatid morphogenesis. Mol. Hum. Reprod. 12, 703–716. doi:10.1093/molehr/gal079.

Williamson, S. H., Hubisz, M. J., Clark, A. G., Payseur, B. A., Bustamante, C. D., and Nielsen, R. (2007). Localizing recent adaptive evolution in the human genome. PLoS Genet. 3, 0901–0915. doi:10.1371/journal.pgen.0030090.

Yates, A. D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., et al. (2020). Ensembl 2020. Nucleic Acids Res. 48, D682–D688. doi:10.1093/nar/gkz966.

Enrichment in intronic enhancers controlling the expression of genes involved in tissue-specific functions and homeostasis

Beatrice Borsari, Pablo Villegas-Mirón, Silvia Perez-Lluch, Isabel Turpin, Hafid Laayouni, Alba Segarra-Casas, Jaume Bertranpetit, Roderic Guigo, Sandra Acosta

Submitted for publication

Preprint citation reference of the first version (currently under second revision):

Borsari B, Villegas-Mirón P, Laayouni H, Segarra-Casas A, Bertranpetit J, Guigó R, Acosta S. 2020. Intronic enhancers regulate the expression of genes involved in tissue-specific functions and homeostasis. bioRxiv doi: https://doi.org/10.1101/2020.08.21.260836

Enrichment in intronic enhancers controlling the expression of genes involved in tissue-specific functions and homeostasis

Beatrice Borsari[1*], Pablo Villegas-Mirón[2*], Silvia Perez-Lluch[1], Isabel Turpin[2], Hafid Laayouni[2,4], Alba Segarra-Casas[2], Jaume Bertranpetit[2], Roderic Guigo[1,3], Sandra Acosta [†5]

[1] Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona 08003, Catalonia, Spain
[2] Institut de Biologia Evolutiva (UPF-CSIC), Universitat Pompeu Fabra, Dr. Aiguader 88, 08003, Barcelona, Catalonia, Spain
[3] Universitat Pompeu Fabra (UPF), Dr. Aiguader 88, Barcelona 08003, Catalonia, Spain
[4] Bioinformatic Studies, ESCI-UPF, Pujades 1, 08003, Barcelona, Spain
[5] Dpt. Pathology and Experimental Therapeutics, Medical School, University of Barcelona, Feixa Llarga, 08907, L'Hospitalet de Llobregat, Barcelona, Spain

* Beatrice Borsari and Pablo Villegas-Mirón contributed equally to this work.
† Corresponding address: Dr.Aiguader 88 room 763, 08003, Barcelona, Spain sandra.acosta@upf.edu

Running title: Intronic enhancers lead tissue-specific regulation

Keywords: enhancers, introns, gene regulation, tissue function, tissue patterning

Abstract

Tissue function and homeostasis reflect the gene expression signature by which the combination of ubiquitous and tissue-specific genes contribute to the tissue maintenance and stimuli-responsive function. Enhancers are central to control this tissue-specific gene expression pattern. Here, we explore the correlation between the genomic location of enhancers and their role in tissue-specific gene expression. We found that enhancers showing tissue-specific activity are highly enriched in intronic regions and regulate the expression of genes involved in tissue-specific functions, while housekeeping genes are more often controlled by intergenic enhancers, common to many tissues. Notably, an intergenic-to-intronic active enhancers continuum is observed in the transition from developmental to adult stages: the most differentiated tissues present higher rates of intronic enhancers, while the lowest rates are observed in embryonic stem cells. Altogether, our results suggest that the genomic location of active enhancers is key for the tissue-specific control of gene expression.

Introduction

Multiple layers of molecular and cellular events tightly control the level, time and spatial distribution of expression of a particular gene. This wide range of mechanisms, known as gene regulation, defines tissue-specific gene expression signatures (Melé et al., 2015), which account for all the processes controlling the tissue function and maintenance, namely tissue homeostasis. Both the level and spatio-temporal pattern of expression of a gene are determined by a combination of regulatory elements (REs) controlling its transcriptional activation. Most genes contributing to tissue-specific expression signatures are actively transcribed in more than one tissue, but at different levels and with distinct patterns of expression in time and space, suggesting that the regulation of these genes is different across tissues. Nevertheless, approximately 10-20% of all genes are ubiquitously expressed *housekeeping genes*, and they are involved in basic cell maintenance functions (Pervouchine et al., 2015; Zabidi et al., 2015; Eisenberg and Levanon, 2013).

cis-REs (CREs) are distributed across the whole genome, and their histone signature correlates with the transcriptional control they exert over their target genes (Chen et al., 2019; Hawkins et al., 2010; Choukrallah et al., 2015). The activation of CREs depends on several epigenetic features, including combinations of different transcription factors' binding sites, and it is positively correlated with the H3K27ac histone modification signal (Heinz et al., 2015; Heintzman et al., 2007). Epigenetic features in specific tissues may change throughout the life-span of individuals. During development, embryos undergo dramatic morphological and functional changes. These changes shape cell fate and identity as a result of tightly regulated transcriptional programs, which in turn are intimately associated with CREs' activity and chromatin dynamics (Shlyueva et al., 2014; Bonev et al., 2017; Rand and Cedar, 2003; Gilbert et al., 2003).

Notably, key CREs known to regulate gene expression have been reported to locate in introns of their target genes (Ott et al., 2009; Kawase et al., 2011). However, it is unknown whether this is either a sporadic feature associated with certain types of genes - for instance long genes, such as HBB (β-globin) (Gillies et al., 1983) or CFTR (Ott et al., 2009) -, a common regulatory mechanism to most genes (Khandekar et al., 2007; Levine, 2010), or a pattern of biological significance. To delve into this question, we analyzed the genomic location of CREs across a panel of 70 adult and embryonic human cell types available from the Encyclopedia of DNA Elements (ENCODE) Project (Abascal et al., 2020).

Results

Enhancer-like regulatory elements define tissue-specific signatures

We leveraged the cell type-agnostic registry of candidate *cis*-Regulatory Elements (cCREs) generated for the human genome (hg19) by the ENCODE Project. We focused on the set of 991,173 cCREs classified as Enhancer-Like Signatures (ELSs), defined as DNAse I hypersensitive sites supported by the $H3K27ac^-$ epigenetic signal, and assessed their presence-absence patterns across 43 adult cell type-specific catalogues (Supplementary Table 1; see Methods). We first explored the data with multidimensional scaling (MDS), which uncovered tissue-specific presence-absence patterns (Supplementary Fig. 1A). Indeed, the separation of samples driven by ELSs' activity was comparable to the one obtained from the analysis of Genotype-Tissue Expression (GTEx) data (Melé et al., 2015), with blood and brain as the most diverging tissues. This suggests a correlation between gene regulatory mechanisms orchestrated by ELSs and tissue-specific gene expression patterns, which has been previously described (Pennacchio et al., 2007; Ernst et al., 2011).

Interestingly, we observed that the proportion of active ELSs located in intergenic regions increases with the number of samples in which ELSs are active (Fig. 1A), suggesting an unexpected role for the genomic location of ELSs. Thus, to untangle the relationship between the genomic location and cell-type specificity of ELSs, we selected a subset of 33 samples that formed 9 main tissue groups, supported by both hierarchical clustering and MDS proximity: brain, iPSCs, blood, digestive system, intestinal mucosa, fibro/myoblasts, aorta, skeletal/cardiac muscle and smooth muscle (Figs. 1B-C; Supplementary Table 1, *Samples' Cluster*). Tissues represented by only one sample (ovary, thyroid gland, lung, esophagus, spleen), or samples that did not cluster consistently with their tissue of origin and function (endocrine pancreas, liver, right lobe of liver, gastrocnemius medialis, bipolar neuron), were not included in the subsequent analyses (Supplementary Table 1; see Methods).

The fact that tissue-specific enhancer signatures contribute to the *ad hoc* tissues' functional clustering suggests a direct link between ELSs' activity and the regulation of tissue-specific functions (Fig. 1C). Thus, we set out to characterize tissue-specific enhancer signatures and to compare them with regulatory mechanisms that are common, i.e. shared among most tissues. Tissue-specific ELSs were defined as those ELSs active in ≥ 80\% of the samples within a given cluster and in at most one sample outside the cluster (Supplementary Table 2; see Methods). For clusters with reduced sample size (≤ 3), we required tissue-specific ELSs to be active exclusively within the corresponding tissue cluster (see Methods). The overlap of tissue-specific ELSs with samples from other clusters (Fig. 1D) is consistent with the samples' MDS proximity observed in Fig. 1C, suggesting a functional relevance of the genes regulated by shared ELSs. In addition, we identified a set of 555 ELSs active in 95% of the 33 samples, herein named as common ELSs (Supplementary Table 2).

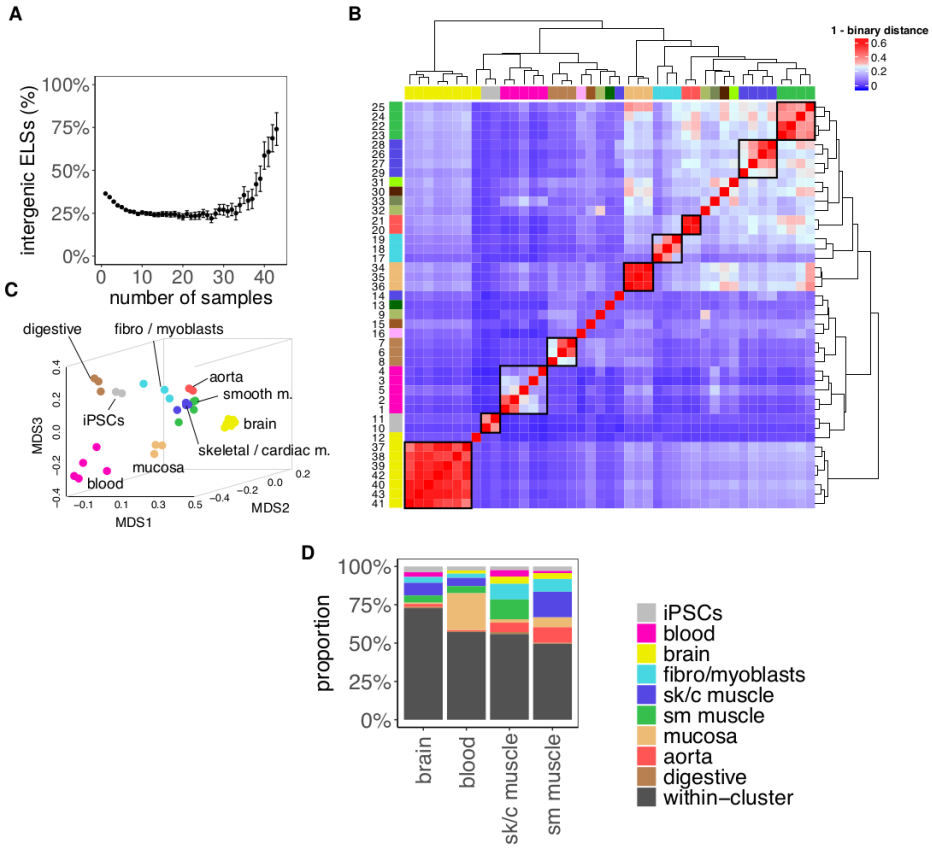


Fig. 1. **A**: Highly-shared ELSs are more frequently located in intergenic regions. The scatter plot represents the proportion of intergenic ELSs active in increasing numbers of human adult samples. Error bars represent the 95% confidence interval. **B**: Samples' clustering defined by ELSs' presence-absence patterns. The heatmap depicts the binary distance between any pair of samples, based on the activity of 921,166 ELSs from any annotated TSS. The correspondence between samples and numbers is reported in Supplementary Table 1 in Supplementary File.pdf. **C**: MDS distribution of human adult samples defined by ELSs' activity. Analogous representation to Supplementary Fig. 1A in Supplementary File.pdf for the subset of 33 selected adult human samples. **D**: Tissue-specific ELSs. The barplot represents the type of samples found within sets of brain-, blood- and muscle-specific ELSs. Most tissue-specific ELSs are only active in the samples of the corresponding cluster ("within-cluster", black), but a few of them may be active in at most one outer sample (i.e. a sample that does not belong to the tissue cluster, coloured). IPSCs-, fibro/myoblasts-, digestive-, mucosa- and aorta-specific ELSs are not represented, since we did not allow outer samples given their small cluster sizes (see Methods).

The genomic location of regulatory elements correlates with their tissue-homeostatic functions

We next explored the genomic location of the sets of common and tissue-specific ELSs. While common ELSs are preferentially located in intergenic regions (58%, Fig. 2A), the majority of aorta, muscle- and brain-specific ELSs fall inside introns (between 63 and 74%; Fig. 2A). These significant differences in genomic distribution between tissue-specific and common regulatory elements (Supplementary Table 3) are consistent with our initial observation of a high sharing rate of intergenic ELSs across samples (Fig. 1A). In contrast, the iPSCs, fibro/myoblasts, mucosa, digestive and blood clusters - which comprise undifferentiated, non-specialized, highly proliferative or more heterogeneous cell types, respectively - showed a more even distribution of tissue-specific ELSs between intergenic and intronic regions (Fig. 2A). Overall, we observed a limited abundance of exonic ELSs (Fig. 2A, Supplementary Tables 3 and 4).

Genes harboring tissue-specific ELSs may present distinctive features, including differences in gene and intron length. To rule out any bias in our analyses, we compared these features between genes hosting common and tissue-specific ELSs. While the number of introns per hosting gene was comparable across groups (Kruskal-Wallis p value test = 0.08), we reported significant differences in gene and median intron length amongst tissues (Kruskal-Wallis p value test < 2.2e-16; Supplementary Fig. S1B). Nevertheless, we did not observe a correlation between such differences and the presence of intronic ELSs (Supplementary Fig. 1B).

We subsequently explored whether the genes harboring tissue-specific intronic ELSs perform functions associated with maintenance of tissue homeostasis and response to stimuli. Indeed, the enrichment of Gene Ontology (GO) terms associated with

tissue-specific cellular components is consistent with the ELSs' tissue identity (Supplementary Table 5). For instance, genes hosting brain-specific ELSs perform functions associated with synapses and axons, while in the case of muscle and blood we found significant terms related to sarcolemma, actin cytoskeleton and contractile fibers, and immunological synapses and cell membranes, respectively. Conversely, genes harboring common ELSs reported terms related to ordinary cell functions and membrane composition (Supplementary Table 5). Although this suggests an implication of intronic ELSs in tissue-specific functions, likely through tissue-specific gene regulation mechanisms, there is no proven evidence of intronic ELSs being direct regulators of their host genes. To identify genes targeted by tissue-specific ELSs, we integrated our ELS analysis with the catalogue of expression Quantitative Trait Loci (eQTLs) provided by the Genotype-Tissue Expression (GTEx) Project (Aguet et al., 2017). eQTLs provide functional information about the changes of expression associated with human variants. We leveraged eQTLs located in both intronic and intergenic ELSs to identify their target genes. Among the 48,555 common and tissue-specific ELSs, 6,349 overlap with a significantly associated eQTL-eGene pair, hereafter referred to as eQTL-ELSs. The proportion of eQTL-ELSs was similar among the tissue samples represented in the GTEx sampling collection, ranging between 10 and 25% (Fig. 2B). In all annotated tissues, gene regulation driven by eQTL-ELSs occurs predominantly in the tissue where the ELS is specifically active (Fig. 2C). In line with the above-mentioned results (Fig. 2A), highly specialized tissues such as brain and muscle show the highest proportion of intronic *vs* intergenic ELSs hosting eQTLs detected in the corresponding tissue (Fig. 2B,C). Conversely, common eQTL-ELSs were more frequently located in intergenic elements (32% *vs* 62%) (Fig. 2C). GO enrichment analysis on the sets of target genes associated with intronic and intergenic eQTL-ELSs showed a clear prevalence of tissue-specific terms for those genes targeted by intronic rather than intergenic eQTL-ELSs - for instance, muscle skeletal/cardiac:

carbohydrate and amino acid metabolism; brain: cell projection and microtubule cytoskeleton organization (Supplementary Table 6). In contrast, common eQTL-ELSs do not show any significantly enriched term neither in intronic nor in intergenic ELSs. Altogether, these results suggest that intronic eQTL-ELSs are involved in the regulation of genes associated with tissue-specific functions, while intergenic ELSs are more devoted to tissue homeostatic processes.

Target genes of intronic ELS identified by HiC regulate tissue-specific functions

The interaction between ELSs and promoters is central for the onset of gene expression. These kinds of interactions are defined in each tissue, and can be identified genome-wide through HiC-seq. Here, we explored the ELS-promoter interactions reported by published HiC datasets in relevant tissues, identifying tissue-specific ELS target genes, and thus improving the annotations of ELSs-target genes with respect to the eQTL analysis (Figure 2D,E) (Jung et al., 2019; Lu et al., 2020; Mifsud et al., 2015). As in the case of eQTL-ELSs, brain and muscle tissues show the highest proportion of intronic *vs* intergenic ELSs intersecting HiC interacting fragments detected in the corresponding tissue, while common HiC-ELSs are enriched in intergenic regions (Fig. 2E). The GO enrichment analysis reported an increase in relevant terms involved in tissue-specific functional roles as well. Notably, intronic HiC-ELS show better enrichment in tissue-specific terms (muscle-I band and Z disc components; brain-pre/postsynaptic assembly and organization; aorta-regulation of smooth muscle cell migration and proliferation), while we observed a broader functionality of intergenic ELSs' interactions (brain-choline catabolic process and copper ion homeostasis, amongst others) (Supplementary Table 7). Moreover, common HiC-ELSs appear to target genes that are enriched in housekeeping functions, like cell adhesion and nucleosome organization (Supplementary Table 7). Overall, these results on ELSs-promoter interactions further support that intronic

ELSs regulate genes controlling tissue-specific functions, while intergenic ELSs are more devoted to tissue homeostatic processes.

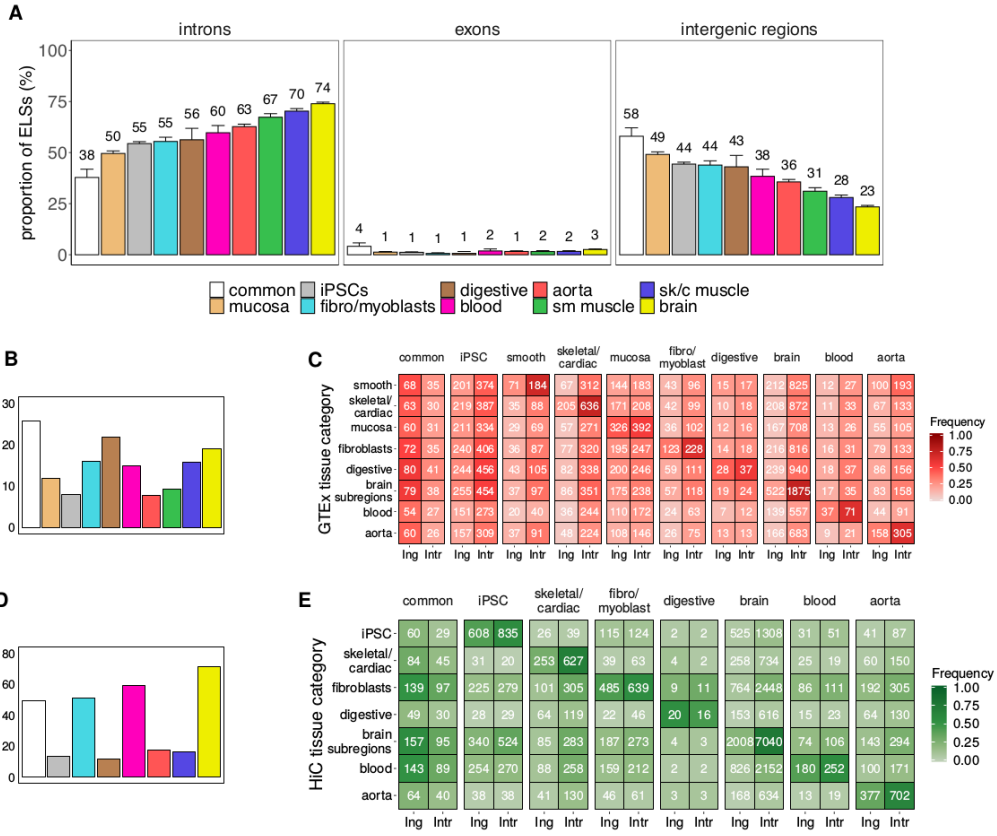


Fig. 2. **A**: Proportions of common and tissue-specific ELSs identified in the 33 selected human adult samples that overlap intronic, exonic and intergenic regions. Error bars represent the 95% confidence interval. **B**: Proportion of eQTL-ELSs with respect to the total amount of ELSs in each cluster. **C**: Number of intergenic (Ing) and intronic (Intr) cluster-specific ELSs harboring eQTLs detected in the analysed GTEx tissue samples except common and iPSC, which were annotated with a composition of tissue-specific significant eQTLs (see methods). Coloured cells represent the proportion of region-specific eQTL-ELSs over the total amount of eQTL-ELSs per cluster. Significant differences were reached between common and tissue-specific annotated eQTL-ELSs (Chi square test $p \leq 0.05$), showing that common annotated ELSs are highly associated with the intergenic part. **D**: Proportion of HiC-ELSs with respect to the total amount of ELSs in each cluster. **E**: Number of intergenic (Ing) and intronic (Intr) cluster-specific ELSs overlapping HiC-based detected fragments in the analysed HiC tissue samples except common, which was annotated with a composition of tissue-specific significant HiC fragments (see methods). Coloured cells represent the proportion of HiC-ELSs over the total amount of tissue-specific HiC-ELSs per cluster.

Significant differences were reached between common and non-common annotated HiC-ELS (Chi square test $p \leq 0.05$).

Intronic ELS regulate the expression of hosting and non-hosting genes

Next, we wanted to understand the relationship between the tissue-specific intronic ELSs and their harboring genes. To do so we focused on the gene expression pattern led by HiC-ELS interactions with the target gene, as a proxy for direct regulation. Of note, the proportion of intronic HiC-ELSs targeting their host genes was comparable among most groups of samples, between 45 and 65%, with the exception of muscle and blood that showed lower levels (Fig. 3A). We compared the gene expression patterns of the HiC-ELSs target genes depending on the type of ELS regulating them: Intergenic, Intronic Host, Intronic Non-Host across all the examined tissues (Fig. 3B). Hierarchical clustering of the genes regulated by each of these three categories indicated that genes regulated by their hosted intronic ELSs are the most efficient category to define tissue-specific expression patterns, while the non-host ELSs are the least efficient (Fig. 3B and Supplementary Fig. 2). Added to that, the tissue clustering always distinguishes the gene expression of the relevant tissue from the other tissue clusters, supporting the importance of the target/host HiC-ELSs interactions in tissue-specific gene expression. Most interestingly, HiC-ELSs regulating the expression of the host gene are associated with tissue-specific functions (Supplementary Table 8), with genes involved in synaptic vesicle clustering and active zone organization for the brain (e.g. PCDH17), regulation of cell division and establishment of cell polarity for fibroblasts (e.g. TGFB2), cardiac myofibril assembly and muscle fiber development terms for muscle skeletal/cardiac (e.g. MEF2A) or regulation of smooth muscle cell migration for aorta (e.g. DOCK5). However, those targeting the expression of non-hosting genes are involved in tissue homeostatic functions not uniquely associated with that tissue, indicating that those genes are not expressed in a tissue-specific manner, although

they are regulated by tissue-specific enhancers. For instance, the brain presents significant terms related to protein monoubiquitination (e.g. PDCD6) and the aorta non-host target genes are involved in cellular response to endogenous stimulus (e.g. TNC) (Fig. 3C). Overall, this suggests that the intronic location of regulatory elements cannot be assumed as the regulation of the host gene. Moreover, the identification of a relevant proportion of non-host intronic ELSs suggests an advantage of the intronic location for the regulation of tissue gene-expression, including non-tissue-specific gene expression.

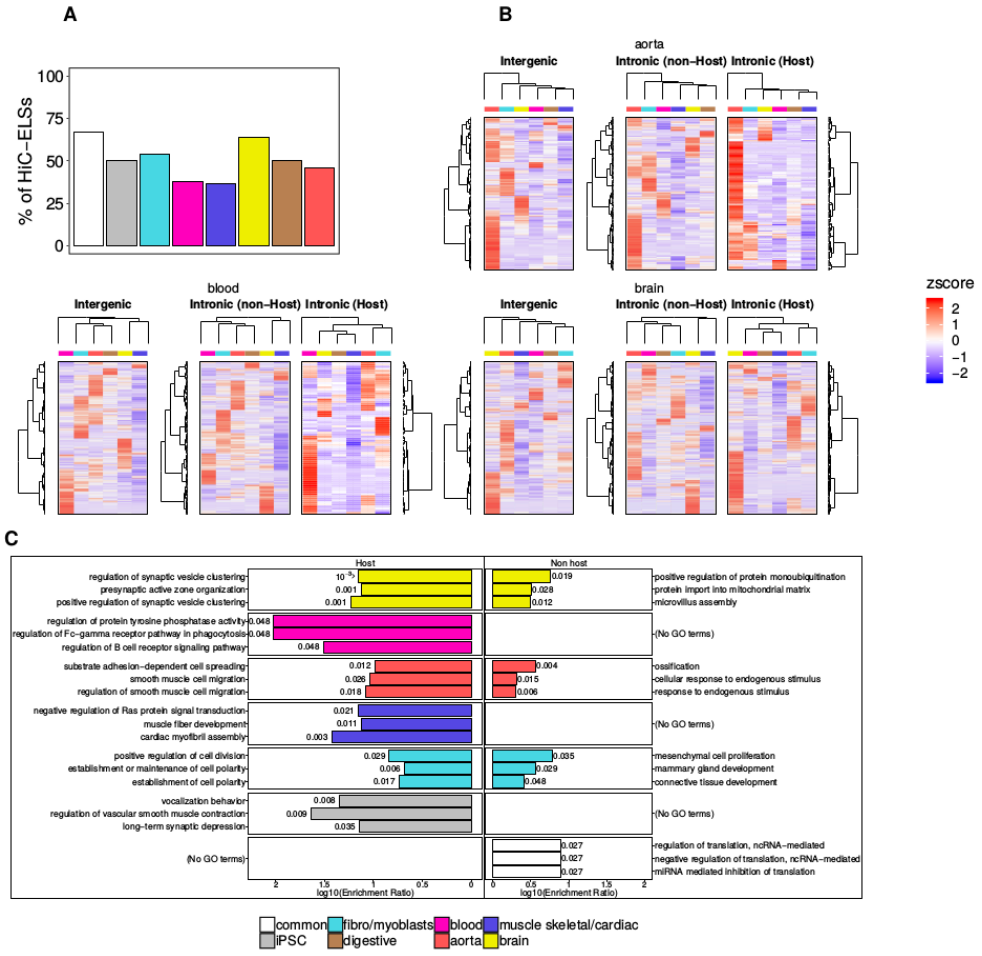


Fig. 3. **A**: Proportions of HiC-ELSs that target their host gene. These proportions were calculated over the total amount of intronic HiC-ELSs per cluster. **B**:

Z-score normalized median gene expression across GTEx tissue categories of the HiC-ELSs target genes in the intergenic and intronic locations. Intronic HiC-ELSs are divided into those that target their host gene (Host) and those that target a gene different from their host (non-Host). Dendrograms show the hierarchical clustering of target genes (rows) and GTEx tissue categories (columns). **C**: Top three significantly enriched GO terms found in the host and non-Host HiC-ELSs targeted genes. P-value (FDR corrected) is shown in each enriched term.

The enrichment of transcription factor binding sites in tissue-specific ELSs is independent of their genomic location

The activation of ELSs is a dynamic process depending, amongst other factors, on its accessible chromatin to be bound by transcription factors (TFs). Thus, tissue-specific gene expression programs may be controlled by the underlying signature of TFs-ELSs pairing (Schmitt et al., 2016). We next wondered whether the specific distribution of ELSs, i.e. intronic *vs* intergenic, was associated with a different transcription factor binding site (TFBS) signature that could account for their tissue-specific activity. To this purpose we explored the enrichment of TFBSs with HOMER (Heinz et al., 2010) for intronic and intergenic ELSs independently across tissues. Indeed, a distinct TFBS signature for each tissue in both intronic and intergenic ELSs can be observed, supporting our previous results that tissue-specific ELSs contribute significantly to the regulation of tissue-specific functions. Notably, when delving into each tissue's TFBS signature, the intronic and intergenic tissue-specific ELSs seem to have a different pattern depending on the tissue. The number of enriched TFBS in intronic regions is higher in highly specialized tissues such as the brain and the muscle, and show no overlap between the intronic and intergenic ELSs. The opposite picture is observed in common ELSs, where the higher enrichment is observed in intergenic ELSs and there is no overlap between the intronic and intergenic ELSs. An intermediate pattern is observed for the highly proliferative tissues like iPSC, fibroblasts, mucosa and blood, in which there is a higher shareness of TFBS and the amount of enriched TFBS is similar between

intronic and intergenic ELSs (Fig. 4A and Supplementary Table 9). Amongst the TFBSs enriched in the tissue-specific intronic and intergenic ELSs we find well-known TFs controlling tissue-specific homeostatic events, such as FLI1 and RUNX in blood controlling adult endothelial hemogenesis (Lis et al., 2017), and POU6F1 (Brn5), SOX4 and SOX8 in brain controlling the adult neural plasticity (McClard et al., 2018) POU5F1 (Oct4) is required for iPSCs. Still, with the exception of the TFs binding the enriched TFBS in iPSC, most TFs are widely expressed across tissues (Fig. 4B). This distinct iPSC TF-ELS binding potential is supported by previous data indicating that iPSC shares the epigenetic signature with early developmental stages than with the original tissue prior to reprogramming. Overall, the TFBS enrichment is different between intronic and intergenic ELSs and amongst different tissues but not the TFs gene expression pattern.

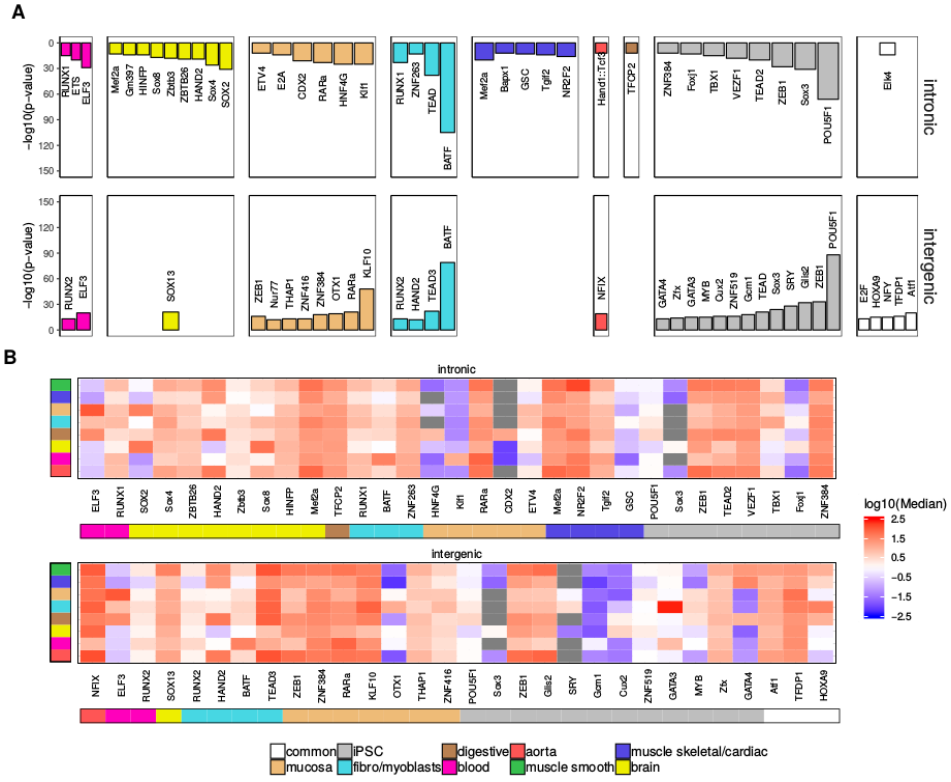


Fig. 4. **A**: Barplots reporting the significantly enriched TFBSs in intronic and intergenic tissue-specific ELSs. **B**: Z-score normalized median gene expression across GTEx tissue categories of the TFs that bind the significantly enriched TFBSs found in each cluster.

Dynamic location of ELSs throughout embryonic development and maturation

Throughout embryonic development, tissues mature to fully reach their functional capacity in adulthood, giving rise to several tissue-specific homeostatic features that dramatically vary among different tissues. For instance, blood comprises a wide number of cell types characterized by heterogeneous functions and high turnover. On the opposite side, we found highly specialized tissues such as the muscle, that are formed by fewer cell types, mainly dedicated to the same function and with limited cell division capacity. During development, tissues share features involving basic tissue homeostasis, proliferation and plasticity but also they are already patterned to their adult functions. Still, whether the regulatory features of a given adult tissue are reminiscent of their developmental lineage and the features of the embryonic ELSs remains largely unknown. To answer this question, we assessed the activity of the 991,173 cell type-agnostic ELSs across 27 embryonic samples (Supplementary Table 10). MDS analysis highlighted three main groups of embryonic samples: stem cells (ESC), neural progenitors, and a larger group of more differentiated cell types (Fig. 5B; Supplementary Table 10, Samples' Group). The three groups of samples are associated with 3,112, 784 and 1,166 specific ELSs, respectively (Supplementary Table 11). Although the majority of these ELSs are active only within the corresponding cluster, we reported that 26% of the neural progenitors-specific ELSs are also active in one ESC sample (Supplementary Fig. 3A). On the contrary, we identified only 94 ELSs common to all embryonic samples (Supplementary Table 11). The proportion of specific intronic ELSs is higher for neural progenitors and differentiated tissues compared to ESC-specific and common ELSs

(Fig. 5C), but lower with respect to clusters such as adult aorta and brain samples (Fig. 2A). As in the case of adult samples, we observed a scarcity of exonic ELSs (Fig. 5C, Supplementary Table 12), while we could not find significant associations between the frequency of group-specific intronic ELSs and features of gene and intron length (Supplementary Fig. 3B). Next, we wanted to validate the dynamics of intronic *vs* intergenic active ELSs throughout development, using brain development as a paradigm (Supplementary Fig. 4A). To this purpose, we identified active ELSs (H3K27ac+/H3K4me3-) in human ESCs and hESC-derived NPCs and neurons, and assessed their degree of overlap with active ELSs detected in ENCODE ESCs, NPCs and adult brain samples. Active ELSs identified by ChIP-seq in ESCs, NPC and neurons overlap with tissue-specific ENCODE ELSs for hESC (86%), NPC (40%) and brain (53%) samples, respectively. Notably, the proportion of active intronic ELSs increases with the degree of differentiation of the samples (55% in ESC, 64% in NPC and 68% in neurons) (Fig. 5D), validating the observed correlation between active tissue-specific ELSs and their intronic location. For common embryonic ELSs we find a high overlap (86% to 98%) with the hESC-differentiation ChIP-seq, which includes known ELSs for housekeeping genes, such as Actin-B (Supplementary Fig. 4B). Expression of genes regulated by a single ELS correlates with the activity of the ELS, being active in a tissue-specific manner in ESCs or common to all ENCODE samples (Supplementary Figs. 4C-E). Although H3K4me3 can be detected overlapping with the H3K27ac in the hESC, NPC and neurons ChIP-seq samples, the corresponding levels of H3K4me3 are much lower compared to those observed at promoter regions (Supplementary Fig. 4F). When analyzing the genes harboring developmental group-specific intronic ELSs, we observed that they are enriched in functions consistent with the corresponding adult tissue (Supplementary Table 13). For instance, the ones hosting neural progenitors-specific ELSs are enriched in neural development-related terms, such as axonogenesis and dendritic spine organization. Notably, genes

harboring developmental common ELSs are enriched in protein complexes like nBAF and SWI/SNF, known developmental chromatin remodelers (Alver et al., 2017).

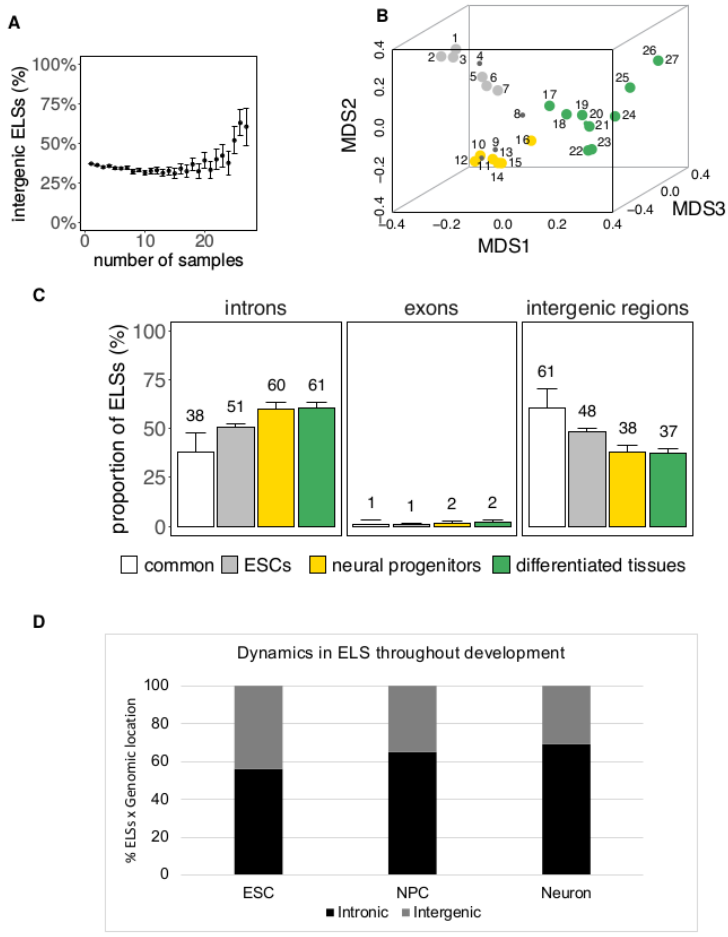


Fig. 5. Dynamic localisation of ELS through embryonic development. **A**. Correlation between shareness amongst embryonic samples and intergenic location of ELS (Spearman's correlation 0.36. Error bars represent the 95% confidence interval). **B**. MDS of embryonic samples allow 3 groups according to their ELS signature (neural tissue, stem cells and differentiated embryonic tissues). **C**. The more differentiated embryonic tissue-specific ELSs have a higher proportion of intronic ELSs while the common ELSs are preferentially intergenic. **D.** Dynamics of the localisation of active ELSs during ESC-derived maturation stages (hESC, neural progenitors (NPCs) and neurons), showing that ELSs increasingly distribute in intronic regions as maturation advances.

Lastly, in an attempt to define the amount of regulatory activity shared by embryonic and adult samples as an indicator of the reminiscent embryonic function in adult tissue homeostasis, we computed, for specific and common embryonic ELSs, the number of adult tissues in which they are found active. As expected, whereas ELSs specific to ESCs and neural progenitors are active in a limited set of adult samples, embryonic differentiated tissues report a higher degree of shared regulatory activity with adult cell types. Moreover, ELSs active in all embryonic samples (common) are also active in the majority of adult samples (Supplementary Fig. 5). Overall, these results show that the genomic location of ELSs is dynamic throughout development, and shifts towards an intronic localization during tissue maturation.

Discussion

In this study, we show the central role of intronic Enhancer-Like Signatures (ELSs) in the control of tissue-specific expression signatures. Since Heitz described in 1928 (Heitz, 1928) euchromatin as transcription permissive chromosomal regions enriched in genes, and heterochromatin as inactive or passive chromatin regions, this dual definition has been shaped throughout the years but it still remains vastly correct (De Laat and Duboule, 2013; DeMare et al., 2013; Ernst and Kellis, 2010). Intergenic regions are often regulatorily silenced, and this happens more frequently in adult than embryonic tissues (Heinz et al., 2015). The ENCODE project reports that about half of the ELSs are intergenic, and 38% are intronic (ENCODE SCREEN Portal: https://screen-v10.wenglab.org/, section “About”). In our study, we report an enrichment in intronic ELSs in the most specialized tissues, which regulate genes involved in tissue-specific functions, suggesting an important role of the genomic location of ELSs. Opposite, in less specialized adult tissues and embryonic samples, ELSs are not as frequently found in intronic elements as in highly specialized tissues, suggesting that the maturation and tissue

commitment correlates with the ELS distribution across the whole genome. One could hypothesize that the enriched presence of intronic ELSs in specialized tissues is advantageous for the control of the gene expression signature of a particular tissue, for instance granting ELSs accessibility in open DNA regions (genes) and avoiding leaky activity of ELSs. Recently, active transcription and nascent RNA have been associated with the maintenance of open chromatin (Hilbert et al., 2021), a process that can be advantageous to the presence of intronic ELSs in actively transcribed genes. Introns have been long observed as gene expression regulators throughout different mechanisms (Rose, 2019; Chorev and Carmel, 2012; Shaul, 2017). Introns regulatory potential has been longly associated with the regulation of the host gene's expression in several different ways, often related to alternative splicing, intron retention (Jacob and Smith, 2017), non-sense mediated decay (Lewis et al., 2003), and even with the control of transcription initiation via recruitment of RNA Polymerase II (Bieberstein et al., 2012). However, here we found that, in most tissues, about half of the ELSs located in introns do not regulate the expression of the host gene, but of genes involved in important tissue homeostasis functions, but whose expression is not restricted to that particular tissue. This is important regulatory information since it disentangles the presence of intronic ELSs from the regulation of the host gene, opening new opportunities to identify the regulatory mechanisms controlling tissue-specific gene expression. Overall, our results suggest that the genomic distribution of tissue-specific active ELSs is not stochastic and mainly overlaps with intronic elements. The opposite happens to active ELSs common to all tissues. These results suggest that intronic enhancers play a role in the regulation of gene expression in a tissue-specific manner.

Methods

The ENCODE registry of candidate *cis*-Regulatory Elements

The cell type-agnostic registry of human candidate *cis*-Regulatory Elements (cCREs) available from the ENCODE portal corresponds to a subset of 1,310,152 representative DNase hypersensitivity sites (rDHSs) in the human genome with epigenetic activity further supported by histone modification (H3K4me3 and H3K27ac) or CTCF-binding data (https://screen-v10.wenglab.org; section "About"). It comprises 991,173 Enhancer-Like Signatures (ELS), 254,880 Promoter-Like Signatures (PLS), and 64,099 CTCF-only Signatures. In addition, cell type-specific catalogues are provided for those cell types with available DNase and ChIP-seq ENCODE data.

Selection of cCREs with enhancer-like signature (ELS) across human samples

We downloaded the set of 1,310,152 cell type-agnostic cCREs for human assembly 19 (hg19) from the ENCODE SCREEN webpage (https://screen-v10.wenglab.org; file ID: ENCFF788SJC). From the ENCODE portal (www.encodeproject.org/matrix/?type=Annotation&encyclopedia_version=ENCODE+v4&annotation_type=candidate+Cis-Regulatory+Elements&assembly=hg19), we retrieved cell type-specific registries of cCREs for 43 adult and 27 embryonic human samples with available DNase data and ChIP-seq H3K4me3 and H3K27ac data. The ENCODE File Identifiers for the adult and embryonic datasets are reported in Supplementary Table 1 and Supplementary Table 8, respectively. No significant changes are expected upon realignment to GRCh38, since main improvements with respect to hg19 have been made in the representation of so-called alternate haplotypes, with a small impact on the definition of genic and intergenic regions (Church et al., 2015). We focused on the 991,173 cell type-agnostic cCREs with ELS activity, and generated a binary table in which we assessed, for a given cCRE, the presence/absence of ELS activity annotation (column 9 = "255, 205, 0") in each of the 43 adult and 27 embryonic samples. A binary distance matrix

between all pairs of adult samples was used to perform multidimensional scaling (MDS) in three dimensions. This resulted in the selection of 33 adult samples, which form 9 tissue groups well supported by hierarchical clustering (Figs. 1B-C) The same procedure was applied, independently, to the embryonic samples. In this case, IMR-90, mesendoderm, mesodermal cell, endodermal cell and ectodermal cell samples were not included in subsequent analyses.

Intersection of ELSs with genes, introns, exons and intergenic regions

Genes, exons and introns' coordinates were obtained from GENCODE v19 annotation (https://www.gencodegenes.org/human/release_19.html). The overlap between ELSs and genes, exons and introns was computed using BEDTools intersectBed v2.27.1 (Quinlan and Hall, 2010). The proportions of ELSs overlapping intronic segments (Figs. 2A, 5C) also include a limited set of ELSs overlapping both intronic and exonic regions. On the other hand, we defined as exonic ELSs those intersecting exclusively exonic regions (Figs. 2A, 5C). The overlap of ELSs with intergenic regions was obtained by intersecting the former with the genes' coordinates using the BEDTools intersectBed option *-v*.

Tissue-specific and common ELSs

Tissue-specific ELSs are ELSs active (see Methods section *Selection of cCREs with enhancer-like signature (ELS) across human samples*) in ≥ 80% of the samples within a given group of samples (blood = 4/5; skeletal/cardiac muscle = 3/4; smooth muscle = 3/4; brain = 6/7; stem cells = 5/6; neural progenitors = 5/6; differentiated tissues = 8/10). Because of the small sample size, we required iPSCs-, fibro/myoblasts-, digestive-, muscosa- and aorta-specific ELSs to be active in 100% of the samples (either 2/2

or 3/3). In addition, tissue-specific ELSs are active in 0 (iPSCs, fibro/myoblasts, digestive, mucosa and aorta) or at most 1 (all other groups) outer samples (i.e. samples outside of the considered group). Common adult and embryonic ELSs are ELSs active in 95% and 100% of the samples, respectively (i.e. 31/33 and 22/22).
To rule out indirect effects of ELS activity related to promoter regions, we discarded common and tissue-specific ELSs overlapping any annotated Transcription Start Site (TSS, ± 2Kb) in GENCODE v19.

Assessing enhancer regulatory activity with GTEx eQTL-eGene significant pairs

ELSs were annotated by using the GTEx v7 (Aguet et al., 2017) significant variant-gene pairs from 46 different tissues (number of samples with genotype ≥ 70), available on the GTEx portal (www.gtexportal.org) Only single-tissue eQTL-eGene associations with a qval ≤ 0.05 were used. Similar GTEx tissues were grouped in unique categories in order to consider the most complete catalogue of eQTL-eGene pairs per group of samples. These categories were named as follows: fibroblasts (Skin Not Sun Exposed Suprapubic, Cells Transformed Fibroblasts), blood (Whole Blood, Spleen), muscle skeletal/cardiac (Skeletal Muscle, Heart Atrial Appendage, Heart Left Ventricle), brain subregions (all brain subregions, Pituitary Gland, Nerve Tibial), Aorta (Artery Aorta), muscle smooth (Artery Coronary, Artery Tibial), digestive (Liver, Pancreas, Small Intestine Terminal Ileum, Stomach, Colon Sigmoid, Colon Transverse, Esophagus Gastroesophageal Junction, Esophagus Muscularis, Adipose Subcutaneous, Adipose Visceral Omentum), mucosa (Esophagus Mucosa), gland (Adrenal Gland, Thyroid, Minor Salivary Gland), breast (Breast Mammary Tissue), lung (Lung), sexual (Ovary, Prostate, Testis, Uterus, Vagina). BEDtools (Quinlan and Hall, 2010) was used to intersect the tissue-specific ELSs' coordinates with the *cis*-eQTLs' positions in the considered genomic locations (intronic and intergenic). We kept all

eQTL-eGene pairs that were found significantly associated with the matching eQTL-ELS's tissue category (muscle skeletal/cardiac, muscle smooth, fibro/myoblast, digestive, mucosa, brain, blood, aorta). In the case of iPSCs-specific and common ELSs, we considered those eQTL-eGene pairs that were significantly reported in at least 50% of all the tissues. The resulting intersected ELSs were considered as being responsible for the regulation of the associated eGene. The functional enrichment of the ELSs' target genes was performed by the online utility WebGestalt (Liao et al., 2019).

Assessing enhancer regulatory activity with HiC-based significant fragment pairs from loop contacts

ELSs were also annotated by using significant HiC-based interacting fragment pairs from three independent datasets (Jung et al., 2019; Lu et al., 2020; Mifsud et al., 2015). Different primary tissue and cell line samples were used to annotate each of the tissue-specific ELS categories in our study, except muscle smooth for which no HiC samples were found. As for the GTEx samples groups in the previous section, we grouped the HiC samples in unique categories in order to consider the most complete catalogue of HiC fragment pairs per group of samples. These categories were named as follows: Muscle skeletal/cardiac (Right ventricle (RV), Right heart atrium (RA3), Psoas (PO3), left ventricle (LV)), Fibro/myoblasts (Fibroblast cells (IMR90)), Brain (Hippocampus, dorsolateral prefrontal cortex, cortext adult, Neuron), Blood (GM12878+GM19240 lymphoblastoid cell line, CD34, GM12878), iPSC (iPSC), Aorta (Aorta), Mucosa (Sigmoid Colon), Digestive (Pancreas, Gastric tissue). In order to identify the significant ELS-gene pairs BEDtools (Quinlan and Hall, 2010) was used to intersect the HiC fragment coordinates with our ELSs in the different genomic locations (intronic and intergenic) and, in those cases in which the other fragment did not belong to any other ELS, we intersected them with the GENCODE annotation (v19),

inferring in this way the target genes of these ELSs. As for the eQTL annotation, only the HiC-based ELS-gene interactions associated with the matching HiC-ELSs' tissue category were kept (iPSC, muscle skeletal/cardiac, fibro/myoblast, digestive, brain, blood, aorta), Mucosa and Muscle smooth tissue-specific ELSs were removed from the analysis due to the lack of intersection with significant fragment pairs and HiC sample tissues, respectively. In the case of common ELSs we considered the ELS-gene pairs reported in at least 50% of all the HiC tissue samples. After the annotation of our ELSs we ended up with a collection of enhancer-gene interactions where the target gene was considered as being regulated by the interacting ELS. In order to define the Host/Non-Host ELSs in Fig. 4A we identified the ELSs' target genes that are also the host gene of that ELS. If a particular ELS presents among their target genes also its own host gene, that ELS is classified as Host, if none of the target genes is hosting the ELS, that element is classified as Non-Host. When considering the interactions ELS-gene in Fig. 4B, we defined an interaction as Host if the target gene is hosting that ELS, otherwise if the same ELS is targeting a gene that is not hosting the element, that interaction is classified as Non-Host. The target gene expression values were obtained from the GTEx expression data (v7) and Z-score normalized across the different GTEx tissue categories. The hierarchical clustering analysis of the Host/Non-Host target genes and GTEx tissue categories were performed with the R function *hclust*. The functional enrichment analysis on the ELSs' target genes and Host/Non-Host target genes were performed by the online utility WebGestalt (Liao et al., 2019).

cis-Regulatory Elements and Transcription Factor Binding Sites

Transcription factor binding sites (TFBSs) were predicted by using the motif discovery software HOMER (Heinz et al., 2010) This program performs a differential motif discovery by taking two sets of genomic regions (*findMotifGenome.pl* script) and identifying the

motifs that are enriched in one set of sequences relative to a background list of regions. We analysed the tissue-specific ELSs' binding motifs by considering the ELS regions from all the other tissues as background. We searched for 6-mer and 7-mer length motifs as a way to focus on enriched core motif sequences and avoid redundancy from longer motifs with similar functions. A hypergeometric test and FDR correction were applied for the motif enrichment. Only significantly enriched motifs were considered in the subsequent analysis. The functionality of the predicted TFBSs was assessed by analysing the tissue-specific expression of the transcription factors that bind to them. GTEx expression data (v7) was analysed for those transcription factors whose TFBSs were reported as significant by HOMER in all tissues and genomic locations. In the expression analysis, some transcription factors were removed due to the lack of expression data. Z-score normalization was performed across the different GTEx tissue categories in all transcription factors.

ChIP-seq data generation and processing

ChIP-seq data generation and processing was performed in hESC line H9 (WiCell), hESC-derived neural progenitors (NPC) and neurons. hESC were maintained in culture in mTESR (Stem Cell Technologies) and NPC and neurons were obtained upon cerebral organoid differentiation (Lancaster et al 2013). Briefly, 9000 H9 hESC were seeded in a low attachment 96-well (Corning) with Rock Inhibitor in mTESR. After 6 days, organoids were induced in induced media for another 6-8 days until the neuroepithelium was detectable and subsequently transferred to the neural expansion in matrigel. Organoids were disaggregated at day 30 post-differentiation and maintained in (N2B27 media supplemented with EGF and FGF2). NPC were harvested after 2 passages. Neurons were terminally differentiated in maturation media (N2B27) for 3 more weeks. Cells were harvested with Cell Dissociation Solution (ESC) and kept at -80C. DNA was

crosslinked with formaldehyde for 10 minutes at room temperature. Fixation was stopped by incubating with PBS / 0.1 % Triton X100 / 0.125 M glycine for 5 minutes at room temperature and chromatin was fragmented in a Q-sonica sonicator (15 minutes constant sonication at 40% Amplitude). H3K27ac (Active Motif reference 39336) and H3K4me3 (Active Motif reference 39916) antibodies were used for immunoprecipitation following the protocol previously described (Perez-Lluch et al., 2015). ChIP libraries were performed following Illumina procedures. Libraries were quantified by Qubit (Thermo Fisher) and visualized in a Fragment Analyzer (Agilent) previous to sequencing. Sequencing was performed in an Illumina NextSeq500, single-end run, following the instructions of the manufacturer. Data was processed using the *ChIP-nf* (https://github.com/guigolab/chip-nf) Nextflow (DI Tommaso et al., 2017) pipeline. Input samples were down-sampled to a number of reads comparable to the ChIP samples with the tool seqtk (https://github.com/lh3/seqtk). ChIP-seq reads were aligned to the human genome assembly (GRCh37) using the GEM (Marco-Sola et al., 2012) mapping software, allowing up to two mismatches. Only alignments for reads mapping to ten or fewer loci were reported. Duplicated reads were removed using Picard (http://broadinstitute.github.io/picard/). Peak calling was performed using Zerone (Cusco and Filion, 2016) with replicates handled internally. Pile-up signal from bigWig files was obtained running MACS2 (Zhang et al., 2008) on individual replicates. No shifting model was built. Instead, fragment length was defined for each experiment and used to extend each read towards the 3' end (using the --extsize option). Pile-up signal was normalized by scaling larger samples to smaller samples (using the default for the --scale-to option) and adjusting signal per million reads (enabling the --SPMR option).

Gene expression analysis

To validate gene expression regulation, target genes regulated by intronic or intergenic ELS were selected upon the following criteria: i) controlled by a single ELS active in brain (for tissue-specific) or in common in the ENCODE analysis, ii) it shows $H3K27ac^{+}/H3k4me3^{-}$ peaks in relevant cell ChIP-seq validation, iii) do not overlap with exons (Supplementary Table 14). RNA was obtained from hESC, NPC and neuron pellets used for ChIP-seq. Retrotranscription was performed using Superscript III retrotranscriptase. qPCR was performed in 10 ng cDNA with the Roche Sybr Green Master Mix. Primers used for qPCR are reported in Supplementary Table 14). Gene expression is reported following the relative expression of the DDCt method. GAPDH and ACTB were used as reference genes. ACTB gene expression showed more stability throughout the differentiation and therefore, it was used as the reference gene for the analysis.

Data access

Newly generated ChIP-seq data are in the process of being submitted to ArrayExpress (https://www.ebi.ac.uk/arrayexpress/).

Acknowledgments

S.A. is a Serra-Hunter Fellow since 2021. S.A. is supported by a fellowship from the Secretaria d'Universitats i Recerca del Departament d'Empresa i Coneixement (Generalitat de Catalunya) (BP-2017-00176). B.B. is supported by the fellowship 2017FI_B00722 from the Secretaria d'Universitats i Recerca del Departament d'Empresa i Coneixement (Generalitat de Catalunya) and the European Social Fund (ESF). P.V-M. is supported by an FPI PhD fellowship (FPI-BES-2016-077706) part of the "Unidad de Excelencia María de Maeztu funded by the MINECO (ref: MDM-2014-0370). We thank the ENCODE and GTEx Consortia for data production. We thank Diego Garrido-Martín (R. Guigó Lab) for valuable statistical advice.

Author Contributions: S.A., B.B, and P.V-M. designed the study, analyzed the data and wrote the manuscript with feedback from all the authors. S.A. and S.P-L performed ChIP-seq experiments, I.T. generated the differentiations for the ChIP-seq. H.L. and A.S-C. performed some of the bioinformatic analyses. J.B and R.G. contributed to the manuscript edition.

References

Abascal, F., Acosta, R., Addleman, N. J., Adrian, J., Afzal, V., Aken, B., Akiyama, J. A., Jammal, O. A., Amrhein, H., Anderson, S. M., et al. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature, 583(7818):699–710.

Aguet, F., Brown, A. A., Castel, S. E., Davis, J. R., He, Y., Jo, B., Mohammadi, P., Park, Y. S., Parsana, P., Segrè A. V., et al. (2017). Genetic effects on gene expression across human tissues. Nature, 454 550(7675):204–213.

Alver, B. H., Kim, K. H., Lu, P., Wang, X., Manchester, H. E., Wang, W., Haswell, J. R., Park, P. J., and Roberts, C. W. (2017). The SWI/SNF chromatin remodelling complex is required for maintenance of lineage specific enhancers. Nature Communications, 8.

Bieberstein, N. I., Oesterreich, F. C., Straube, K., and Neugebauer, K. M. (2012). First exon length controls active chromatin signatures and transcription. Cell Reports, 2(1):62–68.

Bonev, B., Mendelson Cohen, N., Szabo, Q., Fritsch, L., Papadopoulos, G. L., Lubling, Y., Xu, X., Lv, X., Hugnot, J. P., Tanay, A., et al. (2017). Multiscale 3D Genome Rewiring during Mouse Neural Development. Cell, 171(3):557–572.

Chen, C., Yu, W., Tober, J., Blobel, G. A., Speck, N. A., and Correspondence, K. T. (2019). Spatial Genome Re-organization between Fetal and Adult Hematopoietic Stem Cells. Cell Reports, 29(12):4200–4211.

Chorev, M. and Carmel, L. (2012). The function of introns. Frontiers in Genetics, 3.

Choukrallah, M. A., Song, S., Rolink, A. G., Burger, L., and Matthias, P. (2015). Enhancer repertoires are reshaped independently of early priming and heterochromatin dynamics during B cell differentiation. Nature Communications, 6.

Church, D. M., Schneider, V. A., Steinberg, K. M., Schatz, M. C., Quinlan, A. R., Chin, C. S., Kitts, P. A., Aken, B., Marth, G. T., Hoffman, M. M., et al. (2015). Extending reference assembly models. Genome Biology, 16(1):13.

Cuscó, P. and Filion, G. J. (2016). Zerone: A ChIP-seq discretizer for multiple replicates with built-in quality control. Bioinformatics, 32(19):2896–2902

De Laat, W. and Duboule, D. (2013). Topology of mammalian developmental enhancers and their regulatory landscapes. Nature, 502(7472):499–506.

DeMare, L. E., Leng, J., Cotney, J., Reilly, S. K., Yin, J., Sarro, R., and Noonan, J. P. (2013). The genomic landscape of cohesin-Associated chromatin interactions. Genome Research, 23(8):1224–1234.

DI Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., and Notredame, C. (2017). Nextflow enables reproducible computational workflows. Nature Biotechnology, 35(4):316–319.

Eisenberg, E. and Levanon, E. Y. (2013). Human housekeeping genes, revisited. Trends in Genetics, 29(10):569–574.

Ernst, J. and Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. Nature Biotechnology, 28(8):817–825.

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature, 473(7345):43–49.

Gilbert, N., Boyle, S., Sutherland, H., Heras, J. d. L., Allan, J., Jenuwein, T., and Bickmore, W. A. (2003). Formation of facultative heterochromatin in the absence of HP1. The EMBO Journal, 22(20):5540–5550.

Gillies, S. D., Morrison, S. L., Oi, V. T., and Tonegawa, S. (1983). A Tissue-specific Transcription Enhancer Element Is Located in the Major lntron of a Rearranged lmmunoglobulin Heavy Chain Gene. Cell, 33(3):717–728.

Hawkins, R. D., Hon, G. C., Lee, L. K., Ngo, Q., Lister, R., Pelizzola, M., Edsall, L. E., Kuan, S., Luu, Y., Klugman, S., et al. (2010). Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. Cell Stem Cell, 6(5):479–491.

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., Barrera, L. O., Van Calcar, S., Qu, C., Ching, K. A., et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nature Genetics, 39(3):311–318.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., and Glass, C. K. (2010). Simple

combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Molecular cell, 38(4):576–89.

Heinz, S., Romanoski, C. E., Benner, C., and Glass, C. K. (2015). The selection and function of cell type-specific enhancers. Nature Reviews Molecular Cell Biology, 16(3):144–154.

Heitz, E. (1928). Das Heterochromatin der Moose. Jahrbucher für wissenschaftliche Botanik, 69.

Hilbert, L., Sato, Y., Kuznetsova, K., Bianucci, T., Kimura, H., Jülicher, F., Honigmann, A., Zaburdaev, V., and Vastenhouw, N. L. (2021). Transcription organizes euchromatin via microphase separation. Nature Communications, 12(1):1360.

Jacob, A. G. and Smith, C. W. (2017). Intron retention as a component of regulated gene expression programs. Human Genetics, 136(9):1043–1057.

Jung, I., Schmitt, A., Diao, Y., Lee, A. J., Liu, T., Yang, D., Tan, C., Eom, J., Chan, M., Chee, S., et al. (2019). A compendium of promoter-centered long-range chromatin interactions in the human genome. Nature Genetics, 51(10):1442–1449.

Kawase, S., Imai, T., Miyauchi-Hara, C., Yaguchi, K., Nishimoto, Y., Fukami, S. I., Matsuzaki, Y., Miyawaki, A., Itohara, S., and Okano, H. (2011). Identification of a novel intronic enhancer responsible for the transcriptional regulation of musashi1 in neural stem/progenitor cells. Molecular Brain, 4(1).

Khandekar, M., Brandt, W., Zhou, Y., Dagenais, S., Glover, T. W., Suzuki, N., Shimizu, R., Yamamoto, M., Lim, K. C., and Engel, J. D. (2007). A Gata2 intronic enhancer confers its

pan-endothelia-specific regulation. Development, 134(9):1703–1712.

Levine, M. (2010). Transcriptional enhancers in animal development and evolution. Current Biology, 20(17):754–763.

Lewis, B. P., Green, R. E., and Brenner, S. E. (2003). Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proceedings of the National Academy of Sciences of the United States of America, 100(1):189–192.

Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z., and Zhang, B. (2019). WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. Nucleic Acids Research, 47(W1):W199–W205.

Lis, R., Karrasch, C. C., Poulos, M. G., Kunar, B., Redmond, D., Duran, J. G., Badwe, C. R., Schachterle, W., Ginsberg, M., Xiang, J., et al. (2017). Conversion of adult endothelium to immunocompetent haematopoietic stem cells. Nature, 545(7655):439–445.

Lu, L., Liu, X., Huang, W. K., Giusti-Rodr´ıguez, P., Cui, J., Zhang, S., Xu, W., Wen, Z., Ma, S., Rosen, J. D., et al. (2020). Robust Hi-C Maps of Enhancer-Promoter Interactions Reveal the Function of Non-coding Genome in Neural Development and Diseases. Molecular Cell, 79(3):521–534.

Marco-Sola, S., Sammeth, M., Guigo´, R., and Ribeca, P. (2012). The GEM mapper: Fast, accurate and versatile alignment by filtration. Nature Methods, 9(12):1185–1188.

McClard, C. K., Kochukov, M. Y., Herman, I., Liu, Z., Eblimit, A., Moayedi, Y., Ortiz-Guzman, J., Colchado, D., Pekarek, B., Panneerselvam, S., et al. (2018). POU6f1 mediates

neuropeptide-dependent plasticity in the adult brain. Journal of Neuroscience, 38(6):1443–1461.

Melé, M., Ferreira, P. G., Reverter, F., DeLuca, D. S., Monlong, J., Sammeth, M., Young, T. R., Goldmann, J. M., Pervouchine, D. D., Sullivan, T. J., et al. (2015). The human transcriptome across tissues and individuals. Science, 348(6235):660–665.

Mifsud, B., Tavares-Cadete, F., Young, A. N., Sugar, R., Schoenfelder, S., Ferreira, L., Wingett, S. W., Andrews, S., Grey, W., Ewels, P. A., et al. (2015). Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nature Genetics, 47(6):598–606.

Ott, C. J., Blackledge, N. P., Kerschner, J. L., Leir, S. H., Crawford, G. E., Cotton, C. U., and Harris, A. (2009). Intronic enhancers coordinate epithelial-specific looping of the active CFTR locus. Proceedings of the National Academy of Sciences of the United States of America, 106(47):19934–19939.

Pennacchio, L. A., Loots, G. G., Nobrega, M. A., and Ovcharenko, I. (2007). Predicting tissue-specific enhancers in the human genome. Genome Research, 17(2):201–211.

Pérez-Lluch, S., Blanco, E., Tilgner, H., Curado, J., Ruiz-Romero, M., Corominas, M., and Guigó, R. (2015). Absence of canonical marks of active chromatin in developmentally regulated genes. Nature Genetics, 47(10):1158–1167.

Pervouchine, D. D., Djebali, S., Breschi, A., Davis, C. A., Barja, P. P., Dobin, A., Tanzer, A., Lagarde, J., Zaleski, C., See, L. H., et al. (2015). Enhanced transcriptome maps from multiple mouse tissues reveal evolutionary constraint in gene expression. Nature Communications, 6.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 26(6):841–842.

Rand, E. and Cedar, H. (2003). Regulation of imprinting: A multi-tiered process. Journal of Cellular Biochemistry, 88(2):400–407.

Rose, A. B. (2019). Introns as gene regulators: A brick on the accelerator. Frontiers in Genetics, 9.

Schmitt, A. D., Hu, M., Jung, I., Xu, Z., Qiu, Y., Tan, C. L., Li, Y., Lin, S., Lin, Y., Barr, C. L., et al. (2016). A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. Cell Reports, 17(8):2042–2059.

Shaul, O. (2017). How introns enhance gene expression. International Journal of Biochemistry and Cell Biology, 91:145–155.

Shlyueva, D., Stampfel, G., and Stark, A. (2014). Transcriptional enhancers: From properties to genome-wide predictions. Nature Reviews Genetics, 15(4):272–286.

Zabidi, M. A., Arnold, C. D., Schernhuber, K., Pagani, M., Rath, M., Frank, O., and Stark, A. (2015). Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. Nature, 518(7540):556–559.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., et al. (2008). Model-based Analysis of ChIP-Seq (MACS). Genome Biology, 9(9):R137.

Signatures of genetic variation in human microRNAs point to processes of positive selection related to population-specific disease risks

Pablo Villegas-Mirón, Alicia Gallego, Jaume Bertranpetit, Hafid Laayouni and Yolanda Espinosa-Parrilla

Submitted for publication

Preprint citation reference:

Villegas-Mirón P, Gallego A, Bertranpetit J, Laayouni H, Espinosa-Parrilla Y. 2021. Signatures of genetic variation in human microRNAs point to processes of positive selection related to population-specific disease risks. bioRxiv doi: https://doi.org/10.1101/2021.05.24.445417

Signatures of genetic variation in human microRNAs point to processes of positive selection related to population-specific disease risks

Pablo Villegas-Mirón[1], Alicia Gallego[2], Jaume Bertranpetit[1], Hafid Laayouni#[1, 3] and Yolanda Espinosa-Parrilla# [4,5,6]

[1]Institut de Biologia Evolutiva (UPF-CSIC), Universitat Pompeu Fabra, Barcelona, Catalonia, Spain.
[2]Centro de Biología Molecular Severo Ochoa, CSIC-UAM, Madrid, Spain.
[3]Bioinformatics Studies, ESCI-UPF, Pg. Pujades 1, 08003 Barcelona, Spain

[4]Escuela de Medicina, Universidad de Magallanes, Punta Arenas, Chile

[5]Laboratorio de Medicina Molecular - LMM, Centro Asistencial Docente y de Investigación - CADI, Universidad de Magallanes, Punta Arenas, Chile

[6]Interuniversity Center on Healthy Aging, Chile

*Author for Correspondence: Hafid Laayouni (hafid.laayouni@upf.edu) and Yolanda Espinosa Parrilla (yolanda.espinosa@umag.cl)

# Equally contributors

Abstract

The occurrence of natural variation in human microRNAs has been the focus of numerous studies during the last twenty years. Most of them have been dedicated to study the role of specific mutations in diseases, like cancer, while a minor fraction seek to analyse the diversity profiles of microRNAs in the genomes of human populations. In the present study we analyse the latest human microRNA annotations in the light of the most updated catalog of genetic variation provided by the 1000 Genomes Project. We show by means of the *in silico* analysis of noncoding variation of microRNAs that the level of evolutionary constraint of these sequences is governed by the interplay of different factors, like their evolutionary age or the genomic location where they emerged. The role of mutations in the shaping of microRNA-driven regulatory interactions is emphasized with the acknowledgement that, while the whole microRNA sequence is highly conserved, the seed region shows a pattern of higher genetic diversity that appears to be caused by the dramatic frequency shifts of a fraction of human microRNAs. We highlight the participation of these microRNAs in population-specific processes by identifying that not only the seed, but also the loop, are particularly differentiated regions among human populations. The quantitative computational comparison of signatures of population differentiation showed that candidate microRNAs with the largest differences are enriched in variants implicated in gene expression levels (eQTLs), selective sweeps and pathological processes. We explore the implication of these evolutionary-driven microRNAs and their SNPs in human diseases, such as different types of cancer, and discuss their role in population-specific disease risk.

Introduction

MicroRNAs (miRNAs) are short (~22 nucleotides) single-stranded regulatory non-protein-coding RNAs that perform a post-transcriptional negative control of the expression of more than 60% of the whole human genome (Friedman et al. 2009). They are involved in the control of almost every cellular process, including development, differentiation, proliferation and apoptosis, and present important roles in diseases. They are transcribed by RNA polymerase II as primary sequences, which are later processed by the proteins Drosha and Dicer into a miRNA duplex formed by two mature miRNA strands, 5p and 3p (Ha et al. 2014). This mature molecule is then loaded onto an AGO protein forming the RNA-induced silencing complex (RISC), promoting the RNA silencing by translation repression or mRNA degradation. Target gene repression is accomplished by the partial sequence complementarity between the target mRNA and the miRNA. In this interaction, a perfect match between the miRNA seed region, expanded across nucleotides 2-8 of the 5' extreme, and the target site, usually located within the mRNA 3' untranslated region, is needed (Lewis et al. 2005; Grimson et al. 2007; Bartel et al. 2009; Berezikov 2011). Other positions of the mature sequence may interfere in the mRNA binding, like the 3' supplementary and compensatory sites, that enhance the seed-matched binding efficiency (Grimson et al. 2007; Friedman et al. 2009; Bartel 2018).

miRNAs have experienced multiple periods of fast turn over and lineage-specific expansions through their evolutionary trajectory (Lu et al. 2008; Iwama et al. 2012). Most of the current human miRNAs originated in two accelerated peaks of miRNA expansion that are reported during mammalian evolution: the first peak of new miRNAs was located at the initial phase of the placental radiation, while the second and highest peak was observed at the beginning of the simian lineage, that originated more than a half of the current repertoire (Iwama et al. 2012; Santpere et al. 2016). These miRNA

expansions were implicated in the acquisition of new regulatory tools that have been directly linked with animal complexity and evolutionary innovations across all lineages (Hertel et al. 2006; Heimberg et al. 2008; Wheeler et al. 2009).

miRNAs can be found either in intergenic regions or being hosted by other elements, like protein-coding and non-coding genes or repetitive elements like transposons. These are the genomic contexts where hairpin-like transcripts initially emerge and are gradually shaped by evolution until they become functional miRNAs (Berezikov 2011). Differences in the genomic environment and location of miRNAs are associated with different evolutionary properties. For example, in França et al. 2016 the authors show the association of the age of the host gene with the breadth expression and evolutionary trajectory of recently emerged hosted miRNAs. Duplication events are one of the main sources of new miRNAs. These can be found close to each other when the duplication is local, forming clusters that are found to be evolutionary related and functionally implicated in similar regulatory pathways (Wang et al. 2016). The origin of miRNAs and their target sites are tightly related to the dynamics of transposable elements (TE). These are sequences that jump, replicate and insert in other parts of the genome, generating mutations. However, apart from the damaging consequences of these changes, they can also incorporate new functional regions in other genomic environments (like miRNA target sites) and modify regulatory networks (Feschotte 2008; Chuong et al. 2017). According to some authors (Piriyapongsa et al. 2007; Qin et al. 2015; Petri et al. 2019), the expansion of new miRNAs in the primate lineage gave birth to a great number of TE-derived miRNAs, highlighting the importance of transposons as a source of genomic innovation.

The computational analysis of human genetic variation has traditionally been focused on protein-coding genes, being non-protein coding sequences neglected from this kind of studies.

However, in recent years, several reports have paid more attention to the consequences of naturally occurring variation in miRNAs (Cammaerts et al. 2015). A signature of purifying selection shapes the miRNA diversity worldwide, revealing that human miRNAs are highly conserved sequences that rarely accept changes in their sequences (Quach et al. 2009), indeed miRNA expression and functionality are usually tightly subjected to the presence of variants within (Quach et al. 2009) and outside (Borel et al. 2011) their hairpin. Sequence changes in the premature and loop regions might generate distorsions in their folding and affect the expression and maturation of the primary sequences (Fernandez et al. 2017). Moreover, the occurrence of changes in the mature region and the seed (Gong et al. 2012; Hill et al. 2014; Gallego et al. 2016; He et al. 2018), which outstands as the most conserved region of the hairpin, can dramatically affect the recognition of their target genes, which is also affected by the presence of variants in their target sites (Li et al. 2012). All these changes might induce massive rewirings of the miRNA regulatory networks and alter the downstream processes, inducing gene expression changes and phenotypic variation that might degenerate in pathogenic processes (Sethupathy and Collins 2008; Rawlings-Goss et al. 2014; Ghanbari et al. 2017; Grigelioniene et al. 2019), but also be the origin of genetic innovations responsible for phenotypic adaptations (Lu et al. 2012). Several authors have reported population-specific variants that affect different dimensions of the miRNA functionality (Saunders et al. 2007; Torruella-Loran et al. 2016) and their target sites (Li et al. 2012) and might be involved in adaptation processes. More recently, it has been reported a clear signal of adaptive evolution in a metabolic-related miRNA responsible for adaptations to past famine periods (Wang et al. 2020).

In this study we revisited the hosting and conservation patterns of the most complete human miRNA catalog to date. We also performed a comprehensive computational analysis of their diversity patterns worldwide, considering the factors that might

contribute the most to the configuration of this variation. We finally studied the population differences and putative positive selection signals of the variable miRNAs, and looked at the potential consequences of this variation in terms of human diseases and recent adaptation.

Results

The genomic context of miRNAs is associated with their evolutionary age

To study the recent evolutionary history of human miRNA genes a total of 1918 precursor miRNAs (miRBase v.22, March 2018) were considered, from which 1904 remained after *liftOver* conversion to the hg19 assembly (Supplementary Table S1). From these miRNA precursors, 50.3% presented a complete annotation of their mature sequences (5p and 3p), while the other half presented only a single mature sequence identified in one of their arms (Fig. 1a and Supplementary Fig. S1). First, we classified these 1904 miRNAs in groups of conservation, according to their evolutionary age, by adapting the categories from Iwama et al. (2013) and Santpere et al. (2016) (see Methods). In total, 1623 (85.2%) miRNAs were classified in four different conservation categories: Primates (985, 51.7%), Eutherians (421, 22.1%), Metatheria-Prototheria (63, 3.3%) and conserved beyond mammals (154, 8%). The remaining miRNAs (281, 14.8%) could not be classified due to the absence of data or discrepancies between studies and were excluded from the subsequent analyses (Supplementary Table S1).

Next, we classified miRNAs in different genomic contexts by identifying the different elements that overlap their precursor sequences. According to GENCODE 19 (v.29) we found that 483 (25%) miRNAs fell in intergenic regions (Intg), while 1421 were located within protein coding genes (PC) (1217, 63.9%) and long non-coding RNAs (LNC) (204, 10.7%), either presenting a single or

multiple overlapping host genes. In our dataset we found that 856 (60%) intragenic miRNAs (protein coding and lncRNA) overlapped introns of the host sequence, while 545 (38%) were located within exonic regions. The remaining 20 (~1%) showed a mixture of intronic/exonic locations (Supplementary Table S1). Further, we used the last release of the RepeatMasker database (Smit et al. 2013-2015) to identify the different forms of transposable elements (TEs) and repetitive sequences that host miRNAs. We found 660 (35%) miRNAs overlapping TEs alone or in combination with other genes, while the remaining 1244 (65%) were either unmasked or overlapping other forms of repetitive sequences and genes. Interestingly, we found a strong correlation between the frequencies of the TE-hosting miRNAs and their evolutionary age, being the primate-specific group the one with the highest presence of miRNAs in this context (440, 23.1%; Fig. 1b). Alu (67, 6.8%), L1 (54, 5.4%), TcMar (42, 4.2%) and the LTR elements ERV1 and ERVL (36, 3.6%) were found mainly among the primate-specific miRNAs, while hAT (3.3%) and L2 (28, 6.6%) elements were also present in the eutherian group (Supplementary Table S2). It is of interest to note that the contribution of MIR (15.3%) and DNA elements like TcMar (14.8%) and hAT (12.8%) families to the miRNA context is higher than to the whole genome (Supplementary Fig. S2a).

We found that the genomic context increased in complexity when different elements appeared hosting the same miRNA simultaneously. We studied the integrated hosting of miRNAs across the conservation groups considering the different combinations of elements (Fig. 1c, Supplementary Table S3). This shared hosting evidences the two main sources of miRNAs: protein coding genes (796; 41.8%) and TEs (196; 10.2%), with 401 miRNAs presenting a combination of both (21%). As expected, the genomic context is associated with the age of miRNAs (Chi square test = 238.25, p = 2.2e-16). This association shows that primate-specific miRNAs present a dominance of overlapping TEs

in comparison with non-primate miRNAs, with the TE and TE + PC hosting categories being the major contributors across environments. On the other hand, lncRNAs are highly associated with the miRNA context among the non-primate groups, mainly in the group of miRNAs conserved beyond mammals (Supplementary Fig. S2b).

We made use of the miRNA expression levels in 16 different human tissues extracted from Panwar et al. (2017) (see Methods) to study their correlation across groups of conservation. As seen in Fig. 1d, the tissue specificity is higher at lower evolutionary ages, which indicates the limited expression breadth of young miRNAs. Also, the expression levels were correlated with age, having the more conserved miRNAs an overall higher expression due to their consolidated role in regulatory networks (Fig. 1e).

Due to the evolutionary relevance of the miRNA organization in the genome, we revisited the clustering patterns of the miRBase annotations. When studying the closeness between miRNAs, an increment of distances ranging 1-10kb was found (Supplementary Fig. S2c), which indicates a high accumulation of close miRNAs in certain regions. According to this, we defined that two miRNAs belong to the same cluster when they are located 10kb or closer from each other. A total of 100 clusters were identified in the whole genome (Fig. 1f and Supplementary Fig. S3), represented by 352 miRNA members. Two thirds of these clusters (64) were constituted only by two genes, while 36 clusters presented more than two. Two main clustering hotspots were observed in the chromosomes 14 (42) and 19 (46), as previously reported by Guo et al. (2014), while the X chromosome presented a similar amount of clustered miRNAs (57) but more widespread in different smaller groups (Fig. 1f). A total of 1552 miRNAs were located in isolated regions. We also found a strong correlation between the clustering patterns of miRNAs and groups of conservation (Fig. 1g). The more conserved miRNAs tend to be found in clusters rather than in isolated regions,

something likely related to the conserved role of clustered miRNAs in similar biological processes (Berezikov 2011; Wang et al. 2016).

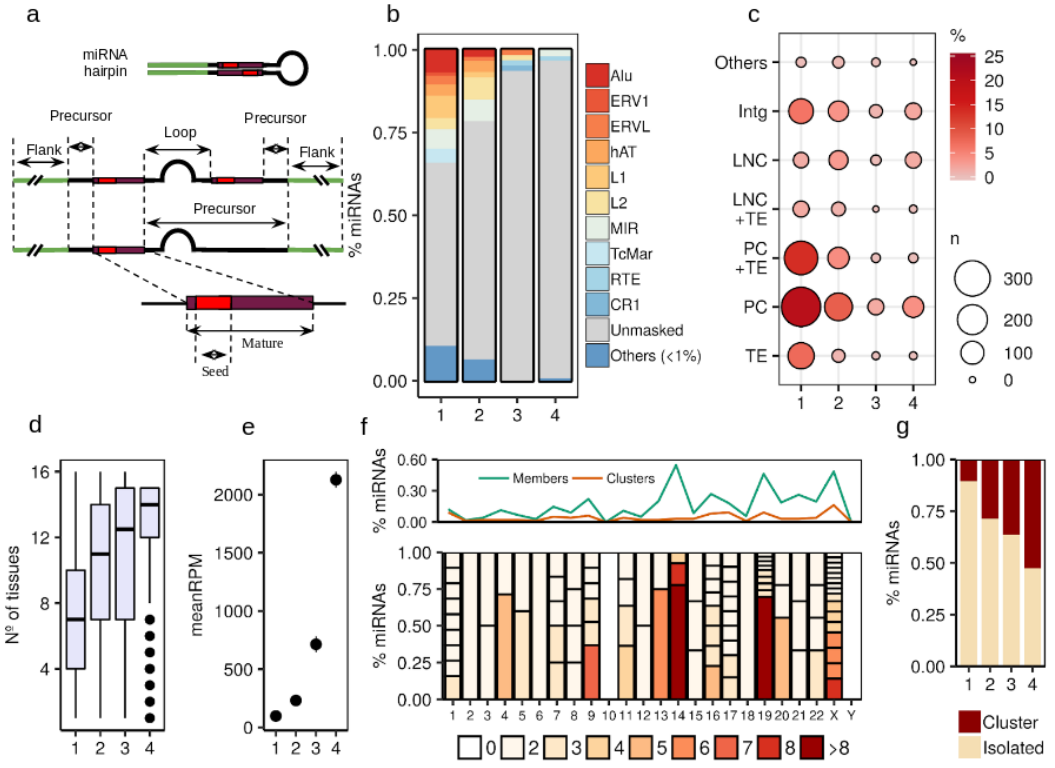


Fig. 1 Description of human miRNAs in terms of genomic context, evolutionary age groups (see Methods), expression levels and clustering. (a) Description of the miRNA hairpin regions identified and analysed in the study. Not all the primary sequences present two mature sequences annotated by miRbase. When the two mature sequences are not given (incomplete annotation), the precursor region is extended from the first mature to the other flanking region. (b) TE-derived miRNA frequencies across conservation groups (Primates, 1; Eutherians, 2; Metatheria and Prototheria, 3; Conserved beyond mammals, 4). (c) Integrated hosting of miRNAs that shows the combination of the different hosting elements that overlap with miRNA sequences. The "Other" group is made with the minor categories (PC+LNC and PC+LNC+TE) that represent less than 1% of the total dataset (Supplementary Table S2). (d) Number of tissues where the miRNA is expressed across evolutionary ages (e) Mean expression level (Reads per million mapped reads; RPM) of miRNAs across evolutionary ages (f) Whole genome clustering patterns of miRNAs. The upper plot represents the frequency of miRNAs that belong to a certain cluster in each chromosome (Members) and the

frequency of clusters in the whole genome (Clusters). The lower plot represents the miRNA clusters per chromosome, according to the number of members and their frequency among the clustered miRNAs. (g) Fraction of clustered and isolated miRNAs across evolutionary ages

Nucleotide diversity of miRNAs is strongly shaped by their age, genomic context and localization

The genetic variation of the miRNA dataset was analysed in the different miRNA functional regions using human genetic variation from the 1000 Genomes project (Fig. 1a; Auton A et al. 2015). A total of 569 single nucleotide polymorphisms (SNPs) were located in 466 miRNA precursors (26.1%), but when considering a region of the same size at both sides of the precursor sequence (5' and 3' flanking regions) the number of SNPs increased to 1994 in 1026 miRNAs (55.9%). Therefore, more than half of the variability found in our miRNAs comes from the neutral-like flanking regions. The mature sequence is considered the most conserved and important functional region of the miRNA, since it regulates the target gene by binding to the 3'UTR mainly through the seed region. In our dataset, 212 SNPs were present in 194 mature sequences (7.5%), while 79 SNPs were present only in the seed region of 75 miRNAs (2.9%).

To study the sequence variation of human miRNAs we analysed the nucleotide diversity of 1904 miRNA precursor sequences described in miRBase in the pooled population sample from the 1000 Genomes project. The genomic context refers to the environment where miRNAs originally emerged, which might be determinant to their level of variation. We calculated the global nucleotide diversity (Pi) in the whole precursor sequence by considering the age, location and clustering of the miRNAs (Fig. 2). We found significant differences when comparing the Pi of miRNAs in the different contexts (Kruskal-Wallis $p = 0.013$). Fig. 2a shows that miRNAs harboured by TEs exhibit a significantly higher Pi than in other genomic contexts. Next, we examined the TE-family specific

diversity of the hosted miRNAs and wondered which TE families contribute more to this high diversity (Supplementary Fig. S3a). We performed a multiple linear regression analysis with the different families as predictors and found that Alu and ERVL are significantly associated with the increase of nucleotide diversity (Alu, p = 0.013; ERVL, p = 5.11e-04).

As expected, the evolutionary age is another determinant factor in the miRNA sequence diversity. We found that Pi presents a clear correlation with the miRNA conservation (Fig. 2b; see Methods), with significant differences among the different groups (Kruskal-Wallis p = 2.373e-11). The highest diversity was seen in the miRNAs classified as primate specific (group 1) and the lowest in those conserved beyond mammals (group 4).

Regarding the clustering patterns of miRNAs, we found that diversity differences between clustered and isolated miRNAs reached significant levels (Wilcoxon p = 3.663e-10) (Fig. 2c) which, as seen before, it might be a reflection of the higher conservation of clusters due to their functionality in cooperative processes (Wang et al. 2016; Kabekkodu et al. 2018) and also the fact that most of the clustered miRNAs have originated after common duplication events (Hertel et al. 2006).

Considering the above, sequence diversity levels of human miRNAs seem to be driven by their location, age and genomic context. These factors might also determine the presence of mutations in miRNA sequences that could affect their expression, hairpin folding and even their ability to bind their target genes and, therefore, be determinant for their evolutionary trajectory. Because of that, we wanted to study the integrated contribution of these factors to the observed diversity differences. We applied a multiple linear regression model to the diversity data and the different miRNA categories (genomic context, evolutionary age and clustering). The regression model showed that age (being primate specific, p =

3.3e-03), clustering (being isolated, p = 3.6e-04) and genomic context (not being intergenic, p = 0.015) are predictors significantly associated with the increase of Pi in human miRNAs.

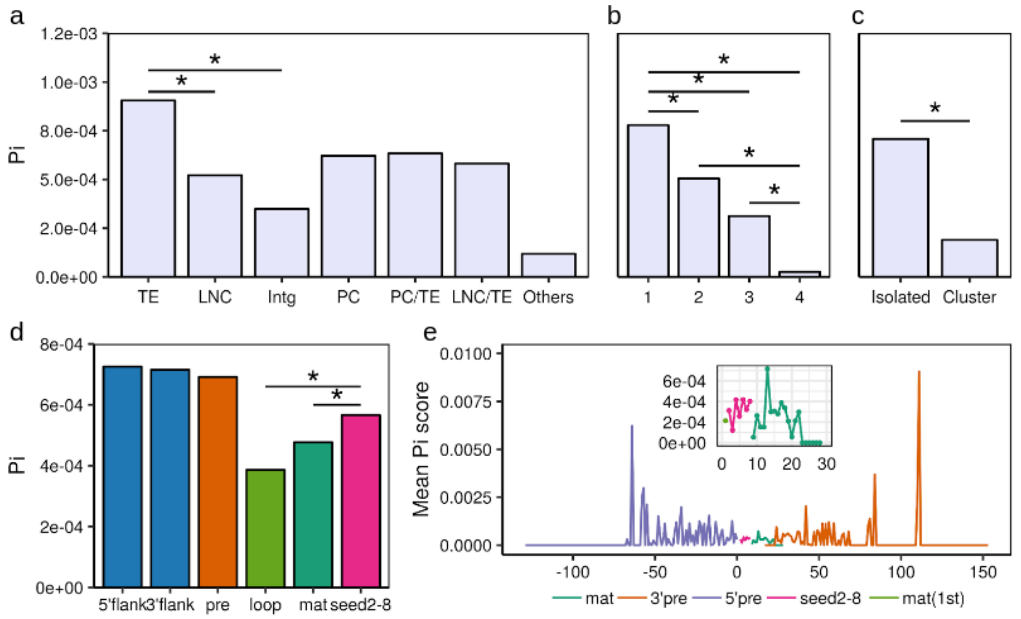


Fig. 2 Mean nucleotide diversity differences between miRNAs in different annotation categories and functional regions. **(a)** Differences between the genomic contexts where the human miRNAs are found. Wilcoxon pairwise comparisons (Bonferroni corrected) show that transposable elements (TE) present a significantly higher diversity than other environments (TE vs LNC, p = 0.022; TE vs Intg, p = 0.022. **(b)** Differences across miRNA conservation groups (see Methods). Primate-specific miRNAs (group 1) show a significantly higher diversity in comparison with the others (1 vs 2, p = 0.00057; 1 vs 3, p = 0.0178; 1 vs 4, p = 3.93e.10; Wilcoxon pairwise comparisons, Bonferroni corrected). Significant differences are also seen for the miRNAs conserved beyond mammals (group 4) (4 vs 3, p = 0.0178; 4 vs 2, p = 2.6e-05; Wilcoxon pairwise comparisons, Bonferroni corrected). **(c)** Differences between miRNAs found isolated and organised in clusters. Isolated miRNAs are associated with a significantly higher diversity than the members of clusters (Wilcoxon test, p = 3.663e-10). **(d)** Diversity comparison between the different functional regions identified in the miRNA hairpins. The seed region (2-8 nucleotides) presents a significantly higher diversity than other conserved regions (seed vs loop, p = 0.0011 and seed vs mat, p = 0.0056; Wilcoxon pairwise comparisons, Bonferroni corrected). **(e)** Mean nucleotide diversity calculated in each relative position of the precursor miRNA. The zoomed region correspond to the diversity per position found in the mature sequence

An excess of diversity in the seed region is driven by a reduced number of miRNAs

The analysis of the nucleotide diversity (Pi) across different miRNA regions indicated an overall higher diversity in the precursor and flanking regions compared to the rest of regions (Wilcoxon test $p < 0.05$). Surprisingly the loop region presented the lowest diversity of the whole miRNA hairpin (Fig. 2d). This might reflect the importance of this region in the hairpin folding, which is determinant for the processing of the primary sequence. Previous studies (Torruella-Loran et al. 2016) showed that the seed is the most conserved region of the miRNA, which has been associated with its functional relevance due its central role in target binding. However, our results showed a higher Pi in the seed than in other conserved regions, like the mature (outside seed) and the loop (Wilcoxon pairwise comparisons $p = 0.0011$ and $p = 0.0056$, respectively). It is worth noting that this level of diversity in the seed comes from the variation of a small set of miRNAs (75, 2.9%), showing that, indeed, most of the human miRNAs are conserved in their seed. On the other hand, the seed region presented values of SNP density similar to those in the mature outside the seed (Supplementary Fig. S4b), which suggests that, considering the values of nucleotide diversity, the seed region is more populated by high frequency variants than the mature region. The region-specific levels of diversity were studied in the whole range of minor allele frequency (MAF), where the seed region was consistently found with diversity levels below the mature region until a frequency ~ 50% (Supplementary Fig. S4c). This shows that no bias in the variant content is confounding these results. Overall, these data suggest that the high diversity observed in this set of miRNAs might be a consequence of the specific targeting of positive selection processes, as discussed below.

Previous reports on miRNA targeting (Grimson et al. 2007; Wheeler et al. 2009) show that not only the seed region but also

certain positions in the mature sequence are involved in target binding. To further analyse the variation in the miRNAs, nucleotide diversity was studied at position basis in the whole precursor sequence (Fig. 2e). As expected, the general pattern shows that the mature sequences are located in a valley of diversity, which confirms their overall conservation. Different levels of diversity are seen in the mature sequence. More specifically a decrease in diversity is seen at the 3' end, corresponding to the region known as participating in the complementary binding of mRNAs.

Highly differentiated miRNA SNPs are enriched in signals of positive selection and expression variation

The excess of diversity found in the seed region may respond to particular processes of positive selection that generate frequency shifts at population level. These population-specific changes could affect the miRNA binding to the target gene and change the targeting profiles. In this line, we wanted to study the population-specific patterns of diversity found within the miRNA seed regions. In Supplementary Fig. S5 we show the Pi values of the seed regions from a total of 60 miRNAs presenting genetic variants (DAF >= 5%) calculated in each of the 26 populations of our study. The clustering pattern of diversity sharing among populations reflects the similarities of demographic and potential evolutionary histories in the same continental group. As expected, African populations (AFR) are clustered separately from the other populations, showing the highest differentiation probably due to the Out-of-Africa event. A higher diversity sharing is seen among the non-African populations. There are some clear continental-specific groups of miRNAs that might be the result of demographic dynamics and/or genetic drift, but also of local processes of positive selection on certain alleles. Considering the group-specific membership of miRNA alleles we found that 37% (22) are exclusively present in AFR, while 13% (8) are found in non-Africans, private or shared among other groups (European

(EUR), American (AMR), EastAsian (EAS) and South Asian (SAS)). The other alleles are shared between African and non-African populations (50%, 30), being 21 (35%) present in all continents.

Next, mean population differentiation (F_{st}) values across all possible population comparisons were calculated for the different miRNA regions (Fig. 3a). As shown, the seed presents an overall F_{st} score higher than the rest of the mature sequence in almost all the compared groups. This tendency is stronger in comparisons including AFR populations than non-African ones. Although demographic dynamics are generally the main cause in the existing differentiation between populations, the high F_{st} values in the seed, compared to other conserved regions like the mature (outside seed) and the loop, suggest that this region could have been particularly targeted by processes of positive selection. Surprisingly, in contrast with the overall low diversity values seen before, the loop region also exhibits particularly high F_{st} scores in some comparisons, especially in the AFR vs SAS populations.

Further, we evaluated the potential functionality of the precursor region-specific SNPs by contrasting their overall Combined Annotation Dependent Depletion (CADD) score distributions, a statistic designed to measure the deleteriousness of human variants (Rentzsch et al. 2019). As shown in Fig. 3b, the CADD scores associated with the loop and seed regions are slightly higher than the rest of the precursor sequence, although non-significant. This evidence reinforces the idea that these regions are specifically implicated in processes potentially involved in adaptive selection.

We wanted to examine the extent to which the top F_{st} scoring SNPs participate in putative signatures of recent positive selection. We focused on signals characterized by the presence of long haplotypes at high (ongoing hard sweeps) and moderate frequencies (soft sweeps) in individual populations, detected by the statistics

integrated haplotype score (iHS) (Voight et al. 2006) and the number of segregating sites by length (nSL) (Ferrer-Admetlla et al. 2014) (see Methods). We pooled the SNP set (100, 16%) that showed extreme F_{st} values (>99%) in the whole miRNA precursor sequence in all population comparisons, and explored their involvement in selective sweeps. Among these top SNPs we found that 23% and 18% present extreme iHS and nSL scores (≥ 2), respectively, in at least one population, while the proportion of highly scoring SNPs in the whole dataset is only 13.8% (iHS) and 11.5% (nSL). This result suggests that highly differentiated SNPs in the precursor miRNA sequence are more likely to be found in genomic regions that hold signatures consistent with recent positive selection signatures (iHS Chi square test = 11.29, p = 7.77e-04; nSL Chi square test = 6.74, p = 9.38e-03).

Nucleotide changes in regions involved in miRNA sequence processing (pre, loop) and target binding (mature, seed) might affect the regulation of their target genes and, therefore, generate expression variation that could lead to genetic disorders, but also to phenotypic adaptations. We used the Genotype-Tissue Expression (GTEx) Project catalog (v7) of associated eQTL-eGene pairs to study the potential impact of our miRNA-harbouring top SNPs in gene expression variation (Aguet F et al. 2017). Among the top 100 SNPs in the precursor sequences, 54% (54) are reported as significant expression Quantitative Trait Loci (eQTLs) by GTEx, while the 24.7% (154) are found in the whole SNP dataset. Also, we used the most recent release of the genome-wide association studies (GWAS) catalog (v1.0) (Buniello et al. 2019) to evaluate the extent to which these highly differentiated SNPs are associated with genetic diseases and traits. In this case, 5% (5) of the top SNPs present significant associations in GWAS studies, while only 1.7% (11) are found in the whole SNP dataset. These results indicate that highly differentiated miRNA-harbouring SNPs are more likely to be reported as significant eQTLs (Chi-square test = 33.994, p = 5.528e-09) and GWAS associated SNPs (Chi square test = 6.7841, p

= 9.19e-03), which suggests their implication in expression variation and human diseases.

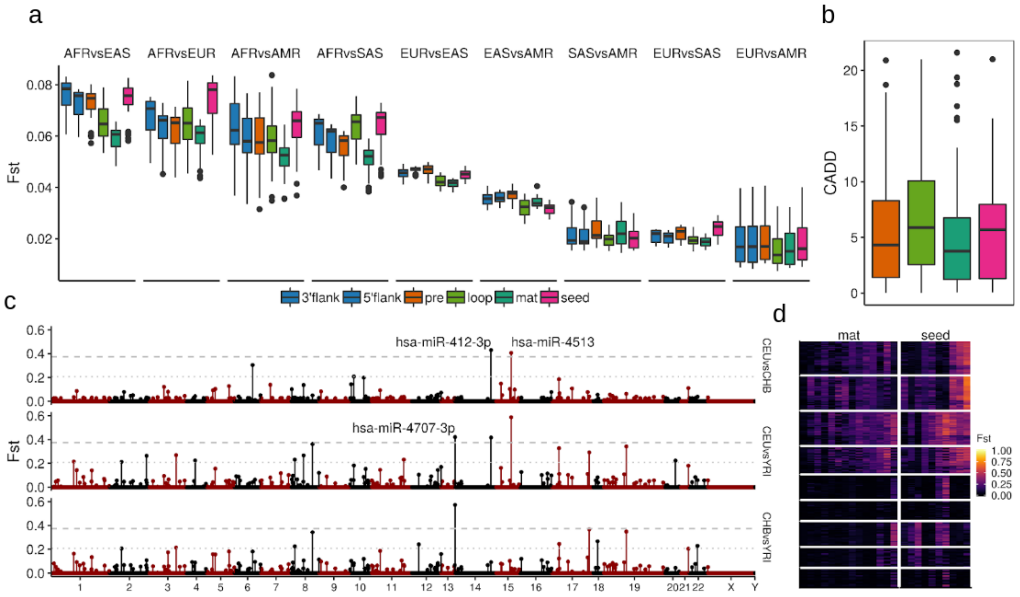


Fig. 3 Analysis of Fst values across miRNA regions and candidates. (a) Mean Fst values per miRNA region across all population comparison groups. The Fst values were calculated in all the variant regions. (b) Combined Annotation Dependent Depletion (CADD) scores distributions, as a measure of the predicted level of deleteriousness of the variants, across miRNA regions (c) Manhattan plot showing the mean Fst values per miRNA mature sequence in the three comparisons of reference. Two Fst thresholds were used to extract the potential miRNA candidates under positive selection (1% and 5%). (d) Heatmap showing the per-SNP Fst values of the variants found in the mature outside seed (14) and seed (10) region of the top 5% miRNA candidates, where the columns correspond to SNPs and rows to all possible population comparisons (243)

miRNA recent evolution might be driven by targeted processes in their seed related to positive selection and disease

In order to identify potential miRNA candidates under the selection pressures of local adaptations, we calculated mean F_{st} values in the whole mature sequence. Fig. 3c shows the genome wide distribution of mature-specific F_{st} values in the three comparisons of reference (Utah Europeans (CEU) vs Han Chinese (CHB), CEU vs Yoruba

(YRI) and CHB vs YRI), where three miRNAs are found in the top 1% (hsa-miR-1269b, hsa-miR-412-3p, hsa-miR-4707-3p) and 22 above the 5% (Table 1). Surprisingly the three most divergent miRNAs belong to conservation groups older than primate specific, which suggests that these population-specific changes might respond to potential adaptations that affect well-established regulatory pathways. These candidate miRNAs harbour 10 SNPs within their seed regions (10 miRNAs) and 14 SNPs in other positions of the mature sequence (14 miRNAs). As seen in Fig. 3d, seed-harboring SNPs like rs2273626 (hsa-miR-4707-3p) present the most extreme F_{st} scores in the candidate mature sequences and reach top values (>99.98%) in the whole miRNA distribution. Among these, seven SNPs in both seed (rs6771809, rs77651740, rs28655823, rs2273626, rs2168518, rs7210937, rs3745198) and mature regions (rs56790095, rs73239138, rs404337, rs2155248, rs61992671, rs12451747, rs73410309) were reported by GTEx as significantly associated to gene expression variation.

Chr	Mature ID	Mature SNP	Seed SNP	Evolutionary Age	Genomic Context	Max. Fst	Max. iHS	Max. nSL	CADD	Disease association
1	hsa-miR-4781-3p	-	rs74085143	Primate	PC;TE	0.21	- / 1.51	- / 1.28	- / 7.85	PD[1], AD[2]
2	hsa-miR-6071	rs56790095	-	Primate	PC;TE	0.21	0.67 / -	0.37 / -	5.64 / -	GB[3], CRC[4,5]
2	hsa-miR-6811-3p	rs2292879	-	Primate	PC;TE	0.26	2.73 / -	1.73 / -	2.71 / -	-
3	hsa-miR-6826-5p	rs115693266	rs6771809	Primate	PC	0.27	0.22 / 1.80	0.88 / 2.22	2.92 / 1.01	CRC[6], BC[7]
4	hsa-miR-1269a	rs73239138	-	Primate	TE	0.22	1.84 / -	2.12 / -	0.70 / -	GC[8], HC[9,10,16], CC[11], BC[12], LC[13,14], CRC[15]
6	hsa-miR-10524-5p	-	rs77651740	Non-classified	TE	0.30	- / 1.69	- / 1.40	- / NA	-
8	hsa-miR-1322	rs59878596	-	Non-classified	PC	0.23	1.33 / -	2.25 / -	NA / -	HC[17], ESC[18]
8	hsa-miR-4472	-	rs28655823	Primate	Intg	0.36	- / 2.02	- / 1.38	- / 2.87	BC[19,20,21], PC[21], CC[21]
8	hsa-miR-8084	rs404337	-	Non-classified	TE	0.27	1.45 / -	1.89 / -	NA / -	BC[22], OC[23]
10	hsa-miR-938	-	rs12416605	Primate	PC	0.21	- / 0.93	- / 1.78	- / 7.68	GC[24,25]
11	hsa-miR-1304-3p	rs2155248	-	Primate	PC;TE	0.23	1.72 / -	1.13 / -	4.44 / -	GC[26], HC[27], HNC[28], EM[29],LC[30]
12	hsa-miR-196a-3p	rs11614913	-	Eutherians	PC	0.24	1.74 / -	1.21 / -	18.77 / -	LC[31,38], HC[31,33], HNC[31], GM[32], OC[33], BC[33,35,37,81], DM1[34], CAD[36], CRC[76], GC[77,78,79,80]
14	hsa-miR-412-3p	rs61992671	-	Eutherians	LNC	0.43	1.44 / -	1.33 / -	15.52 / -	OS[39], CC[40]
14	hsa-miR-4707-3p	-	rs2273626	Eutherians	PC	0.57	- / 2.33	- / 1.22	- / 10.85	POAG[41], ESC[42]
15	hsa-miR-4513	-	rs2168518	Meta/Prototheria	PC	0.59	- / 2.09	- / 1.40	- / 5.01	CAD[43,46], LC[44,45], GC[47], BC[48], OSCC[49]
17	hsa-miR-548h-5p	rs9913045	-	Primate	PC;TE	0.24	2.56 / -	3.28 / -	1.31 / -	GM[50]
17	hsa-miR-1269b	rs12451747	rs7210937	Primate	PC;TE	0.33	1.67 / 1.11	2.23 / 1.27	0.31 / 0.39	OPSCC[51], LC[52]
17	hsa-miR-4739	rs73410309	-	Primate	LNC;TE	0.37	2.01 / -	3.08 / -	12.73 / -	PF[53], PC[54], DM1[55,56], GC[57], AML[58]
18	hsa-miR-4741	-	rs7227168	Eutherians	PC	0.27	- / 3.37	- / 1.74	- / 13.31	MY[59], HC[60], CRC[61], CC[61]
19	hsa-miR-6796-3p	-	rs3745198	Primate	PC	0.35	- / 2.39	- / 1.63	- / 3.67	UR[62]
20	hsa-miR-646	rs6513497	-	Primate	LNC;TE	0.22	2.99 / -	3.06 / -	6.33 / -	GC[63,68], HC[64], LAC[65], LC[66,71], BC[67], CRC[69], RC[70], OS[72]
22	hsa-miR-3928-5p	rs5997893	-	Non-classified	TE	0.23	1.36 / -	2.01 / -	NA / -	HD[73], HNC[74], OS[75]

Table 1. Top 5% miRNA candidates of different ages and genomic contexts under putative positive selection. The Max. Fst value represents the maximum mean Fst of the mature sequence among the three comparisons of reference. The selection test values (iHS and nSL) correspond to the population that exhibit the maximum value of the mature (left) and seed SNP (right). The CADD column provides the predicted deleteriousness scores (see Methods) of the mature (left) and seed SNP (right). Disease association for most of the candidates are indicated in the disease column and some examples are described in the main text: AD (Alzheimer's disease; (2) Satoh et al. 2015), AML (acute myeloid leukemia; (58) Cattaneo et al. 2015), BC (breast cancer; (7) Danková et al. 2020, (12) Sarabandi et al. 2021, (19) Li et al. 2020, (20) Wang et al. 2018, (21) Kim et al. 2012, (22) Gao et al. 2018, (33) Choupani et al. 2019, (35) Ahmad and Shah 2020, (37) Zhao et al. 2016, (48) Li et al. 2019, (67) Darvishi et al. 2020, (81) Qi et al. 2015), CAD (coronary artery disease; (36) Fragoso et al. 2019, (43) Mir et al. 2019, (46) Li et al. 2015), CC (colon cancer; (11) Mao et al. 2017, (21) Kim et al. 2012, (40) Zhu et al. 2020), CRC (colorectal cancer; (4,5) Slattery et al. 2018, (6) Kijima et al. 2017; (15) Bu et al. 2015, (61) Cojocneanu et al. 2020, (69) Dai et al. 2017, (76) Yan et al. 2017), DM1 (type 1 diabetes mellitus; (34) Ibrahim et al. 2019, (55) Delić et al. 2016, (56) Li et al. 2018), EM (endometriosis; (29) Xu et al. 2017), ESC (esophageal squamous cell carcinoma; (18) Zhang et al. 2013, (42) Bi et al. 2020), GB (glioblastoma; (3) Zhou et al. 2020), GC (gastric cancer; (8) Li et al. 2017, (24) Torruella‐Loran et al. 2019, (25) Arisawa et al. 2012, (26) Kurata and Lin 2018, (47) Ding et al. 2019, (57) Dong et al. 2015, (63) Cai et al. 2016, (68) Zhang et al. 2017, (77) Ni et al. 2015, (78) Yan et al. 2017, (79) Peng et al., 2010, (80) Wang et al 2013), GM (glioma; (32) Yang et al. 2020, (50) Ji et al. 2020), HC (hepatocellular carcinoma; (9) Min et al. 2017, (10) Xiong et al. 2015, (16) Wang et al. 2019, (17) Zhao et al. 2020, (27) Oura et al. 2019, (60) Liu et al. 2019, (64) Wang et al. 2014), HD (Huntington disease; (73) Reed et al. 2018), HNC (head and neck squamous cell carcinoma; (28) Petronacci et al. 2020, (74) Fadhil et al. 2020), LAC (laryngeal carcinoma; (65) Yuan et al. 2020), LC (lung cancer; (13) Jin et al. 2018, (14) Wang et al. 2020; (30) Othman et al. 2013, (31) Liu et al. 2018, (38) Wang et al. 2017, (44) Ghanbari M et al. 2014, (45) Ghanbari M et al. 2017, (52) Yang et al. 2020, (66) Wang et al. 2020, (71) Pan et al. 2016), MY (myeloma; (59) Zhang et al. 2019), OC (ovarian cancer; (23) Chong et al. 2015, (33) Choupani et al. 2019), OPSCC (oral and pharyngeal squamous carcinoma; (51) Chen et al. 2016), OS (osteosarcoma; (39) Martin-Guerrero et al. 2018, (72) Sun et al. 2015, (75) Xu et al. 2014), OSCC (oral squamous cell carcinoma; (49) Xu et al. 2019), PC (prostate cancer; (21) Kim et al. 2012, (54) Wang et al. 2020), PD (Parkinson disease; (1) Beecham et al. 2015), PF (pleural fibrosis; (53) Wang et al. 2019), POAG (open-angle glaucoma; (41) Ghanbari, et al. 2017), RC (renal carcinoma; (70) Li et al. 2014), UR (urolithiasis; (62) Liang et al. 2019).

As seen before, the presence of SNPs in the seed region might lead to variations of the miRNA targeting profiles. In order to evaluate the degree of change that a single SNP might generate, we adapted the *TargetScanHuman* (Agarwal et al. 2015) pipeline to predict the

allele-specific targets of the seed-variant candidates. When comparing the sets of target genes due to the ancestral and derived alleles we observed that, among the top ten miRNAs with SNPs in their seed, only two present a cosine similarity (see Methods) above 70% (hsa-miR-10524-5p and hsa-miR-4513), while the other candidates fall below 23%. This indicates the dramatic target shift that a single SNP generates and might be involved in regulatory adaptations (Table 2).

Mature ID	SNP	AA	DA	Targets (AA)	Targets (DA)	Overlapping targets	Cosine similarity
hsa-miR-938	rs12416605	C	T	2678	2594	573	0.22
hsa-miR-4472	rs28655823	G	C	3257	835	322	0.19
hsa-miR-4513	rs2168518	G	A	2532	2693	2118	0.81
hsa-miR-1269b	rs7210937	G	C	2437	3167	626	0.23
hsa-miR-4707-3p	rs2273626	C	A	1167	2592	356	0.20
hsa-miR-4741	rs7227168	C	T	3665	2231	676	0.23
hsa-miR-4781-3p	rs74085143	A	G	2339	2724	558	0.22
hsa-miR-6796-3p	rs3745198	C	G	2331	2855	484	0.19
hsa-miR-6826-5p	rs6771809	C	T	3191	2032	517	0.20
hsa-miR-10524-5p	rs77651740	G	T	2853	3332	2234	0.72

Table 2. TargetScanHuman predicted target genes of the seed-variant miRNA candidates. Two sets of target genes were predicted for each candidate holding both ancestral (AA) and derived alleles (DA). The overlap between these two lists of target genes is provided and the similarity is estimated with the cosine similarity (see Methods)

Next, we wanted to examine these candidate miRNAs with SNPs showing the highest population differentiation more in depth. We reviewed the literature looking for particular phenotypes in human populations and potential regulatory processes where these variants might be associated with. Among the ten miRNA candidates with SNPs located in the seed, all except one (hsa-miR-10524-5p) have been related to disease and, specially, with different types of cancers (Table 1), showing some of them differences among populations attributable to genetic risk factors, like in breast cancer (BC), colorectal cancer (CRC) and gastric cancer (GC) (Sung et al. 2021). Particularly, three of these miRNAs (hsa-miR-4472, hsa-miR-4513

and hsa-miR-6826-5p) were associated with BC, two (hsa-miR-4472 and hsa-miR-4741) with CRC and two (hsa-miR-938, hsa-miR-4513) with GC. In four out of the nine miRNAs related to disease the miRNA association was linked to the presence of the variant (rs12416605 in hsa-miR-938, rs7210937 in hsa-miR-1269b, rs2168518 in hsa-miR-4513 and rs2273626 in hsa-miR-4707-3p) (Table 1). When considering the 14 miRNAs candidates with SNPs located in the mature regions we observed that, all except one, for which no previous data have been reported (hsa-miR-6811), have been previously related to disease (Table 1). Among the associations with cancers showing differences on their risk among populations, five (hsa-miR-196a-3p, hsa-miR-646, hsa-miR-1269a, hsa-miR-6826-5p and hsa-miR-8084) have been associated with BC, five (hsa-miR-196a-3p, hsa-miR-646, hsa-miR-1269a, hsa-miR-6071 and hsa-miR-6826-5p) with CRC, and four (hsa-miR-196a-3p, hsa-miR-646, hsa-miR-1269a and hsa-miR-1304-3p) with GC. In four out of the 13 miRNAs related to disease the miRNAs association was linked to the presence of the variant (rs11614913 in hsa-miR-196a-3p, rs61992671 in hsa-miR-412-3p, rs6513497 in hsa-miR-646 and rs73239138 in hsa-miR-1269a) (Table 1).

In particular, for rs11614913 in hsa-miR-196a-3p (F_{st} = 0.24) the derived T allele has been associated with a decreased risk of different types of cancers, including breast and gastrointestinal cancers, principally in Asian populations. The frequency of the derived T allele is higher in East Asians (~ 54%) than in Europeans (CEU ~ 44%) and remarkably higher than in Africans (~13%) which may explain differences in the presentation of these types of cancer among populations and would agree with selective processes in this SNP. Similarly, for rs12416605 in hsa-miR-938 (F_{st} = 0.21), the derived T allele has been reported as a protective factor for the susceptibility to suffer a diffuse subtype of gastric cancer with the finding of a higher frequency of the T allele in Europeans compared with Asians (~29% vs. ~2%), which would agree with the reported

higher predisposition to gastric cancer in Asian populations (Torruella-Loran et al. 2019). In this regard, also the T allele of rs73239138 in hsa-miR-1269a (F_{st} = 0.22) has been significantly associated with a decreased risk of gastric cancer in a chinese population (Table 1).

Although most of the literature is centered on cancer diseases, other pathologies showing population differences worldwide have been linked to some of these miRNA candidates and SNPs. The T allele of rs11614913 in hsa-miR-196a-3p (highest frequency in Asian populations: 54%) shows a pleiotropic effect being not only associated with cancer but also with the risk of developing coronary artery disease (CAD) (Fragoso et al. 2019), as well as the T allele of rs2168518 in hsa-miR-4513 (highest frequency in European populations: 61%), which has been strongly associated with increased susceptibility to CAD and other related pathologies and physiological states showing risk differences among populations such as glucose homeostasis, blood pressure, and age-related macular degeneration (Mir et al. 2019; Ghanbari et al. 2014 and 2017; Li et al. 2015).

Additionally, among the SNP candidates with the highest F_{st} scores in the top 1% is rs2273626 (F_{st} = 0.57), located in the seed region of hsa-miR-4707-3p. A neuroprotective role for the derived T allele in the progression of glaucoma has been reported (Ghanbari et al. 2017), which goes in line with the negative association of rs2273626 with the disease (Springelkamp et al. 2017). This SNP shows a derived allele frequency of ~3% in African populations and more than 50% in non-Africans (Fig. 4a), which would be in agreement with the higher incidence of glaucoma in Africans (Abu-Amero et al. 2015). Furthermore, the extended haplotype homozygosity (EHH) decay on this variant indicates the presence of longer haplotypes harbouring the derived allele in non-African populations (Fig. 4b), which is consistent with the occurrence of

positive selection processes favouring the neuroprotective allele since the Out-of-Africa event.

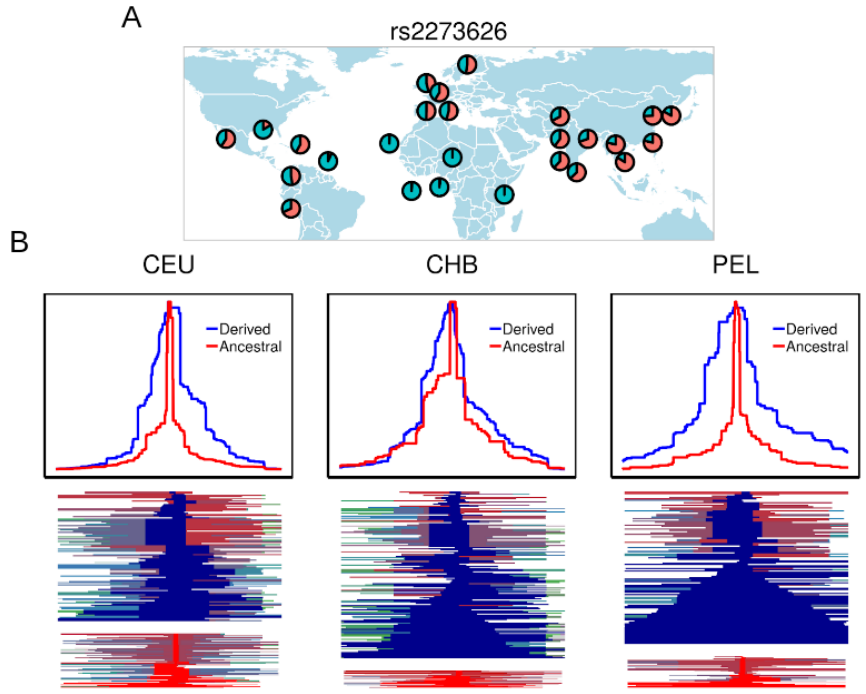


Fig. 4 Analysis of signatures of positive selection in the candidate SNP rs2273626. (a) World wide MAF distribution of rs2273626. (b) Extended haplotype homozygosity (EHH) decay in both ancestral and derived alleles of rs2273626 (upper plot) and haplotype patterns around the ancestral and derived alleles (bottom plot) in Utah Europeans (CEU), Han Chinese (CHB) and Peruvian (PEL) populations

Discussion

The increasing discovery of naturally occurring variation in the human genome, together with the improvement in annotation strategies of non-protein coding genes, has made it possible to study the potential consequences of mutations in the human miRNAs. As a dense layer of post-transcriptional regulation, miRNAs are expected to be highly susceptible to the occurrence of mutations in their sequences. However, in this analysis, along with previous studies (Carbonell et al. 2012), we discuss the unexpected level of

variation in the critical regions of these regulatory molecules and its possible relationship with evolutionary processes associated with disease.

We implemented a computational pipeline to annotate and analyse the nucleotide diversity and selection signatures of the most updated catalog of genetic variation from 1000 Genomes Project (phase III), in the most complete collection of annotated human miRNAs to date (miRBase, v.22). We integrated the analysis of miRNA variation with the most sophisticated software for target prediction to date, *TargetScanHuman*, which was adapted to predict allele-specific target genes in seed-harbouring SNP miRNAs. This method, unlike others previously published (Riffo-Campos et al. 2016), incorporates multiple features from target conservation to sequence context to generate more accurate prediction scores. As a result, this provided a robust approach to compare the allele-driven targeting and estimate the extent of the shift generated in the gene target profiles of seed-harbouring SNP miRNAs. We also integrated novedous statistical methods sensitive to different modes of selective sweeps (hard and soft) to capture a wider range of selection signatures than previously reported for human miRNAs.

Until now very few studies have considered the integrated role of the different genomic factors that might have shaped the global diversity of the human microRNAome (Gallego et al. 2016). Here we show that the expansion of new miRNAs in the primate lineage, their location in the genome and the role of hosting transposable elements are significantly associated with the increase in miRNA diversity, something that might be related with the evolutionary boost of the miRNA system in the human genome. Furthermore, against the common belief, here we report a global excess of variation in the seed, which appears as the most diverse among the traditionally conserved functional regions of miRNAs. This is in contrast with the low diversity found in the loop, which evidences the evolutionary constraints due to its role in hairpin folding. This

evidence stresses the importance of the secondary structure in maintaining the stability of the RNA molecule and determining the balance between miRNA biogenesis, particularly binding of the miRNA with the Drosha-DGCR8 complex, and miRNA turnover (Han et al. 2006; Guo et al. 2015). Moreover, the population differences found in these two regions are among the highest in the whole precursor sequence, something compatible with targeted evolutionary-driven processes that might be implicated in regulatory advantages. These processes are evaluated in the present study by identifying a global enrichment in positive selection signals (selective sweeps) among the highest differentiated SNPs across populations, showing the potential of these miRNAs and their regulatory networks to drive population-specific adaptations in agreement with some previously reported works (Quach et al., 2009; Li et al. 2012; Torruella-Loran et al. 2016).

Either by changing their targeting profiles or modifying their expression levels, it is clear that miRNA networks are more versatile to sequence changes than reported until now. We show that a significant fraction of human miRNAs participate in gene expression variation driven by the presence of eQTLs in their sequences. This goes in line with the regulatory plasticity that miRNAs have proven to hold and that might be determinant in adaptive changes at regulatory level. However, the phenotypic consequences of adaptive changes in these molecules are far to be properly understood. The great target breadth of miRNAs and the massive complexity of their regulatory networks make changes in their sequences affect multiple pathways simultaneously. Therefore, selective forces that rewire these networks might also be behind population-specific susceptibilities to different disorders. In this line, here we show that human miRNAs are also enriched in variants associated with specific human traits and diseases reported by GWAS studies. In this paper we provide a collection of miRNA alleles that were reported to affect individuals differently depending on their genetic ancestries.

In this regard, some of the miRNAs with SNPs showing the highest population differentiation have been found associated with diseases that show different population prevalence worldwide. One of the clearest examples is the case of rs12416605 in hsa-miR-938, whose derived T allele has been reported to confer protection against the diffuse subtype of gastric cancer (GC) through one of its targets, the chemokine *CXCL12* (Torruella-Loran et al. 2019), reported as playing a critical role in cell migration and invasion (Izumi et al. 2016). This cancer seems to be promoted by the amplified repression of *CXCL12,* mediated by the rs12416605 ancestral C allele (Torruella-Loran et al. 2019), which makes C-allele carriers more susceptible to develop GC metastasis. This would be in agreement with the finding of a higher frequency of the T allele in European compared with Asian populations, which is reflected by a high-global fixation index (F_{st}), and may influence the existing geographical clinical differences between Asian and non-Asian populations (Lin et al. 2015).

Among non-cancer diseases we found the T alleles of rs11614913 in hsa-miR-196a-3p and rs2168518 in hsa-miR-4513, associated with increased susceptibility to coronary artery disease (CAD). Although this disease seems to be highly dependent on environmental factors, with over 60% of current cases occurring in developing countries (Beltrame et al. 2012), population differences in CAD susceptibility are envisaged. In that context, the most striking finding is for primary open-angle glaucoma (POAG), a complex neurodegenerative disorder, dependent on environmental and genetic factors, that causes irreversible blindness and affects approximately 70 million people worldwide. Recent studies report a highly biased prevalence of the disease towards individuals with African ancestry, followed by Asians and Europeans (Abu-Amero et al. 2015). Several genes have been found associated with the progression of the disease by diverse GWAS studies. Among them, the caspase recruitment domain family member 10 (*CARD10*)

seems to confer a neuroprotective role by increasing the survival and proliferation of retinal ganglion cells (Khor et al. 2011), whose apoptosis is enhanced in POAG. In Ghanbari et al. (2017), the authors demonstrated by allele-specific *in vitro* validation that the rs2273626 derived T-allele generates a lower repression of *CARD10*. A weaker binding to the target seems to be behind this expression change, which we further validated with *TargetScanHuman*, reporting a greater repression score by the ancestral allele (0.632) than the derived allele (0.124). The authors suggest that the neuroprotective role of *CARD10* in the progression of glaucoma is associated with this lower repression, supported by the negative association of rs2273626 with the disease (Springelkamp et al. 2017). Here we report that the allele-specific regulation of *CARD10* through hsa-miR-4707-3p might contribute to the ethnic disparities prevalence of POAG and that this differential regulation is driven by processes of positive selection that promote the neuroprotective role of rs2273626 derived T-allele in non-African populations.

Here we show that, despite the strong selective pressures that maintain miRNA conservation, several miRNA variants might have suffered the effect of positive selection and may account for phenotypic diversity among human populations being, in some cases, related to disease. Even though we identify some of these miRNA variants and, in certain cases, functional data shows allele-specific regulation of specific target genes, the extent to which most of these miRNA mutations contribute to differences in disease risk among populations remains to be investigated. One of the main limitations of the analysis of positive selection in miRNAs is their small size. Haplotype-based statistics like iHS and nSL rely on the detection of long unbroken haplotypes that might span thousands of base pairs on both sides of the selected locus, which hinder the identification of the true target of selection. The intronic origin of a substantial number of human miRNAs also makes difficult the identification of the causal genomic locus of the

selection signature, potentially being originated either by the miRNA or the hosting gene. The conclusive evidence to understand the contribution of miRNAs to the recent evolutionary history of humans is the experimental validation of the genotype-phenotype association. However, the multiple potential targets of miRNAs and the side effects generated by sequence changes in the non-selected cellular processes makes this validation a difficult task. New methods and more data are needed to fill this gap between the genetic change and the phenotypic adaptation.

Materials and Methods

Human miRNA coordinates and functional region annotation

The human miRNA genomic coordinates were downloaded from the last release of the miRBase annotation database (v.22, March 2018) (Kozomara et al. 2019, http://www.mirbase.org/). This dataset contains the coordinates of 1918 human miRNA precursor transcripts and their mature sequences that were converted to hg19 genome assembly with liftOver (Hinrichs et al. 2006). From this conversion, four miRNA genes were dropped from the original dataset, and 10 were not able to be located in any chromosome, being also removed and leaving a total of 1904 precursor sequences. A custom script was designed to extract the individual functional regions of each miRNA. As shown in Fig. 1a we differentiated the "seed" region (positions 2-8), the mature ("mat") region outside the seed, the "loop" (region between two mature sequences) and the precursor regions (5' and 3' sides) outside the mature and loop. We also considered precursor flanking regions on both sides (5' and 3') of each miRNA hairpin, having the same length as the whole precursor sequences. An additional category was created in order to accommodate the regions that overlap between different miRNAs ("ovlp"), these miRNAs are treated differently due to the difficulty of analysing the overlapping regions. In the analysis of region-specific diversity the miRNAs with "ovlp" regions (71) were

discarded. A different degree of mature annotation is seen in the miRBase transcripts: 959 transcripts out of the 1904 (50.3%) present both mature sequences annotated (5p and 3p arms), allowing to completely describe the different regions of the precursor sequences. However, in 945 transcripts (49.7%) only one mature sequence is reported. In these cases, the description of the whole precursor sequence is limited to the boundaries of the single mature described (the specific boundaries of the loop region are not able to be defined). Therefore, when extracting the functional regions of the miRNA genes, the precursor region is considered as the whole portion that encompasses from the end of the given mature sequence to the start of the opposite flanking region (this would retain as "precursor" the "loop" region, the unannotated mature region and the actual premature region of that arm). The "loop" region is only extracted when the two mature sequence coordinates are given. These inconsistencies in the annotation of the miRNA transcripts are taken into account throughout the analysis (Fig. 1a).

Computational analyses of genomic context, evolutionary age and clustering annotation

A computational pipeline was used to integrate the tools to annotate miRNAs, locate variants in the miRNA sequences and perform the statistical calculations for the analysis of diversity, positive selection and target prediction. This pipeline was adapted to work in a high performance computing (HPC) environment based on the cluster management and job scheduling system SLURM. In order to obtain the genomic context of miRNAs, we intersected the GENCODE 19 protein coding gene and lncRNA gene annotations (v.29) (Frankish et al. 2019) with the miRNA coordinates with the multipurpose software *Bedtools* (Quinlan et al. 2010), which allow us to find coordinate overlaps between two or more sets of genomic regions with a minimum overlap of 1bp (*Bedtools intersect* functionality). The RepeatMasker open-4.0.5 database (repeat library 20140131)

(Smit et al. 2013-2015), which looks for interspersed repeats and low-complexity DNA (simple repeats, microsatelites), was also used in order to define the overlap of miRNAs with repetitive elements. miRNAs were classified based on their evolutionary age by merging the classifications obtained in Iwama et al. (2013) and Santpere et al. (2016). We grouped the miRNAs in the following categories: Primate-specific (group 1, previous 5 to 12 groups in Iwama et al. (2013)); Eutherians (group 2, previous 1 to 4); Metatheria and prototheria (group 3, previous -1 to 0) and Conserved beyond mammals (group 4, previous -2 to -3). The remaining 281 miRNAs were non-classified due to absence of data or discrepancies between the two studies in their evolutionary age. In order to obtain the miRNA clusters, a python-based custom script was designed to calculate the closest distance of each miRNA to any other in the same strand and chromosome. We defined miRNA clusters as groups of two or more miRNA genes separated by 10000 bp or less (Guo et al. 2014). The contributions of the genomic context, evolutionary age and clustering to the nucleotide diversity were obtained by applying a multiple linear regression model (*lm*), which is based on the programming language R (R Core Team 2020) and seeks to estimate the relationships between these factors (predictors) and the response variable (diversity).

miRNA genetic variation and nucleotide diversity

Human variation data from The 1000 Genomes project (third phase) (Auton A et al. 2015) was used to annotate the human miRNA dataset. 26 different human populations accounting for a total of 2504 individuals were considered in the analysis, including the admixed populations from South Asia (SAS) and the Americas (AMR). We used the last version of the program *BCFtools* (v.1.11) (Danecek et al. 2021), for processing and analysing high-throughput sequencing data, to extract the variants located within the miRNA sequences. Only biallelic SNPs with a MAF greater or equal than 1% in individual populations and 0.5% in the global population

were taken into account. In the case of unnamed variants, these were kept and corrected by using the physical position preceded by "rs_" as provisional SNP ID. When computing the derived allele frequency and haplotype-based statistics, the human ancestral alleles annotated in the original VCF files were used to format the REF and ALT fields and the corresponding genotypes of the individuals. Any SNP whose ancestral status was unknown or did not match with the reference or alternative alleles were removed from the dataset. The overall pairwise mismatches per SNP (pi) were calculated with *BCFtools* in the whole miRNA SNP dataset, after that the nucleotide diversity (Pi) per region was computed by obtaining the diversity per nucleotide in the whole length (L) of each functional region (Pi = pi/L). In this way we consider each category of region (flank, pre, mat, seed, loop) as a single sequence instead of calculating the nucleotide diversity in the regions of the individual miRNAs. The nucleotide diversity per position was calculated by aligning the precursor transcripts of the whole miRNA dataset and obtaining the mean pi value at each site. In this analysis, the “ovlp” regions were not taken into account due to the difficulty of interpreting the diversity properties of such overlaps.

Pathogenicity and disease associations of miRNA variants

The catalog of Combined Annotation Dependent Depletion (CADD) scores (Rentzsch et al. 2019) provides a quantitative way to measure the deleteriousness of single nucleotide polymorphisms (SNPs) in the human genome by prioritizing the functionality and diseases causing variants. This catalog was used to assess the level of pathogenicity of miRNA-harbouring SNPs as aproxy of their functionality. According to Kircher et al. (2014) a threshold of PHRED-scaled CADD score ≥ 10 is normally used to discern the 1% most deleterious SNPs of the whole human genome. We also leveraged the GWAS (v1.0) catalog (Buniello et al. 2019) to evaluate the participation of miRNA-harbouring SNPs in human traits.

Calculation of F_{st}, iHS and nSL scores

Population fixation indexes (F_{st}) were computed by using the Hudson estimator of the F_{st} statistic, which is not affected by the sample size and does not overestimate the F_{st} scores in comparison with others (Bhatia et al. 2013), in all the variant miRNAs. The calculations were performed by pairwise comparison between the 26 populations used from the 1000 Genome project dataset. These F_{st} scores were normalized by frequency by performing a linear regression of the estimator values and the global MAF, the residual values were used as the final F_{st} scores. We extended the analysis of selection with two haplotype-based statistics: iHS (Voight et al. 2006) and nSL (Ferrer-Admetlla et al. 2014). These tests rely on the detection of blocks of homozygosity by the EHH statistic (Extended Haplotype Homozygosity) introduced by (Sabeti et al. 2002). A recent positive selection signal is found when these blocks present moderately high or intermediate frequency of derived alleles. The iHS test is designed to detect ongoing hard sweep signals, signatures characterized by the presence of a single sweeping haplotype at high frequency in their way to fixation. On the other hand, nSL was designed to detect either ongoing hard and soft sweep signatures with a greater power than iHS. In the case of soft sweeps, these are signatures of selection on standing variation, where more than one haplotype is sweeping at intermediate frequencies. The calculations of iHS and nSL were computed with the software *selscan* (Szpiech et al. 2014), an application that implements different haplotype-based statistics in a multithreaded framework. We allowed for a maximum gap of 20kb and kept only SNPs with a minor allele frequency (MAF) higher than 5%. This statistic is standardized (mean 0, variance 1) by the distribution of observed scores over a range of SNPs with similar derived allele frequencies. The standardization was performed in each population separately by using the *norm* function, also contained in the *selscan* package (Voight et al. 2006).

Target predictions

The program TargetScanHuman (TSH, release 7.2) (Agarwal et al. 2015) was used to perform the miRNA target predictions. The perl-based pipeline used by the authors (http://www.targetscan.org/cgi-bin/targetscan/data_download.vert72.cgi), together with the ViennaRNA package (Lorenz et al. 2011), were implemented locally and adapted to our needs of performing predictions from a custom miRNA dataset. This pipeline is composed by three different steps: (i) target site identification across the set of 3'UTR regions of the human genome, (ii) the probability of conserved targeting (P_{ct}) calculations and (iii) the calculations of the context++ scores, which integrates different genomic features implicated in targeting efficiency. miRNA families and species information were downloaded from the *targetscan.org* Data Download page. In order to calculate the P_{ct} parameters, the 3'UTR dataset from the GENCODE version 19 (Ensembl 75) was obtained as a 84-way alignment from the same download page. As described in Agarwal et al. (2015), only the longest 3'UTR isoform of each gene was used as representative transcripts. In order to account for the miRNA variation in the target predictions, the variable positions in the miRNA seed regions (ancestral and derived states) were considered and incorporated into the TSH pipeline. Two different miRNA datasets were obtained when accounting for the ancestral and derived alleles of the SNPs found in the seed regions. As described in Agarwal et al. (2015), the accumulated weighted-scores per target gene were calculated as the sum of the individual target site weighted-scores, which is the final score associated with each target gene. As suggested by the authors, in order to remove the potential false positives we applied a custom per-site-based filtering strategy. Since negative weighted scores are associated with mRNA repression, only the per-site weighted scores below zero are considered and, from these, the per-miRNA 50th percentile was used as threshold to obtain the putative true target

sites in each miRNA. In order to analyse the overlap between the predicted targets of the derived and ancestral miRNA alleles we used the cosine similarity (Hill et al. 2014), which is calculated by the total number of overlapping genes divided by the square root of the product of the number of targets of both alleles.

Analysis of expression levels and expression variation

The catalogue of expression Quantitative Trait Loci (eQTLs) provided by the Genotype-Tissue Expression (GTEx) Project (Aguet F et al. 2017) was used to assess the implication of miRNA-harbouring variants in expression variation. Expression data from 16 different human tissues (bladder, blood, brain, breast, hair follicle, liver, lung, nasopharynx, pancreas, placenta, plasma, saliva, semen, serum, sperm and testis) was taken from Panwar et al. (2017). We used 2085 mature miRNAs from this dataset for which evolutionary age was available. Reads per million (RPM) values were analyzed for each mature miRNA separately, whose conservation status were determined by the precursor molecule following the classification criteria described before. A miRNA was considered to be expressed in a specific tissue when its reads were unequal to zero in at least one sample from that tissue. For the comparative analyses of the expression levels among conservation groups we took the total number of reads in the 16 tissues for all the miRNAs within each group.

References

Abu-Amero K, Kondkar AA, Chalam K V. (2015) An updated review on the genetics of primary open angle glaucoma. Int. J. Mol. Sci. 16:28886–28911

Agarwal V, Bell GW, Nam JW, Bartel DP (2015) Predicting effective microRNA target sites in mammalian mRNAs. Elife 4:. https://doi.org/10.7554/eLife.05005

Aguet F, Brown AA, Castel SE, et al (2017) Genetic effects on gene expression across human tissues. Nature 550:204–213. https://doi.org/10.1038/nature24277

Ahmad M, Shah AA (2020) Predictive role of single nucleotide polymorphism (rs11614913) in the development of breast cancer in Pakistani population. Per Med 17:213–227. https://doi.org/10.2217/pme-2019-0086

Arisawa T, Tahara T, Shiroeda H, et al (2012) Genetic polymorphisms of IL17A and pri-microRNA-938, targeting IL17A 3'-UTR, influence susceptibility to gastric cancer. Hum Immunol 73:747–752. https://doi.org/10.1016/j.humimm.2012.04.011

Auton A, Abecasis GR, Altshuler DM, et al (2015) A global reference for human genetic variation. Nature 526:68–74

Bartel DP (2018) Metazoan MicroRNAs. Cell 173:20–51

Bartel DP (2009) MicroRNAs: Target Recognition and Regulatory Functions. Cell 136:215–233

Beecham GW, Dickson DW, Scott WK, et al (2015) PARK10 is a major locus for sporadic neuropathologically confirmed Parkinson disease. Neurology 84:972–980. https://doi.org/10.1212/WNL.0000000000001332

Berezikov E (2011) Evolution of microRNA diversity and regulation in animals. Nat. Rev. Genet. 12:846–860

Bhatia G, Patterson N, Sankararaman S, Price AL (2013) Estimating and interpreting FST: The impact of rare variants. Genome Res 23:1514–1521. https://doi.org/10.1101/gr.154831.113

Bi Y, Guo S, Xu X, et al (2020) Decreased ZNF750 promotes angiogenesis in a paracrine manner via activating DANCR/miR-4707-3p/FOXC2 axis in esophageal squamous cell carcinoma. Cell Death Dis 11:. https://doi.org/10.1038/s41419-020-2492-2

Borel C, Deutsch S, Letourneau A, et al (2011) Identification of cis- and trans-regulatory variation modulating microRNA expression levels in human fibroblasts. Genome Res 21:68–73. https://doi.org/10.1101/gr.109371.110

Bu P, Wang L, Chen KY, et al (2015) miR-1269 promotes metastasis and forms a positive feedback loop with TGF-β. Nat Commun 6:. https://doi.org/10.1038/ncomms7879

Buniello A, Macarthur JAL, Cerezo M, et al (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res 47:D1005–D1012. https://doi.org/10.1093/nar/gky1120

Cai M, Zhang Y, Ma Y, et al (2015) Association between microRNA-499 polymorphism and gastric cancer risk in Chinese population. Bull Cancer 102:973–978. https://doi.org/10.1016/j.bulcan.2015.09.012

Cammaerts S, Strazisar M, Rijk P De, Del Favero J (2015) Genetic variants in microRNA genes: Impact on microRNA expression, function, and disease. Front. Genet. 6

Carbonell J, Alloza E, Arce P, et al (2012) A map of human microRNA variation uncovers unexpectedly high levels of variability. Genome Med 4:. https://doi.org/10.1186/gm363

Cattaneo M, Pelosi E, Castelli G, et al (2015) A miRNA signature in human cord blood stem and progenitor cells as potential

biomarker of specific acute myeloid leukemia subtypes. J Cell Physiol 230:1770–1780. https://doi.org/10.1002/jcp.24876

Chen HC, Tseng YK, Chi CC, et al (2016) Genetic variants in microRNA-146a (C > G) and microRNA-1269b (G > C) are associated with the decreased risk of oral premalignant lesions, oral cancer, and pharyngeal cancer. Arch Oral Biol 72:21–32. https://doi.org/10.1016/j.archoralbio.2016.08.010

Chong GO, Jeon HS, Han HS, et al (2015) Differential MicroRNA Expression Profiles in Primary and Recurrent Epithelial Ovarian Cancer. Anticancer Res 35(5):2611-2617.

Choupani J, Nariman-Saleh-Fam Z, Saadatian Z, et al (2019) Association of mir-196a-2 rs11614913 and mir-149 rs2292832 polymorphisms with risk of cancer: An updated meta-analysis. Front. Genet. 10

Chuong EB, Elde NC, Feschotte C (2017) Regulatory activities of transposable elements: From conflicts to benefits. Nat. Rev. Genet. 18:71–86

Cojocneanu R, Braicu C, Raduly L, et al (2020) Plasma and tissue specific miRNA expression pattern and functional analysis associated to colorectal cancer patients. Cancers (Basel) 12:. https://doi.org/10.3390/cancers12040843

Dai H, Hou K, Cai Z, et al (2017) Low-level miR-646 in colorectal cancer inhibits cell proliferation and migration by targeting NOB1 expression. Oncol Lett 14:6708–6714. https://doi.org/10.3892/ol.2017.7032

Danková Z, Grendár M, Dvorská D, et al. (2020) miRNA profile of luminal breast cancer subtypes in Slovak women. Ceska Gynekol 85(3):174-180.

Danecek P, Bonfield JK, Liddle J, et al (2021) Twelve years of SAMtools and BCFtools. Gigascience 10:. https://doi.org/10.1093/gigascience/giab008

Darvishi N, Rahimi K, Mansouri K, et al (2020) MiR-646 prevents proliferation and progression of human breast cancer cell lines by suppressing HDAC2 expression. Mol Cell Probes 53:. https://doi.org/10.1016/j.mcp.2020.101649

Delić D, Eisele C, Schmid R, et al (2016) Urinary exosomal miRNA signature in type II diabetic nephropathy patients. PLoS One 11:. https://doi.org/10.1371/journal.pone.0150154

Ding H, Shi Y, Liu X, Qiu A (2019) MicroRNA-4513 Promotes Gastric Cancer Cell Proliferation and Epithelial-Mesenchymal Transition Through Targeting KAT6B. Hum Gene Ther Clin Dev 30:142–148. https://doi.org/10.1089/humc.2019.094

Dong L, Deng J, Sun ZM, et al (2015) Interference with the β-catenin gene in gastric cancer induces changes to the miRNA expression profile. Tumor Biol 36:6973–6983. https://doi.org/10.1007/s13277-015-3415-1

Fadhil RS, Wei MQ, Nikolarakos D, et al (2020) Salivary microRNA miR-let-7a-5p and miR-3928 could be used as potential diagnostic bio-markers for head and neck squamous cell carcinoma. PLoS One 15:. https://doi.org/10.1371/journal.pone.0221779

Fernandez N, Cordiner RA, Young RS, et al (2017) Genetic variation and RNA structure regulate microRNA biogenesis. Nat Commun 8:. https://doi.org/10.1038/ncomms15114

Ferrer-Admetlla A, Liang M, Korneliussen T, Nielsen R (2014) On detecting incomplete soft or hard selective sweeps using haplotype

structure. Mol Biol Evol 31:1275–1291. https://doi.org/10.1093/molbev/msu077

Feschotte C (2008) Transposable elements and the evolution of regulatory networks. Nat. Rev. Genet. 9:397–405

Fragoso JM, Ramírez-Bello J, Martínez-Ríos MA, et al (2019) miR-196a2 (rs11614913) polymorphism is associated with coronary artery disease, but not with in-stent coronary restenosis. Inflamm Res 68:215–221. https://doi.org/10.1007/s00011-018-1206-z

França GS, Vibranovski MD, Galante PAF (2016) Host gene constraints and genomic context impact the expression and evolution of human microRNAs. Nat Commun 7:. https://doi.org/10.1038/ncomms11438

Frankish A, Diekhans M, Ferreira AM, et al (2019) GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res 47:D766–D773. https://doi.org/10.1093/nar/gky955

Friedman RC, Farh KKH, Burge CB, Bartel DP (2009) Most mammalian mRNAs are conserved targets of microRNAs. Genome Res 19:92–105. https://doi.org/10.1101/gr.082701.108

Gallego A, Melé M, Balcells I, et al (2016) Functional Implications of Human-Specific Changes in Great Ape microRNAs. PLoS One 11:. https://doi.org/10.1371/journal.pone.0154194

Gao Y, Ma H, Gao C, et al (2018) Tumor-promoting properties of miR-8084 in breast cancer through enhancing proliferation, suppressing apoptosis and inducing epithelial-mesenchymal transition. J Transl Med 16:. https://doi.org/10.1186/s12967-018-1419-5

Ghanbari M, de Vries PS, de Looper H, et al (2014) A Genetic variant in the seed region of miR-4513 shows pleiotropic effects on lipid and glucose homeostasis, blood pressure, and coronary artery disease. Hum Mutat 35:1524–1531. https://doi.org/10.1002/humu.22706

Ghanbari M, Erkeland SJ, Xu L, et al (2017) Genetic variants in microRNAs and their binding sites within gene 3′UTRs associate with susceptibility to age-related macular degeneration. Hum Mutat 38:827–838. https://doi.org/10.1002/humu.23226

Ghanbari M, Iglesias AI, Springelkamp H, et al (2017) A genome-wide scan for microrna-related genetic variants associated with primary open-angle glaucoma. Investig Ophthalmol Vis Sci 58:5368–5377. https://doi.org/10.1167/iovs.17-22410

Gong J, Tong Y, Zhang HM, et al (2012) Genome-wide identification of SNPs in MicroRNA genes and the SNP effects on MicroRNA target binding and biogenesis. Hum Mutat 33:254–263. https://doi.org/10.1002/humu.21641

Grigelioniene G, Suzuki HI, Taylan F, et al (2019) Gain-of-function mutation of microRNA-140 in human skeletal dysplasia. Nat Med 25:583–590. https://doi.org/10.1038/s41591-019-0353-2

Grimson A, Farh KKH, Johnston WK, et al (2007) MicroRNA Targeting Specificity in Mammals: Determinants beyond Seed Pairing. Mol Cell 27:91–105. https://doi.org/10.1016/j.molcel.2007.06.017

Guo L, Zhao Y, Zhang H, et al (2014) Integrated evolutionary analysis of human miRNA gene clusters and families implicates evolutionary relationships. Gene 534:24–32. https://doi.org/10.1016/j.gene.2013.10.037

Guo Y, Liu J, Elfenbein SJ, et al (2015) Characterization of the mammalian miRNA turnover landscape. Nucleic Acids Res 43:2326–2341. https://doi.org/10.1093/nar/gkv057

Han J, Lee Y, Yeom KH, et al (2006) Molecular Basis for the Recognition of Primary microRNAs by the Drosha-DGCR8 Complex. Cell 125:887–901. https://doi.org/10.1016/j.cell.2006.03.043

He S, Ou H, Zhao C, Zhang J (2018) Clustering pattern and functional effect of SNPs in human miRNA seed regions. Int J Genomics 2018:. https://doi.org/10.1155/2018/2456076

Heimberg AM, Sempere LF, Moy VN, et al (2008) MicroRNAs and the advent of vertebrate morphological complexity. Proc Natl Acad Sci U S A 105:2946–2950. https://doi.org/10.1073/pnas.0712259105

Hertel J, Lindemeyer M, Missal K, et al (2006) The expansion of the metazoan microRNA repertoire. BMC Genomics 7:. https://doi.org/10.1186/1471-2164-7-25

Hill CG, Jabbari N, Matyunina L V., McDonald JF (2014) Functional and evolutionary significance of human microRNA seed region mutations. PLoS One 9:. https://doi.org/10.1371/journal.pone.0115241

Hinrichs AS, Karolchik D, Baertsch R, et al (2006) The UCSC Genome Browser Database: update 2006. Nucleic Acids Res 34:. https://doi.org/10.1093/nar/gkj144

Ibrahim AA, Ramadan A, Wahby AA, et al (2019) Micro-RNA 196a2 expression and miR-196a2 (rs11614913) polymorphism in T1DM: A pilot study. J Pediatr Endocrinol Metab 32:. https://doi.org/10.1515/jpem-2019-0226

Iwama H, Kato K, Imachi H, et al (2013) Human microRNAs originated from two periods at accelerated rates in mammalian evolution. Mol Biol Evol 30:613–626. https://doi.org/10.1093/molbev/mss262

Izumi D, Ishimoto T, Miyake K, et al (2016) CXCL12/CXCR4 activation by cancer-associated fibroblasts promotes integrin β1 clustering and invasiveness in gastric cancer. Int J Cancer 138:1207–1219. https://doi.org/10.1002/ijc.29864

Ji B, Chen L, Cai Q, et al (2020) Identification of an 8-miRNA signature as a potential prognostic biomarker for glioma. PeerJ 8:. https://doi.org/10.7717/peerj.9943

Jin RH, Yu DJ, Zhong M (2018) MiR-1269a acts as an onco-miRNA in non-small cell Lung cancer via down-regulating SOX6. Eur Rev Med Pharmacol Sci 22:4888–4897. https://doi.org/10.26355/eurrev_201808_15625

Kabekkodu SP, Shukla V, Varghese VK, et al (2018) Clustered miRNAs and their role in biological functions and diseases. Biol Rev 93:1955–1986. https://doi.org/10.1111/brv.12428

Khor CC, Ramdas WD, Vithana EN, et al (2011) Genome-wide association studies in Asians confirm the involvement of ATOH7 and TGFBR3, and further identify CARD10 as a novel locus influencing optic discarea. Hum Mol Genet 20:1864–1872. https://doi.org/10.1093/hmg/ddr060

Kijima T, Hazama S, Tsunedomi R, et al (2017) MicroRNA-6826 and-6875 in plasma are valuable non-invasive biomarkers that predict the efficacy of vaccine treatment against metastatic colorectal cancer. Oncol Rep 37:23–30. https://doi.org/10.3892/or.2016.5267

Kim HK, Prokunina-Olsson L, Chanock SJ (2012) Common Genetic Variants in miR-1206 (8q24.2) and miR-612 (11q13.3) Affect Biogenesis of Mature miRNA Forms. PLoS One 7:. https://doi.org/10.1371/journal.pone.0047454

Kircher M, Witten DM, Jain P, et al (2014) A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46:310–315. https://doi.org/10.1038/ng.2892

Kozomara A, Birgaoanu M, Griffiths-Jones S (2019) MiRBase: From microRNA sequences to function. Nucleic Acids Res 47:D155–D162. https://doi.org/10.1093/nar/gky1141

Kurata JS, Lin RJ (2018) MicroRNA-focused CRISPR-Cas9 library screen reveals fitness-associated miRNAs. RNA 24:966–981. https://doi.org/10.1261/rna.066282.118

Lewis BP, Burge CB, Bartel DP (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 120:15–20

Li Q, Chen L, Chen D, Wu X, Chen M (2015) Influence of microRNA-related polymorphisms on clinical outcomes in coronary artery disease. Am J Transl Res 7(2):393-400.

Li JY, Cheng B, Wang XF, et al (2018) Circulating MicroRNA-4739 May Be a Potential Biomarker of Critical Limb Ischemia in Patients with Diabetes. Biomed Res Int 2018:. https://doi.org/10.1155/2018/4232794

Li W, Liu M, Feng Y, et al (2014) Downregulated miR-646 in clear cell renal carcinoma correlated with tumour metastasis by targeting the nin one binding protein (NOB1). Br J Cancer 111:1188–1200. https://doi.org/10.1038/bjc.2014.382

Li J, Liu Y, Xin X, et al (2012) Evidence for positive selection on a number of microRNA regulatory interactions during recent human evolution. PLoS Genet 8:. https://doi.org/10.1371/journal.pgen.1002578

Li W, Zhang H, Min P, et al (2017) Downregulated miRNA-1269a variant (rs73239138) decreases the susceptibility to gastric cancer via targeting ZNF70. Oncol Lett 14:6345–6354. https://doi.org/10.3892/ol.2017.7091

Li Y, Wang YW, Chen X, et al (2020) MicroRNA-4472 Promotes Tumor Proliferation and Aggressiveness in Breast Cancer by Targeting RGMA and Inducing EMT. Clin Breast Cancer 20:e113–e126. https://doi.org/10.1016/j.clbc.2019.08.010

Li Y, Zhu H, Wang J, Qian X, Li N (2019) miR-4513 promotes breast cancer progression through targeting TRIM3. Am J Transl Res 11(4):2431-2438.

Liang X, Lai Y, Wu W, et al (2019) LncRNA-miRNA-mRNA expression variation profile in the urine of calcium oxalate stone patients. BMC Med Genomics 12:57. https://doi.org/10.1186/s12920-019-0502-y

Lin SJ, Gagnon-Bartsch JA, Tan IB, et al (2015) Signatures of tumour immunity distinguish Asian and non-Asian gastric adenocarcinomas. Gut 64:1721–1731. https://doi.org/10.1136/gutjnl-2014-308252

Liu J, Yan J, Zhou C, et al (2015) miR-1285-3p acts as a potential tumor suppressor miRNA via downregulating JUN expression in hepatocellular carcinoma. Tumor Biol 36:219–225. https://doi.org/10.1007/s13277-014-2622-5

Liu Y, He A, Liu B, et al (2018) Rs11614913 polymorphism in miRNA-196a2 and cancer risk: An updated meta-analysis. Onco. Targets. Ther. 11:1121–1139

Lorenz R, Bernhart SH, Höner zu Siederdissen C, et al (2011) ViennaRNA Package 2.0. Algorithms Mol Biol 6:. https://doi.org/10.1186/1748-7188-6-26

Lu J, Clark AG (2012) Impact of microRNA regulation on variation in human gene expression. Genome Res 22:1243–1254. https://doi.org/10.1101/gr.132514.111

Lu J, Shen Y, Wu Q, et al (2008) The birth and death of microRNA genes in Drosophila. Nat Genet 40:351–355. https://doi.org/10.1038/ng.73

Mao Y, Zou C, Meng F, et al (2018) The SNPs in pre-miRNA are related to the response of capecitabine-based therapy in advanced colon cancer patients. Oncotarget 9:6793–6799. https://doi.org/10.18632/oncotarget.23190

Martin-Guerrero I, Bilbao-Aldaiturriaga N, Gutierrez-Camino A, et al (2018) Variants in the 14q32 miRNA cluster are associated with osteosarcoma risk in the Spanish population. Sci Rep 8:. https://doi.org/10.1038/s41598-018-33712-4

Min P, Li W, Zeng D, et al (2017) A single nucleotide variant in microRNA-1269a promotes the occurrence and process of hepatocellular carcinoma by targeting to oncogenes SPATS2L and LRP6. Bull Cancer 104:311–320. https://doi.org/10.1016/j.bulcan.2016.11.021

Mir R, Jha C k, Elfaki I, et al (2019) Incidence of MicroR-4513C/T Gene Variability in Coronary Artery Disease - A case-Control Study. Endocrine, Metab Immune Disord - Drug Targets

19:1216–1223.
https://doi.org/10.2174/1871530319666190417111940

Ni Q, Ji A, Yin J, et al (2015) Effects of two common polymorphisms rs2910164 in miR-146a and rs11614913 in miR-196a2 on gastric cancer susceptibility. Gastroenterol. Res. Pract. 2015

Othman N, In LLA, Harikrishna JA, Hasima N (2013) Bcl-xL silencing induces alterations in hsa-miR-608 expression and subsequent cell death in A549 and SK-LU1 human lung adenocarcinoma cells. PLoS One 8:. https://doi.org/10.1371/journal.pone.0081735

Oura K, Fujita K, Morishita A, et al (2019) Serum microRNA-125a-5p as a potential biomarker of HCV-associated hepatocellular carcinoma. Oncol Lett 18:882–890. https://doi.org/10.3892/ol.2019.10385

Pan Y, Chen Y, Ma D, et al (2016) miR-646 is a key negative regulator of EGFR pathway in lung cancer. Exp Lung Res 42:286–295. https://doi.org/10.1080/01902148.2016.1207726

Panwar B, Omenn GS, Guan Y (2017) MiRmine: A database of human miRNA expression profiles. Bioinformatics 33:1554–1560. https://doi.org/10.1093/bioinformatics/btx019

Peng S, Kuang Z, Sheng C, et al (2010) Association of MicroRNA-196a-2 gene polymorphism with gastric cancer risk in a Chinese population. Dig Dis Sci 55:2288–2293. https://doi.org/10.1007/s10620-009-1007-x

Petri R, Brattås PL, Sharma Y, et al (2019) LINE-2 transposable elements are a source of functional human microRNAs and target sites. PLoS Genet 15:. https://doi.org/10.1371/journal.pgen.1008036

Petronacci CMC, García AG, Iruegas EP, et al (2020) Identification of prognosis associated microRNAs in HNSCC subtypes based on TCGA dataset. Med 56:1–10. https://doi.org/10.3390/medicina56100535

Piriyapongsa J, Mariño-Ramírez L, Jordan IK (2007) Origin and evolution of human microRNAs from transposable elements. Genetics 176:1323–1337. https://doi.org/10.1534/genetics.107.072553

Qi P, Wang L, Zhou B, et al (2015) Associations of miRNA polymorphisms and expression levels with breast cancer risk in the Chinese population. Genet Mol Res 14:6289–6296. https://doi.org/10.4238/2015.June.11.2

Qin S, Jin P, Zhou X, et al (2015) The role of transposable elements in the origin and evolution of microRNAs in human. PLoS One 10:. https://doi.org/10.1371/journal.pone.0131365

Quach H, Barreiro LB, Laval G, et al (2009) Signatures of Purifying and Local Positive Selection in Human miRNAs. Am J Hum Genet 84:316–327. https://doi.org/10.1016/j.ajhg.2009.01.022

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. 2020. https://www.r-project.org/.

Rawlings-Goss RA, Campbell MC, Tishkoff SA (2014) Global population-specific variation in miRNA associated with cancer risk and clinical biomarkers. BMC Med Genomics 7:. https://doi.org/10.1186/1755-8794-7-53

Reed ER, Latourelle JC, Bockholt JH, et al (2018) MicroRNAs in CSF as prodromal biomarkers for Huntington disease in the

PREDICT-HD study. Neurology 90:E264–E272. https://doi.org/10.1212/WNL.0000000000004844

Rentzsch P, Witten D, Cooper GM, et al (2019) CADD: Predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res 47:D886–D894. https://doi.org/10.1093/nar/gky1016

Riffo-Campos ÁL, Riquelme I, Brebi-Mieville P (2016) Tools for sequence-based miRNA target prediction: What to choose? Int. J. Mol. Sci. 17

Sabeti PC, Reich DE, Higgins JM, et al (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832–837. https://doi.org/10.1038/nature01140

Santpere G, Lopez-Valenzuela M, Petit-Marty N, et al (2016) Differences in molecular evolutionary rates among microRNAs in the human and chimpanzee genomes. BMC Genomics 17:. https://doi.org/10.1186/s12864-016-2863-3

Sarabandi S, Sattarifard H, Kiumarsi M, et al (2021) Association between Genetic Polymorphisms of miR-1307, miR-1269, miR-3117 and Breast Cancer Risk in a Sample of South East Iranian Women. Asian Pacific J Cancer Prev 22:201–208. https://doi.org/10.31557/APJCP.2021.22.1.201

Satoh JI, Kino Y, Niida S (2015) MicroRNA-Seq data analysis pipeline to identify blood biomarkers for alzheimer's disease from public data. Biomark Insights 2015:21–31. https://doi.org/10.4137/BMI.S25132

Saunders MA, Liang H, Li WH (2007) Human polymorphism at microRNAs and microRNA target sites. Proc Natl Acad Sci U S A 104:3300–3305. https://doi.org/10.1073/pnas.0611347104

Sethupathy P, Collins FS (2008) MicroRNA target site polymorphisms and human disease. Trends Genet 24:489–497. https://doi.org/10.1016/j.tig.2008.07.004

Slattery ML, Mullany LE, Sakoda LC, et al (2018) The MAPK-signaling pathway in colorectal cancer: Dysregulated genes and their association with micrornas. Cancer Inform 17:. https://doi.org/10.1177/1176935118766522

Slattery ML, Mullany LE, Sakoda LC, et al (2018) The PI3K/AKT signaling pathway: Associations of miRNAs with dysregulated gene expression in colorectal cancer. Mol Carcinog 57:243–261. https://doi.org/10.1002/mc.22752

Smit, AFA, Hubley, R & Green, P. *RepeatMasker Open-4.0*. 2013-2015 <http://www.repeatmasker.org>

Springelkamp H, Iglesias AI, Mishra A, et al (2017) New insights into the genetics of primary open-angle glaucoma based on meta-analyses of intraocular pressure and optic disc characteristics. Hum Mol Genet 26:438–453. https://doi.org/10.1093/hmg/ddw399

Sun X hui, Geng X lin, Zhang J, Zhang C (2015) miRNA-646 suppresses osteosarcoma cell metastasis by downregulating fibroblast growth factor 2 (FGF2). Tumor Biol 36:2127–2134. https://doi.org/10.1007/s13277-014-2822-z

Sung H, Ferlay J, Siegel RL, et al (2021) Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 71:. https://doi.org/10.3322/caac.21660

Szpiech ZA, Hernandez RD (2014) Selscan: An efficient multithreaded program to perform EHH-based scans for positive

selection. Mol Biol Evol 31:2824–2827. https://doi.org/10.1093/molbev/msu211

Torruella-Loran I, Laayouni H, Dobon B, et al (2016) MicroRNA Genetic Variation: From Population Analysis to Functional Implications of Three Allele Variants Associated with Cancer. Hum Mutat 37:1060–1073. https://doi.org/10.1002/humu.23045

Torruella-Loran I, Ramirez Viña MK, Zapata-Contreras D, et al (2019) rs12416605:C>T in MIR938 associates with gastric cancer through affecting the regulation of the CXCL12 chemokine gene. Mol Genet Genomic Med 7:. https://doi.org/10.1002/mgg3.832

Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS Biol 4:0446–0458. https://doi.org/10.1371/journal.pbio.0040072

Wang F, Sun GP, Zou YF, et al (2013) Quantitative assessment of the association between miR-196a2 rs11614913 polymorphism and gastrointestinal cancer risk. Mol Biol Rep 40:109–116. https://doi.org/10.1007/s11033-012-2039-4

Wang J, Shu H, Guo S (2020) MiR-646 suppresses proliferation and metastasis of non-small cell lung cancer by repressing FGF2 and CCND2. Cancer Med 9:4360–4370. https://doi.org/10.1002/cam4.3062

Wang J, Liu Q, Yuan S, et al (2017) Genetic predisposition to lung cancer: Comprehensive literature integration, meta-analysis, and multiple evidence assessment of candidate-gene association studies. Sci Rep 7:. https://doi.org/10.1038/s41598-017-07737-0

Wang L, Sinnott-Armstrong N, Wagschal A, et al (2020) A MicroRNA Linking Human Positive Selection and Metabolic

Disorders. Cell 183:684-701.e14. https://doi.org/10.1016/j.cell.2020.09.017

Wang M, Xiong L, Jiang LJ, et al (2019) miR-4739 mediates pleural fibrosis by targeting bone morphogenetic protein 7. EBioMedicine 41:670–682. https://doi.org/10.1016/j.ebiom.2019.02.057

Wang R, Zhang J, Jiang W, et al (2014) Association between a variant in MicroRNA-646 and the susceptibility to hepatocellular carcinoma in a large-scale population. Sci World J 2014:. https://doi.org/10.1155/2014/312704

Wang X, Chen Q, Wang X, et al (2020) ZEB1 activated-VPS9D1-AS1 promotes the tumorigenesis and progression of prostate cancer by sponging miR-4739 to upregulate MEF2D. Biomed Pharmacother 122:. https://doi.org/10.1016/j.biopha.2019.109557

Wang X, Gao J, Zhou B, et al (2019) Identification of prognostic markers for hepatocellular carcinoma based on miRNA expression profiles. Life Sci 232:. https://doi.org/10.1016/j.lfs.2019.116596

Wang X, Jiang X, Li J, et al (2020) Serum exosomal miR-1269a serves as a diagnostic marker and plays an oncogenic role in non-small cell lung cancer. Thorac Cancer 11:3436–3447. https://doi.org/10.1111/1759-7714.13644

Wang YW, Zhang W, Ma R (2018) Bioinformatic identification of chemoresistance-associated microRNAs in breast cancer based on microarray data. Oncol Rep 39:1003–1010. https://doi.org/10.3892/or.2018.6205

Wang Y, Luo J, Zhang H, Lu J (2016) MicroRNAs in the Same Clusters Evolve to Coordinately Regulate Functionally Related

Genes. Mol Biol Evol 33:2232–2247. https://doi.org/10.1093/molbev/msw089

Wheeler BM, Heimberg AM, Moy VN, et al (2009) The deep evolution of metazoan microRNAs. Evol Dev 11:50–68. https://doi.org/10.1111/j.1525-142X.2008.00302.x

Xiong G, Wang Y, Ding Q, Yang L. (2015) Hsa-mir-1269 genetic variant contributes to hepatocellular carcinoma susceptibility through affecting SOX6. Am J Transl Res ;7(10):2091-2098.

Xu H, Liu X, Zhao J (2014) Down-regulation of mir-3928 promoted osteosarcoma growth. Cell Physiol Biochem 33:1547–1556. https://doi.org/10.1159/000358718

Xu X, Li Z, Liu J, et al (2017) MicroRNA expression profiling in endometriosis-associated infertility and its relationship with endometrial receptivity evaluated by ultrasound. J Xray Sci Technol 25:523–532. https://doi.org/10.3233/XST-17286

Xu YX, Sun J, Xiao WL, et al (2019) MiR-4513 mediates the proliferation and apoptosis of oral squamous cell carcinoma cells via targeting CXCL17. Eur Rev Med Pharmacol Sci 23:3821–3828. https://doi.org/10.26355/eurrev_201905_17809

Yan W, Gao X, Zhang S (2017) Association of miR-196a2 rs11614913 and miR-499 rs3746444 polymorphisms with cancer risk: A meta-analysis. Oncotarget 8:114344–114359. https://doi.org/10.18632/oncotarget.22547

Yang S, Zheng Y, Zhou L, et al (2020) miR-499 rs3746444 and miR-196a-2 rs11614913 Are Associated with the Risk of Glioma, but Not the Prognosis. Mol Ther - Nucleic Acids 22:340–351. https://doi.org/10.1016/j.omtn.2020.08.038

Yang W, Xiao W, Cai Z, et al (2020) miR-1269b drives cisplatin resistance of human non-small cell lung cancer via modulating the PTEN/PI3K/AKT signaling pathway. Onco Targets Ther 13:109–118. https://doi.org/10.2147/OTT.S225010

Zhang P, Tang WM, Zhang H, et al (2017) MiR-646 inhibited cell proliferation and EMT-induced metastasis by targeting FOXK1 in gastric cancer. Br J Cancer 117:525–534. https://doi.org/10.1038/bjc.2017.181

Zhang T, Zhao D, Wang Q, et al (2013) MicroRNA-1322 regulates ECRG2 allele specifically and acts as a potential biomarker in patients with esophageal squamous cell carcinoma. Mol Carcinog 52:581–590. https://doi.org/10.1002/mc.21880

Zhang Z yao, Li Y chen, Geng C ying, et al (2019) Serum exosomal microRNAs as novel biomarkers for multiple myeloma. Hematol Oncol 37:409–417. https://doi.org/10.1002/hon.2639

Zhao H, Xu J, Zhao D, et al (2016) Somatic Mutation of the SNP rs11614913 and Its Association with Increased MIR 196A2 Expression in Breast Cancer. DNA Cell Biol 35:81–87. https://doi.org/10.1089/dna.2014.2785

Zhao M, Dong G, Meng Q, et al (2020) Circ-HOMER1 enhances the inhibition of miR-1322 on CXCL6 to regulate the growth and aggressiveness of hepatocellular carcinoma cells. J Cell Biochem 121:4440–4449. https://doi.org/10.1002/jcb.29672

Zhou Y, An H, Wu G (2020) Microrna-6071 suppresses glioblastoma progression through the inhibition of pi3k/akt/ mtor pathway by binding to ulbp2. Onco Targets Ther 13:9429–9441. https://doi.org/10.2147/OTT.S265791

Zhu K, Wang Y, Liu L, et al (2020) Long non-coding RNA MBNL1-AS1 regulates proliferation, migration, and invasion of cancer stem cells in colon cancer by interacting with MYL9 via sponging microRNA-412-3p. Clin Res Hepatol Gastroenterol 44:101–114. https://doi.org/10.1016/j.clinre.2019.05.001

Zhu M, Wang F, Mi H, et al (2020) Long noncoding RNA MEG3 suppresses cell proliferation, migration and invasion, induces apoptosis and paclitaxel-resistance via miR-4513/PBLD axis in breast cancer cells. Cell Cycle 19:3277–3288. https://doi.org/10.1080/15384101.2020.1839700

IV. DISCUSSION

Discussion

The first two decades of the twenty-first century have witnessed revolutionary advances in a multitude of ambits of the human dimension. The exponential development of computing and sequencing technologies during the last twenty years has put humans in a very unique and interesting position in our history, with the acknowledgment of the complexity of our genome and its relationship with phenotypic traits and local environments. The publication of the first draft of the human genome was just the beginning of an increasing emergence of computational, sequencing and statistical methods devoted to disentangle the diversity of elements, genetic changes and regulatory interactions that have driven the evolutionary trajectory of *Homo sapiens*.

The study of natural selection in human populations has been possible due to the application of these sequencing technologies to groups of people from diverse genetic backgrounds. Although the ethnic bias in projects that seek to catalog human variation is still a subject to be fully addressed by research institutions, the extent currently reached by available human genetic data has made possible the analysis of signatures left by natural selection in virtually all the globe. A project of reference used in this kind of analysis is the 1000 Genomes Project which, as described in the introduction of this thesis, offers a very complete description of the genetic variation in a wide range of human populations. An example of the use of this database in the analysis of selection signatures is the recent creation of a publicly available population genomics-oriented genome browser called *PopHuman* (Casillas et al. 2018). This work, listed as part of the projects developed during my thesis, catalogues a series of metrics that seek to describe the population-specific nucleotide diversity and selection signatures, among others.

Nevertheless, the analysis of signatures of natural selection has suffered from a traditional preference of focusing on protein-coding genes, since their signatures generally outstand in genomic scans and are revealed as clear outliers from a neutral background. They can also be linked more easily to phenotypic traits that help to understand the underlying causes of these adaptations. The remarkable impact of coding mutations and the relatively simple interpretation of their consequences have positioned these signatures as the protagonists of these screenings. However, the early predictions made by King and Wilson in 1975 about the major role of gene regulation in the evolution of the Homo sapiens seems to have been gradually fulfilled during these last years. Again, as the major discovery engine in recent molecular biology, high throughput sequencing technologies have revealed the complexities of the regulatory genome with projects like ENCODE, which provides an increasing and rich catalog of regulatory elements in different human tissues.

In the first work presented in this thesis, "*Chromosome X-wide analysis of positive selection in human populations: from common and private signals to selection impact on inactivated genes and enhancers-like signatures*", we addressed both the analysis of X-linked signatures of positive selection using the most updated catalog of genetic variation in human populations (1000 Genomes, phase III), and the overly implication of regulatory regions in adaptive processes in the sexual chromosome. In this study we wanted to cover the necessity of capturing not only the classical hard sweep signatures but also other modes of selection not well represented in genomic scans as the previous ones, like signatures characterized by the presence of multiple sweeping haplotypes (soft sweeps). We generated a comprehensive catalog of positive selection signals across the three main continental groups (Sub-Saharan Africa, Europe and East Asia), which resulted in sets of genes overly enriched in neural development and reproduction-related processes. This catalog implies an extension of

previously reported lists of genes under positive selection, with the remarkable presence of a global signature of ancient selection in a key gene involved in neural development in the olfactory system, the *TENM1*. Two of the main characteristic phenomena in the X chromosome are represented in our selection scan: the faster-X effect and the incomplete inactivation of female-biased genes. In the comparison with autosomes, the X chromosome reveals slightly stronger signatures of positive selection, which corroborate the possibility of higher efficiency of selection processes due to the hemizygous state of males. Also, the set of escape genes identified by previous studies seem to be specifically targeted by hard sweep signatures, which indicates a potential advantageous effect of female traits driven by the overexpression and higher exposure of these incompletely inactivated genes. The implication of regulatory elements in selection processes are comprehensively addressed in this study. The overly description of selection signatures in non-genic parts of the X chromosome is revealed as caused by the presence of regulatory elements. We specifically analysed the regulatory effect of three of the top enhancer candidates under positive selection. The luciferase-based assays we performed revealed that there are population-specific distributions of haplotypes that drive differential regulatory activities in different geographic locations, something that indicates local processes of adaptive evolution at regulatory level.

Apart from participating in processes of adaptive evolution at population level, *cis*-regulatory elements (cREs) have long been believed to play a key role in the control of basic cellular functions and the determination of tissue identity. Considering that all cell types in an organism present the same genetic information, how the different temporal and spatial regulatory programs are carried out in order to generate such a wide repertoire of tissue-specific functions? A factor that seems to contribute to the tissue-specific regulatory control is the genomic location of cREs, in particular the role of elements identified by enhancer-like signatures (ELSs). Different

reports have associated the regulation of key genes by enhancers located in their introns. Also, studies in *Drosophila* show a distinctive location signature of enhancers that perform a differential regulation of tissue-specific and housekeeping genes.

In the second work of this thesis, "*Enrichment in intronic enhancers controlling the expression of genes involved in tissue-specific functions and homeostasis"*, we delved into this question by analysing the genomic location of a collection of human ELSs reported by the ENCODE project in a tissue-specific manner. In this work, we reported a correlation between highly shared active enhancers across tissues and their presence in intergenic regions. This seemed to indicate that enhancers that regulate ubiquitously expressed genes tend to be located in intergenic regions. The patterns of enhancer activation allowed us to classify these ELSs in different tissue-specific clusters, which showed a preferential location for introns in the case of highly specialized tissues like muscle and brain, while the majority of common enhancers fall within intergenic regions. The analysis of eQTLs from the GTEx project and enhancer-promoter loop contacts from HiC datasets allowed to identify the potential target genes of these enhancers. This analysis revealed that intronic enhancers tend to regulate genes involved in the specific processes of the tissue where they are active, while intergenic enhancers are more devoted to regulate genes with more basic cellular functions. The regulation of tissue-specific processes seems to be more efficient by those enhancers that are hosted by their target genes. The expression patterns of these tissue-specific interactions identify more remarkably the identity of the tissue and, added to that, the role of these hosting target genes revealed a higher enrichment in tissue-specific functions. On the other hand, enhancers regulating non-host genes appear controlling broader homeostatic processes not only associated with tissue-specific functions but also with basic cellular maintenance processes. This regulatory specificity of intronic enhancers is not exclusive of adult tissues. Our results on

embryonic samples indicate that the genomic location of enhancers is a dynamic feature that experiences a shift towards intronic regions through development. The preferential location of enhancers in intronic regions seems to play an evolutionary advantageous role in terms of regulatory efficiency. The chromatin accessibility of these active enhancers might facilitate the transcription of their hosting target genes and therefore contribute to the establishment of tissue-specific regulatory programs.

The human regulatory genome does not only depend on the interplay of *cis*-regulatory elements in the transcriptional control of gene expression. The emergence of the "*omics*" methodologies by the mid 90's created a new way of doing biology. In this context, the traditional way of understanding gene regulation as unidimensional processes was definitely drowned by the model of regulatory networks and complex systems. In this new paradigm, the usage of genetic information by the genome started to be understood as the result of the cross-talk between multiple dimensions of regulatory control. One of the levels that has arisen during the last years as an essential regulatory layer in the cell is the post-transcriptional level. In this line, the advent of high throughput sequencing technologies has revealed a vast and complex transcriptional landscape populated by a multitude of RNA sequence species that participate in numerous biological processes of the post-transcriptional dimension. One of the most studied classes of post-transcriptional regulatory players are the miRNAs, which constituted a burst of functional innovation in the human lineage. This class of RNAs forms a dense, diverse, temporal and location faceted layer of regulation that presents a potential ability to fine-tune gene expression programs in order to make regulatory networks adaptive to multiple changes in the environment.

In the third part of this thesis, "*Signatures of genetic variation in human microRNAs point to processes of positive selection related to population-specific diseases*", we have focused on the role of this

class of small RNAs in the regulatory adaptations of human populations. In this study we performed a comprehensive description of the diversity profiles of the most recent human miRNA repertoire annotated to date. Firstly, we wanted to delve into the question of what genomic and evolutionary features might drive the nucleotide diversity in the human miRNAs. Our results suggest that miRNA genomic location and evolutionary age are the main factors that contribute to the increase of miRNA diversity. The emergence of new miRNAs in the primate lineage, the presence of transposable elements in their genomic context and their location outside clusters appear as the main contributors of this diversity. As a proxy for conservation, nucleotide diversity also exhibits differential signatures in the different functional regions of miRNAs. The seed region, traditionally considered as the most conserved part of the sequence due to its crucial role in target recognition, outstands in our study as the most diverse among the conserved miRNA regions (loop, mature outside the seed and seed). This higher level of genetic variation indicates the capacity of miRNAs to accept nucleotide changes in this part of the sequence, which potentially reshape their targeting profiles and regulatory behaviour. The analysis of population differentiation scores (F_{st}) supports this hypothesis by showing the highest F_{st} values across population comparisons in the seed and loop regions, being the latter an essential part devoted to the proper folding of the sequence. These highly differentiated miRNAs are therefore candidates of being under positive selection since frequency shifts in different populations might be a response to adaptive processes in their local environments. We show that the top differentiated SNPs are indeed enriched in signals of expression variation (eQTLs), signatures of recent positive selection reported by haplotype-based statistics (iHS and nSL) and associated with human diseases (GWAS). Given the high overlap between miRNA locations and coding gene regions, signatures of positive selection may be, at least in part, the result of gene adaptation and not specific selection signals of miRNAs. Further analysis is needed to assess this result and identify specific

miRNAs signals of selection. We further evaluated the implication of the top differentiated miRNAs in adaptive processes and human diseases. As expected, most of our top 5% candidates present a remarkable shift of their targeting profiles, suggesting massive regulatory changes driven by these miRNAs. Also, these candidates seem to participate in different types of cancers with remarkable differences in population prevalence and glaucoma, which appears associated with the candidate *miR-4707-3p* and presents remarkable susceptibility differences among African and non-African populations.

Limitations and caveats

It is worth mentioning the limitations that this work presents in their methodologies and interpretation of the results. Like in many other genomic scans, the identification of true selection signals in the X chromosome is a rather difficult task. The appropriate association between the genomic signal of a putative selective sweep with the phenotypic effect behind the selection process is the major bottleneck of this kind of "hypothesis free" studies. The usual next step, upon the identification of a candidate region under positive selection, is the experimental validation of the phenotypic effect. The validation of genic signals are normally focused on specific non-synonymous changes, which is the most likely type of variant that generates an appreciable phenotypic effect. However, the identification of the true target of selection (causal variant) would be rather complicated due to the presence of many other variants in the region under selection. The validation of selection signatures in regulatory regions presents additional difficulties. As described in Chapter 1, we performed a luciferase-based assay of the top candidate signals captured in human enhancers. We found a significant differential activity between the ancestral and derived haplotypes in different populations, which indicates the phenotypic effect of the selective sweep at regulatory level. Although the most probable target of enhancers are the closest gene, they may present

other targets located in further genomic locations. Therefore the selection processes behind the change in the enhancer activities remain obscure due to the complexity of the regulatory interactions that a particular enhancer may have, and the extent in which its activity may affect.

The analysis of the role of enhancers in tissue-specific regulatory programs presented in Chapter 2 depends partially on the availability of association datasets between the enhancer and the target gene. In our study we used two different approaches to infer these regulatory interactions: eQTLs, which provides indirect evidence of the genes affected by the enhancers, and the more direct HiC loop contacts, based on the physical interaction between the enhancer and the target promoter. Our results indicate a clear association between intronic enhancers and the regulation of genes implicated in tissue-specific functions. However, for some tissues, we were not able to obtain HiC-based interactions and provide a more detailed implication of enhancers in the regulation of their specific genes. In this line, the statistical power of this study would increase significantly if more interaction datasets were added to the analysis and, therefore, we would provide a more complete picture of the tissue-specific regulatory programs.

The limitations described for Chapter 1 are extensive to the analysis of noncoding regulatory sequences in general and miRNAs in particular. As described in Chapter 3, human miRNAs are predicted to target thousands of genes and form dense networks of regulatory interactions. Therefore, the identification of the regulatory pathways where these population-specific changes are implicated is rather difficult. Adaptive changes in ancient populations are sometimes associated with maladaptations to current environments. Also, due to the complexity of miRNA regulatory networks, the phenotypic advantage that a genetic change confers to the population might be also implicated with other regulatory pathways, generating pathological consequences. However, in this context we could still

identify probable phenotypic consequences of some of these changes, being most of them related to common diseases with population-specific prevalence. The signatures of genetic variation found in our work might likely correspond to processes of positive selection, but also the frequency shifts generated by demographic events like bottlenecks or expansion are likely affecting the interpretation of these results. A comprehensive analysis of not only the miRNA sequence changes, but also the target site variation, would approximate the interpretation of signals of selection to more accurate hypotheses. Additionally, the experimental validation of these changes would identify the true target of selection and narrow the interpretation behind the potential phenotypic effect.

Future perspectives

There is no doubt that the exponential improvement in sequencing and computational technologies during the last two decades has led us to remarkable breakthroughs in the understanding of the evolutionary history of Homo sapiens. However, the current knowledge on the complexity of the human genome and the incompleteness of genetic information in a multitude of populations suggests that we are still dealing with the tip of the iceberg that remains to be fully understood.

In this thesis we have delved into the implication of the X chromosome in processes of positive selection and the role of the regulatory genome at both miRNA and enhancer levels in processes of regulatory adaptation and tissue-specific control. However, a more comprehensive analysis of the noncoding genome in selection studies are required to generate a faithful picture of the evolutionary implication of such regions. The remarkable presence of GWAS and selection signals outside genes suggests that there is still much work to be done in order to disentangle the role of these parts of the genome in traits like complex diseases. The experimental validation of these signals is the true bottleneck in this kind of analysis. In this

line, the development of massive experimental screening methodologies would accelerate the discovery of true targets of positive selection. However, the complexity of regulatory processes would still hinder the proper association with the phenotype behind the adaptation signal.

The expansion of the sample catalogs, the development of artificial intelligence (AI) and the constant increase of computational power are our allies in the task of deciphering the dark corners of our genome. Deep learning techniques are already providing evolutionary models with an unprecedented level of sophistication. The application of these methodologies in the discovery of subtle signals of positive selection, like those left by regulatory adaptations or polygenic selection, would completely change the paradigm of natural selection studies. Added to that, the development of quantum computing would lead us, in a near future, to horizons of knowledge very difficult to predict. In any case, one of the convergent outcomes of this technological progress is the development of personalized medicine. Although significant improvements must be made in the legal, social and ethical aspects of this new paradigm, it is clear that these advances would mean a key step in the evolution of human societies.

V. BIBLIOGRAPHY

Bibliography

Abascal, F., Acosta, R., Addleman, N. J., Adrian, J., Afzal, V., Aken, B., et al. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature 583, 699–710. doi:10.1038/s41586-020-2493-4.

Abbott, J. K., Nordén, A. K., and Hansson, B. (2017). Sex chromosome evolution: Historical insights and future perspectives. Proc. R. Soc. B Biol. Sci. 284. doi:10.1098/rspb.2016.2806.

Agarwal, V., Bell, G. W., Nam, J. W., and Bartel, D. P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. Elife 4. doi:10.7554/eLife.05005.

Arbiza, L., Gottipati, S., Siepel, A., and Keinan, A. (2014). Contrasting X-linked and autosomal diversity across 14 human populations. Am. J. Hum. Genet. 94, 827–844. doi:10.1016/j.ajhg.2014.04.011.

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Bentley, D. R., Chakravarti, A., et al. (2015). A global reference for human genetic variation. Nature 526, 68–74. doi:10.1038/nature15393.

Balaton, B. P., and Brown, C. J. (2016). Escape Artists of the X Chromosome. Trends Genet. 32, 348–359. doi:10.1016/j.tig.2016.03.007.

Balaton, B. P., Cotton, A. M., and Brown, C. J. (2015). Derivation of consensus inactivation status for X-linked genes from genome-wide studies. Biol. Sex Differ. 6. doi:10.1186/s13293-015-0053-7.

Banerji, J., Olson, L., and Schaffner, W. (1983). A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. Cell 33, 729–740. doi:10.1016/0092-8674(83)90015-6.

Banerji, J., Rusconi, S., and Schaffner, W. (1981). Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. Cell 27, 299–308. doi:10.1016/0092-8674(81)90413-X.

Bartel, D. P. (2018). Metazoan MicroRNAs. Cell 173, 20–51. doi:10.1016/j.cell.2018.03.006.

Berezikov, E. (2011). Evolution of microRNA diversity and regulation in animals. Nat. Rev. Genet. 12, 846–860. doi:10.1038/nrg3079.

Bergström, A., Stringer, C., Hajdinjak, M., Scerri, E. M. L., and Skoglund, P. (2021). Origins of modern human ancestry. Nature 590, 229–237. doi:10.1038/s41586-021-03244-5.

Bersaglieri, T., Sabeti, P. C., Patterson, N., Vanderploeg, T., Schaffner, S. F., Drake, J. A., et al. (2004). Genetic signatures of strong recent positive selection at the lactase gene. Am. J. Hum. Genet. 74, 1111–1120. doi:10.1086/421051.

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., et al. (2011). The evolution of gene expression levels in mammalian organs. Nature 478, 343–348. doi:10.1038/nature10532.

Britten, R. J., and Davidson, E. H. (1969). Gene regulation for higher cells: A theory. Science (80-.). 165, 349–357. doi:10.1126/science.165.3891.349.

Bryk, J., Hardouin, E., Pugach, I., Hughes, D., Strotmann, R., Stoneking, M., et al. (2008). Positive selection in East Asians for an EDAR allele that enhances NF-κB activation. PLoS One 3. doi:10.1371/journal.pone.0002209.

Casillas, S., Mulet, R., Villegas-Mirón, P., Hervas, S., Sanz, E., Velasco, D., et al. (2018). PopHuman: The human population genomics browser. Nucleic Acids Res. 46, D1003–D1010. doi:10.1093/nar/gkx943.

Casto, A. M., Li, J. Z., Absher, D., Myers, R., Ramachandran, S., and Feldman, M. W. (2010). Characterization of X-Linked SNP genotypic variation in globally-distributed human populations. Genome Biol. 11. doi:10.1186/gb-2010-11-1-r10.

Chen, K., and Rajewsky, N. (2007). The evolution of gene regulation by transcription factors and microRNAs. Nat. Rev. Genet. 8, 93–103. doi:10.1038/nrg1990.

Chen, K., and Rajewsky, N. (2006). Natural selection on human microRNA binding sites inferred from SNP data. Nat. Genet. 38, 1452–1456. doi:10.1038/ng1910.

Cotter, D. J., Brotman, S. M., and Wilson Sayres, M. A. (2016). Genetic diversity on the human x chromosome does not support a strict pseudoautosomal boundary. Genetics 203, 485–492. doi:10.1534/genetics.114.172692.

Craig Venter, J., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., et al. (2001). The sequence of the human genome. Science (80-.). 291, 1304–1351. doi:10.1126/science.1058040.

CRICK, F. H. (1958). On protein synthesis. Symp. Soc. Exp. Biol. 12, 138–163. Available at:

https://europepmc.org/article/med/13580867 [Accessed May 21, 2021].

Duan, D., Goemans, N., Takeda, S., Mercuri, E., and Aartsma-Rus, A. (2021). Duchenne muscular dystrophy. Nat. Rev. Dis. Prim. 7. doi:10.1038/s41572-021-00248-3.

Dutheil, J. Y., Munch, K., Nam, K., Mailund, T., and Schierup, M. H. (2015). Strong Selective Sweeps on the X Chromosome in the Human-Chimpanzee Ancestor Explain Its Low Divergence. PLoS Genet. 11. doi:10.1371/journal.pgen.1005451.

Eisenberg, E., and Levanon, E. Y. (2013). Human housekeeping genes, revisited. Trends Genet. 29, 569–574. doi:10.1016/j.tig.2013.05.010.

Fan, S., Hansen, M. E. B., Lo, Y., and Tishkoff, S. A. (2016). Going global by adapting local: A review of recent human adaptation. Science (80-.). 354, 54–59. doi:10.1126/science.aaf5098.

Ferrer-Admetlla, A., Liang, M., Korneliussen, T., and Nielsen, R. (2014). On detecting incomplete soft or hard selective sweeps using haplotype structure. Mol. Biol. Evol. 31, 1275–1291. doi:10.1093/molbev/msu077.

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., and Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in caenorhabditis elegans. Nature 391, 806–811. doi:10.1038/35888.

França, G. S., Vibranovski, M. D., and Galante, P. A. F. (2016). Host gene constraints and genomic context impact the expression and evolution of human microRNAs. Nat. Commun. 7. doi:10.1038/ncomms11438.

Fu, W., and Akey, J. M. (2013). Selection and adaptation in the human genome. Annu. Rev. Genomics Hum. Genet. 14, 467–489. doi:10.1146/annurev-genom-091212-153509.

Fu, Y. X., and Li, W. H. (1993). Statistical tests of neutrality of mutations. Genetics 133.

Garud, N. R., Messer, P. W., Buzbas, E. O., and Petrov, D. A. (2015). Recent Selective Sweeps in North American Drosophila melanogaster Show Signatures of Soft Sweeps. PLoS Genet. 11, 1–32. doi:10.1371/journal.pgen.1005004.

Georges, M., Clop, A., Marcq, F., Takeda, H., Pirottin, D., Hiard, S., et al. (2006). Polymorphic microRNA-target interactions: A novel source of phenotypic variation. in Cold Spring Harbor Symposia on Quantitative Biology (Cold Spring Harb Symp Quant Biol), 343–350. doi:10.1101/sqb.2006.71.056.

Georges, M., Coppieters, W., and Charlier, C. (2007). Polymorphic miRNA-mediated gene regulation: contribution to phenotypic variation and disease. Curr. Opin. Genet. Dev. 17, 166–176. doi:10.1016/j.gde.2007.04.005.

Gerbault, P., Liebert, A., Itan, Y., Powell, A., Currat, M., Burger, J., et al. (2011). Evolution of lactase persistence: An example of human niche construction. Philos. Trans. R. Soc. B Biol. Sci. 366, 863–877. doi:10.1098/rstb.2010.0268.

Gilbert, W., and Muller-Hill, B. (1966). ISOLATION OF THE LAC REPRESSOR. Proc. Natl. Acad. Sci. 56, 1891–1898. doi:10.1073/pnas.56.6.1891.

Gittelman, R. M., Hun, E., Ay, F., Madeoy, J., Pennacchio, L., Noble, W. S., et al. (2015). Comprehensive identification and

analysis of human accelerated regulatory DNA. Genome Res. 25, 1245–1255. doi:10.1101/gr.192591.115.

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. Nat. Rev. Genet. 17, 333–351. doi:10.1038/nrg.2016.49.

Grimson, A., Farh, K. K. H., Johnston, W. K., Garrett-Engele, P., Lim, L. P., and Bartel, D. P. (2007). MicroRNA Targeting Specificity in Mammals: Determinants beyond Seed Pairing. Mol. Cell 27, 91–105. doi:10.1016/j.molcel.2007.06.017.

Gu, T. J., Yi, X., Zhao, X. W., Zhao, Y., and Yin, J. Q. (2009). Alu-directed transcriptional regulation of some novel miRNAs. BMC Genomics 10. doi:10.1186/1471-2164-10-563.

Guo, L., Zhao, Y., Zhang, H., Yang, S., and Chen, F. (2014). Integrated evolutionary analysis of human miRNA gene clusters and families implicates evolutionary relationships. Gene 534, 24–32. doi:10.1016/j.gene.2013.10.037.

Hammer, M. F., Woerner, A. E., Mendez, F. L., Watkins, J. C., Cox, M. P., and Wall, J. D. (2010). The ratio of human X chromosome to autosome diversity is positively correlated with genetic distance from genes. Nat. Genet. 42, 830–831. doi:10.1038/ng.651.

Heinz, S., Romanoski, C. E., Benner, C., and Glass, C. K. (2015). The selection and function of cell type-specific enhancers. Nat. Rev. Mol. Cell Biol. 16, 144–154. doi:10.1038/nrm3949.

Iwai, N., and Naraba, H. (2005). Polymorphisms in human pre-miRNAs. Biochem. Biophys. Res. Commun. 331, 1439–1444. doi:10.1016/j.bbrc.2005.04.051.

Iwama, H., Kato, K., Imachi, H., Murao, K., and Masaki, T. (2013). Human microRNAs originated from two periods at accelerated rates in mammalian evolution. Mol. Biol. Evol. 30, 613–626. doi:10.1093/molbev/mss262.

Jacob, F., and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. J. Mol. Biol. 3, 318–356. doi:10.1016/S0022-2836(61)80072-7.

Jacquier, A. (2009). The complex eukaryotic transcriptome: Unexpected pervasive transcription and novel small RNAs. Nat. Rev. Genet. 10, 833–844. doi:10.1038/nrg2683.

Kamm, G. B., Pisciottano, F., Kliger, R., and Franchini, L. F. (2013). The developmental brain gene NPAS3 contains the largest number of accelerated regulatory sequences in the human genome. Mol. Biol. Evol. 30, 1088–1102. doi:10.1093/molbev/mst023.

Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., and Segal, E. (2007). The role of site accessibility in microRNA target recognition. Nat. Genet. 39, 1278–1284. doi:10.1038/ng2135.

King, M. C., and Wilson, A. C. (1975). Evolution at two levels in humans and chimpanzees. Science (80-.). 188, 107–116. doi:10.1126/science.1090005.

Kouprina, N., Mullokandov, M., Rogozin, I. B., Collins, N. K., Solomon, G., Otstot, J., et al. (2004). The SPANX gene family of cancer/testis-specific antigens: Rapid evolution and amplification in African great apes and hominids. Proc. Natl. Acad. Sci. U. S. A. 101, 3077–3082. doi:10.1073/pnas.0308532100.

Kozomara, A., Birgaoanu, M., and Griffiths-Jones, S. (2019). MiRBase: From microRNA sequences to function. Nucleic Acids Res. 47, D155–D162. doi:10.1093/nar/gky1141.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome. Nature 409, 860–921. doi:10.1038/35057062.

Lee, R. C., Feinbaum, R. L., and Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 75, 843–854. doi:10.1016/0092-8674(93)90529-Y.

Levine, M., and Tjian, R. (2003). Transcription regulation and animal diversity. Nature 424, 147–151. doi:10.1038/nature01763.

Lewis, B. P., Burge, C. B., and Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 120, 15–20. doi:10.1016/j.cell.2004.12.035.

Lewis, B. P., Shih, I. H., Jones-Rhoades, M. W., Bartel, D. P., and Burge, C. B. (2003). Prediction of Mammalian MicroRNA Targets. Cell 115, 787–798. doi:10.1016/S0092-8674(03)01018-3.

Li, J., Liu, Y., Xin, X., Kim, T. S., Cabeza, E. A., Ren, J., et al. (2012). Evidence for positive selection on a number of microRNA regulatory interactions during recent human evolution. PLoS Genet. 8. doi:10.1371/journal.pgen.1002578.

Lu, J., and Wu, C. I. (2005). Weak selection revealed by the whole-genome comparison of the X chromosome and autosomes of human and chimpanzee. Proc. Natl. Acad. Sci. U. S. A. 102, 4063–4067. doi:10.1073/pnas.0500436102.

Lu, Z., Carter, A. C., and Chang, H. Y. (2017). Mechanistic insights in X-chromosome inactivation. Philos. Trans. R. Soc. B Biol. Sci. 372. doi:10.1098/rstb.2016.0356.

Lyon, M. F. (1961). Gene action in the X-chromosome of the mouse (mus musculus L.). Nature 190, 372–373. doi:10.1038/190372a0.

McDougall, I., Brown, F. H., and Fleagle, J. G. (2005). Stratigraphic placement and age of modern humans from Kibish, Ethiopia. Nature 433, 733–736. doi:10.1038/nature03258.

Messer, P. W., and Petrov, D. A. (2013). Population genomics of rapid adaptation by soft selective sweeps. Trends Ecol. Evol. 28, 659–669. doi:10.1016/j.tree.2013.08.003.

Miga, K. H., Koren, S., Rhie, A., Vollger, M. R., Gershman, A., Bzikadze, A., et al. (2020). Telomere-to-telomere assembly of a complete human X chromosome. Nature 585, 79–84. doi:10.1038/s41586-020-2547-7.

Minster, R. L., Hawley, N. L., Su, C. T., Sun, G., Kershaw, E. E., Cheng, H., et al. (2016). A thrifty variant in CREBRF strongly influences body mass index in Samoans. Nat. Genet. 48, 1049–1054. doi:10.1038/ng.3620.

Moran, Y., Agron, M., Praher, D., and Technau, U. (2017). The evolutionary origin of plant and animal microRNAs. Nat. Ecol. Evol. 1. doi:10.1038/s41559-016-0027.

Nam, K., Munch, K., Hobolth, A., Dutheil, J. Y., Veeramah, K. R., Woerner, A. E., et al. (2015). Extreme selective sweeps independently targeted the X chromosomes of the great apes. Proc. Natl. Acad. Sci. U. S. A. 112, 6413–6418. doi:10.1073/pnas.1419306112.

Nielsen, R., Akey, J. M., Jakobsson, M., Pritchard, J. K., Tishkoff, S., and Willerslev, E. (2017). Tracing the peopling of the world through genomics. Nature 541, 302–310. doi:10.1038/nature21347.

Pandey, R., Bhattacharya, A., Bhardwaj, V., Jha, V., Mandal, A. K., and Mukerji, M. (2016). Alu-miRNA interactions modulate transcript isoform diversity in stress response and reveal signatures of positive selection. Sci. Rep. 6. doi:10.1038/srep32348.

Park, C., Carrel, L., and Makova, K. D. (2010). Strong purifying selection at genes escaping X chromosome inactivation. Mol. Biol. Evol. 27, 2446–2450. doi:10.1093/molbev/msq143.

Pennisi, E. (2008). Deciphering the genetics of evolution. Science (80-.). 321, 760–763. doi:10.1126/science.321.5890.760.

Petri, R., Brattås, P. L., Sharma, Y., Jonsson, M. E., Pircs, K., Bengzon, J., et al. (2019). LINE-2 transposable elements are a source of functional human microRNAs and target sites. PLoS Genet. 15. doi:10.1371/journal.pgen.1008036.

Piriyapongsa, J., Mariño-Ramírez, L., and Jordan, I. K. (2007). Origin and evolution of human microRNAs from transposable elements. Genetics 176, 1323–1337. doi:10.1534/genetics.107.072553.

Prabhakar, S., Noonan, J. P., Pääbo, S., and Rubin, E. M. (2006). Accelerated evolution of conserved noncoding sequences in humans. Science (80-.). 314, 786. doi:10.1126/science.1130738.

Prescott, S. L., Srinivasan, R., Marchetto, M. C., Grishina, I., Narvaiza, I., Selleri, L., et al. (2015). Enhancer Divergence and cis-Regulatory Evolution in the Human and Chimp Neural Crest. Cell 163, 68–83. doi:10.1016/j.cell.2015.08.036.

Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., et al. (2014). The complete genome sequence of a

Neanderthal from the Altai Mountains. Nature 505, 43–49. doi:10.1038/nature12886.

Quach, H., Barreiro, L. B., Laval, G., Zidane, N., Patin, E., Kidd, K. K., et al. (2009). Signatures of Purifying and Local Positive Selection in Human miRNAs. Am. J. Hum. Genet. 84, 316–327. doi:10.1016/j.ajhg.2009.01.022.

Rawlings-Goss, R. A., Campbell, M. C., and Tishkoff, S. A. (2014). Global population-specific variation in miRNA associated with cancer risk and clinical biomarkers. BMC Med. Genomics 7. doi:10.1186/1755-8794-7-53.

Rees, J. S., Castellano, S., and Andrés, A. M. (2020). The Genomics of Human Local Adaptation. Trends Genet. 36. doi:10.1016/j.tig.2020.03.006.

Rees, J. S., Castellano, S., and Andrés, A. M. (2020). The Genomics of Human Local Adaptation. Trends Genet. 36. doi:10.1016/j.tig.2020.03.006.

Reich, D., Green, R. E., Kircher, M., Krause, J., Patterson, N., Durand, E. Y., et al. (2010). Genetic history of an archaic hominin group from Denisova cave in Siberia. Nature 468, 1053–1060. doi:10.1038/nature09710.

Reilly, S. K., and Noonan, J. P. (2016). Evolution of Gene Regulation in Humans. Annu. Rev. Genomics Hum. Genet. 17, 45–67. doi:10.1146/annurev-genom-090314-045935.

Reinhart, B. J., Slack, F. J., Basson, M., Pasquienelll, A. E., Bettlnger, J. C., Rougvle, A. E., et al. (2000). The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. Nature 403, 901–906. doi:10.1038/35002607.

Riffo-Campos, Á. L., Riquelme, I., and Brebi-Mieville, P. (2016). Tools for sequence-based miRNA target prediction: What to choose? Int. J. Mol. Sci. 17. doi:10.3390/ijms17121987.

Ross, M. T., Grafham, D. V., Coffey, A. J., Scherer, S., McLay, K., Muzny, D., et al. (2005). The DNA sequence of the human X chromosome. Nature 434, 325–337. doi:10.1038/nature03440.

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., et al. (2002). Detecting recent positive selection in the human genome from haplotype structure. Nature 419, 832–837. doi:10.1038/nature01140.

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., et al. (2007). Genome-wide detection and characterization of positive selection in human populations. Nature 449, 913–918. doi:10.1038/nature06250.

Samstein, R. M., Arvey, A., Josefowicz, S. Z., Peng, X., Reynolds, A., Sandstrom, R., et al. (2012). Foxp3 exploits a pre-existent enhancer landscape for regulatory T cell lineage specification. Cell 151, 153–166. doi:10.1016/j.cell.2012.06.053.

Saunders, M. A., Liang, H., and Li, W. H. (2007). Human polymorphism at microRNAs and microRNA target sites. Proc. Natl. Acad. Sci. U. S. A. 104, 3300–3305. doi:10.1073/pnas.0611347104.

Schrider, D. R., Mendes, F. K., Hahn, M. W., and Kern, A. D. (2015). Soft shoulders ahead: Spurious signatures of soft and partial selective sweeps result from linked hard sweeps. Genetics 200, 267–284. doi:10.1534/genetics.115.174912.

Shlyueva, D., Stampfel, G., and Stark, A. (2014). Transcriptional enhancers: From properties to genome-wide predictions. Nat. Rev. Genet. 15, 272–286. doi:10.1038/nrg3682.

Sirugo, G., Williams, S. M., and Tishkoff, S. A. (2019). The Missing Diversity in Human Genetic Studies. Cell 177, 26–31. doi:10.1016/j.cell.2019.02.048.

Skoglund, P., and Jakobsson, M. (2011). Archaic human ancestry in East Asia. Proc. Natl. Acad. Sci. U. S. A. 108, 18301–18306. doi:10.1073/pnas.1108181108.

Spengler, R. M., Oakley, C. K., and Davidson, B. L. (2014). Functional microRNAs and target sites are created by lineage-specific transposition. Hum. Mol. Genet. 23, 1783–1793. doi:10.1093/hmg/ddt569.

Stevenson, B. J., Iseli, C., Panji, S., Zahn-Zabal, M., Hide, W., Old, L. J., et al. (2007). Rapid evolution of cancer/testis genes on the X chromosome. BMC Genomics 8. doi:10.1186/1471-2164-8-129.

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., et al. (2015). An integrated map of structural variation in 2,504 human genomes. Nature 526, 75–81. doi:10.1038/nature15394.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123, 585–595. doi:10.1093/genetics/123.3.585.

Tishkoff, S. A., Reed, F. A., Ranciaro, A., Voight, B. F., Babbitt, C. C., Silverman, J. S., et al. (2007). Convergent adaptation of human lactase persistence in Africa and Europe. Nat. Genet. 39, 31–40. doi:10.1038/ng1946.

Torres, R., Szpiech, Z. A., and Hernandez, R. D. (2018). Human demographic history has amplified the effects of background selection across the genome. PLoS Genet. 14. doi:10.1371/journal.pgen.1007387.

Torruella-Loran, I., Laayouni, H., Dobon, B., Gallego, A., Balcells, I., Garcia-Ramallo, E., et al. (2016). MicroRNA Genetic Variation: From Population Analysis to Functional Implications of Three Allele Variants Associated with Cancer. Hum. Mutat. 37, 1060–1073. doi:10.1002/humu.23045.

Tukiainen, T., Villani, A. C., Yen, A., Rivas, M. A., Marshall, J. L., Satija, R., et al. (2017). Landscape of X chromosome inactivation across human tissues. Nature 550, 244–248. doi:10.1038/nature24265.

Turchin, M. C., Chiang, C. W. K., Palmer, C. D., Sankararaman, S., Reich, D., and Hirschhorn, J. N. (2012). Evidence of widespread selection on standing variation in Europe at height-associated SNPs. Nat. Genet. 44, 1015–1019. doi:10.1038/ng.2368.

Veeramah, K. R., Gutenkunst, R. N., Woerner, A. E., Watkins, J. C., and Hammer, M. F. (2014). Evidence for increased levels of positive and negative selection on the X chromosome versus autosomes in humans. Mol. Biol. Evol. 31, 2267–2282. doi:10.1093/molbev/msu166.

Vernot, B., and Akey, J. M. (2015). Complex history of admixture between modern humans and neanderthals. Am. J. Hum. Genet. 96, 448–453. doi:10.1016/j.ajhg.2015.01.006.

Vicoso, B., and Charlesworth, B. (2006). Evolution on the X chromosome: Unusual patterns and processes. Nat. Rev. Genet. 7, 645–653. doi:10.1038/nrg1914.

Vitti, J. J., Grossman, S. R., and Sabeti, P. C. (2013). Detecting natural selection in genomic data. Annu. Rev. Genet. 47, 97–120. doi:10.1146/annurev-genet-111212-133526.

Voight, B. F., Kudaravalli, S., Wen, X., and Pritchard, J. K. (2006). A map of recent positive selection in the human genome. PLoS Biol. 4, 0446–0458. doi:10.1371/journal.pbio.0040072.

Wang, L., Sinnott-Armstrong, N., Wagschal, A., Wark, A. R., Camporez, J. P., Perry, R. J., et al. (2020). A MicroRNA Linking Human Positive Selection and Metabolic Disorders. Cell 183, 684-701.e14. doi:10.1016/j.cell.2020.09.017.

Wang, Y., Luo, J., Zhang, H., and Lu, J. (2016). MicroRNAs in the Same Clusters Evolve to Coordinately Regulate Functionally Related Genes. Mol. Biol. Evol. 33, 2232–2247. doi:10.1093/molbev/msw089.

Young, R. S., Hayashizaki, Y., Andersson, R., Sandelin, A., Kawaji, H., Itoh, M., et al. (2015). The frequent evolutionary birth and death of functional promoters in mouse and human. Genome Res. 25, 1546–1557. doi:10.1101/gr.190546.115.

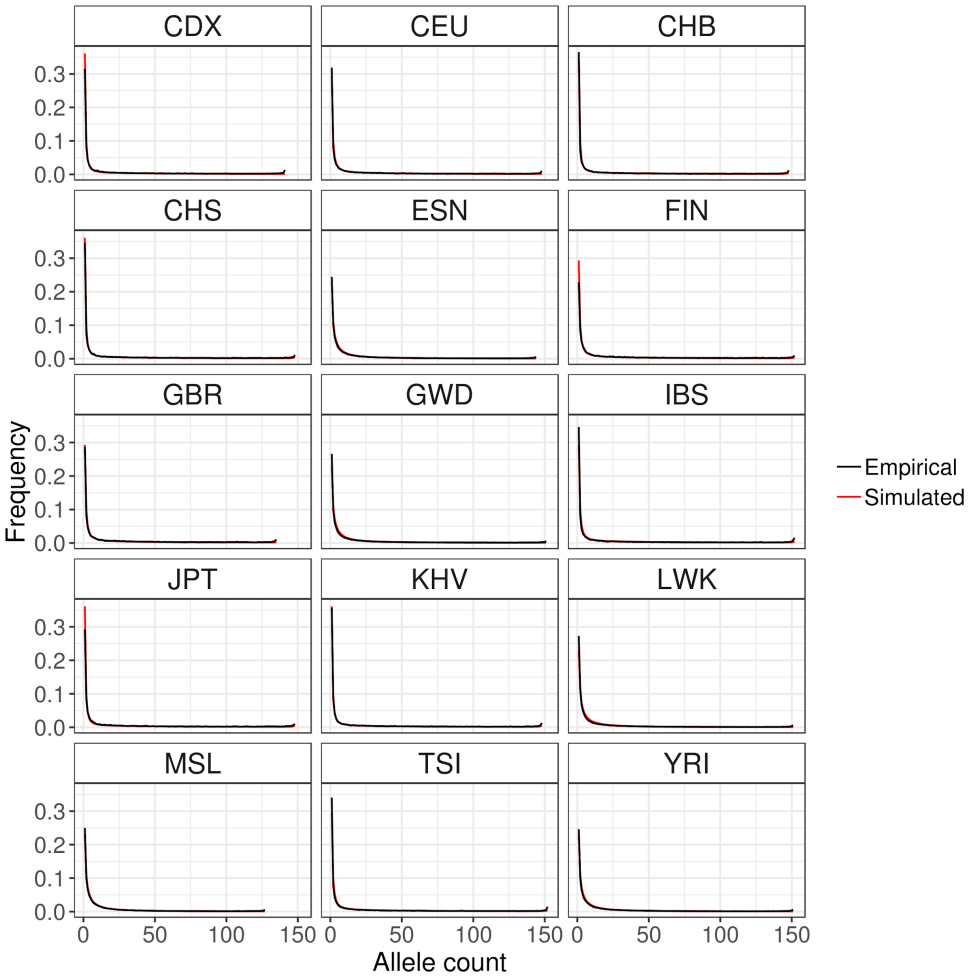
Zaret, K. S., Watts, J., Xu, J., Wandzioch, E., Smale, S. T., and Sekiya, T. (2008). Pioneer factors, genetic competence, and inductive signaling: Programming liver and pancreas progenitors from the endoderm. in Cold Spring Harbor Symposia on Quantitative Biology (Cold Spring Harb Symp Quant Biol), 119–126. doi:10.1101/sqb.2008.73.040.

VI. SUPPLEMENTARY MATERIAL

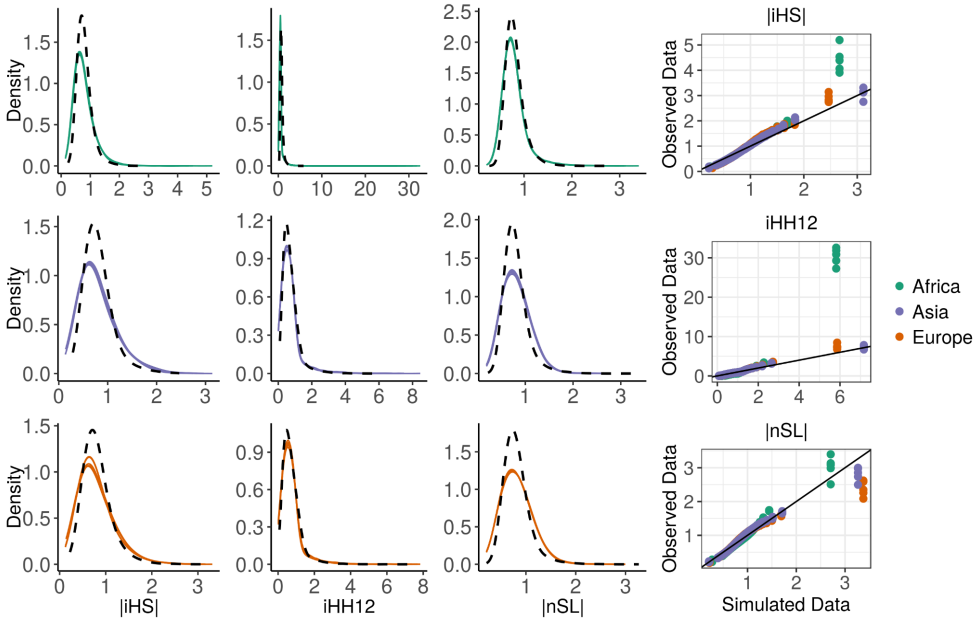
Supplementary Material:

Chromosome X-wide analysis of positive selection in human populations: from common and private signals to selection impact on inactivated genes and enhancers-like signatures

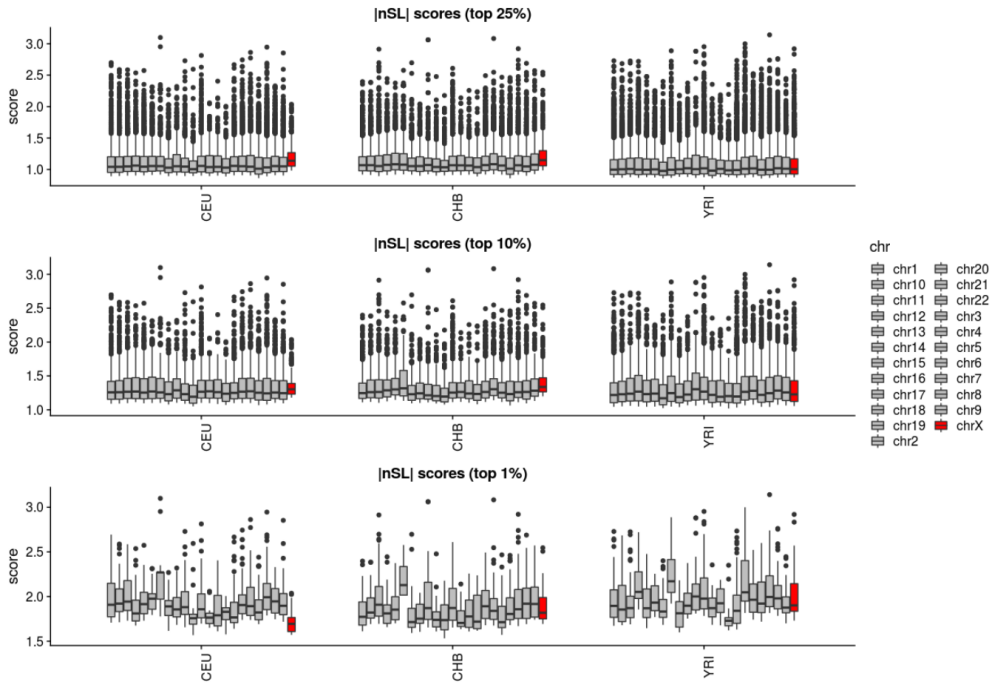
Pablo Villegas-Mirón, Sandra Acosta, Jessica Nye, Jaume Bertranpetit and Hafid Laayouni



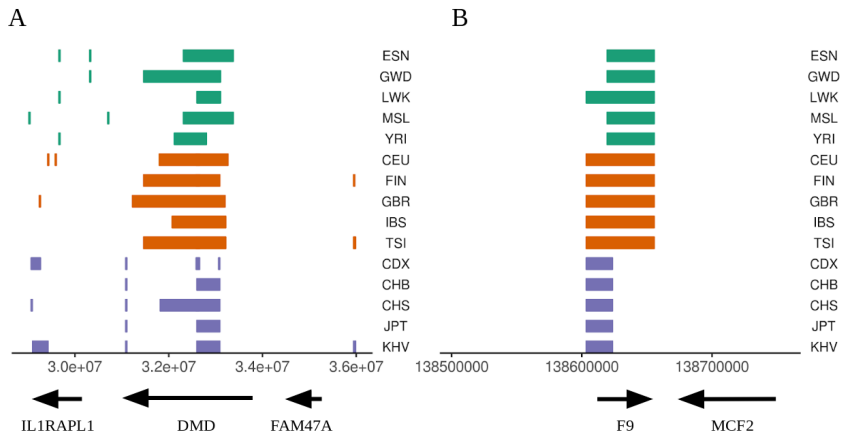
Supplementary Figure 1. Comparison of site frequency spectrums between empirical and simulated data across all populations. Fixed sites have been pruned.



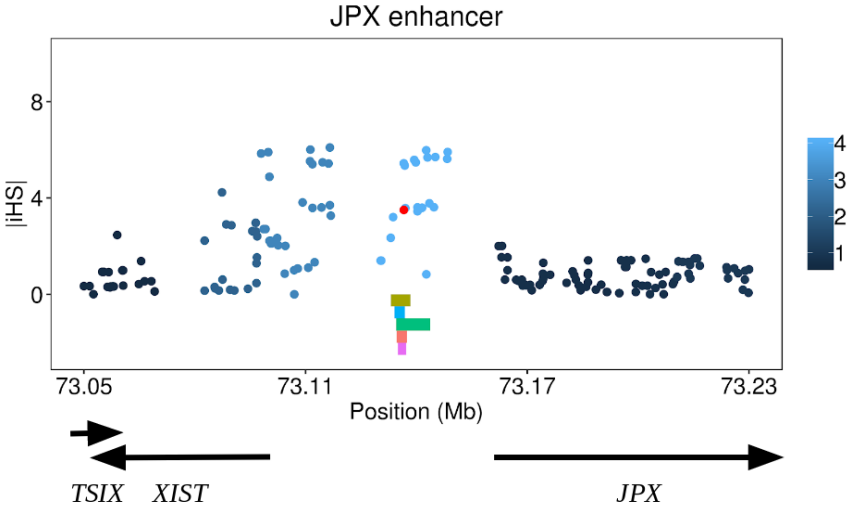
Supplementary Figure 2. A) iHS, iHH12 and nSL distributions (dashed lines as simulated scores) and B) QQ plots of the window-based score distributions in the three geographical groups (Sub-Saharan Africa, Europe, Asia). The QQ plots indicate an overall agreement between the observed and simulated scores. An enrichment of high values in some groups (iHH12 ~ 30) are due to the presence of extreme outliers in the empirical distribution (≥99%), this can be seen in the density plots by the long tail towards positive values.



Supplementary Figure 3. Comparison between nSL extreme tail distributions of autosomes and the X chromosome.



Supplementary Figure 4. Sweeping regions under putative positive selection in the three continental groups in the dystrophin gene (*DMD*; A) and coagulation factor 9 (*F9*; B). The reported sweeps are a result of merging the overlapping windows under positive selection in the 99th in all the statistics used (iHS, iHH12 and nSL).



Supplementary Figure 5. Manhattan plot showing the putative positive selection signal in African populations reported by iHS in the enhancer located at ~23kb from the *JPX* gene. Marked in red is the SNP rs112977454 reported as eQTL by GTEx. Colour bars at the bottom represent the active enhancers in five different cell lines. Legend shows the window-based score where the SNPs belong to.

		iHS		nSL		iHH12	
Group	Population	99th	99.9th	99th	99.9th	99th	99.9th
AFR	ESN	142	25	185	39	172	51
	GWD	128	32	180	42	166	46
	MSL	153	40	188	39	168	51
	LWK	173	28	190	45	171	59
	YRI	143	31	172	35	181	53
EUR	CEU	61	12	23	0	84	9
	FIN	64	13	29	1	70	15
	GBR	83	13	19	4	88	12
	IBS	58	7	28	8	83	13
	TSI	36	8	21	1	81	12
ASI	CDX	53	9	29	4	57	19
	CHB	53	6	32	3	67	11
	CHS	57	4	28	6	55	12
	JPT	55	7	38	1	54	16
	KHV	56	6	31	4	60	14

Supplementary Table 1. Windows under putative positive selection in the extreme simulated 99th and 99.9th percentiles across the 15 populations under study and the three selection statistics accounting for hard and soft sweeps (iHS, iHH12 and nSL).

	iHS			nSL			iHH12		
Chr	YRI	CEU	CHB	YRI	CEU	CHB	YRI	CEU	CHB
1	180	81	74	229	51	23	277	108	96
2	191	98	118	325	51	41	273	130	122
3	228	46	99	205	33	35	224	101	87
4	193	88	83	275	59	29	207	91	98
5	146	68	72	248	52	16	205	110	95
6	119	60	48	230	35	23	239	114	98
7	123	53	55	190	51	24	180	84	70
8	120	39	51	202	38	26	178	81	62
9	69	41	71	161	32	25	132	52	68
10	123	48	54	156	42	19	145	72	70
11	120	45	45	169	42	30	165	56	68
12	86	32	60	196	24	14	170	62	53
13	72	35	34	130	32	19	124	65	53
14	79	34	34	117	30	18	109	50	39
15	66	18	28	85	16	8	66	35	27
16	61	16	30	95	17	8	87	42	36
17	50	27	28	63	17	7	71	40	22
18	65	13	13	95	23	6	92	36	30
19	31	9	9	73	5	6	54	18	19
20	41	9	21	67	6	6	65	26	24
21	31	11	14	45	6	7	42	18	23
22	27	5	6	33	7	6	34	11	12

Supplementary Table 2. Windows under putative positive selection in the extreme simulated 99th percentiles in the human autosomes of the three populations of reference (YRI, CEU and CHB) across the three selection statistics accounting for hard and soft sweeps (iHS, iHH12 and nSL).

Population	Test	GO term	Description	FDR
ESN	iHS,nSL	MF:GO:0004065	sulfuric ester hydrolase activity	0.00074
	iHS,nSL	CC:GO:0004065	endoplasmic reticulum lumen	0.046
	iHH12	CC:GO:0045211	postsynaptic membrane	0.017
	iHH12	CC:GO:1902495	transmembrane transporter complex	0.017
	iHH12	CC:GO:0030425	dendrite	0.0475
	nSL	MF:GO:0042043	neurexin family protein binding	0.0085
GWD	iHS	MF:GO:0008484	sulfuric ester hydrolase activity	0.00000117
	iHS	CC:GO:0005587	collagen type IV trimer	0.032
	iHS	CC:GO:0031968	organelle outer membrane	0.023
	iHS	CC:GO:0044432	endoplasmic reticulum part	0.0093
	iHH12	CC:GO:0005884	actin filament	0.039
	iHH12	CC:GO:1902495	transmembrane transporter complex	0.021
	iHH12	CC:GO:0044463	cell projection part	0.0217
	nSL	MF:GO:0004065	arylsulfatase activity	0.021
	nSL	MF:GO:0042043	neurexin family protein binding	0.0059
MSL	iHS	MF:GO:0008484	sulfuric ester hydrolase activity	0.0000043
	iHH12,nSL	MF:GO:0004065	arylsulfatase activity	0.0197
	iHH12	CC:GO:0097060	synaptic membrane	0.0163
	iHH12	CC:GO:1902495	transmembrane transporter complex	0.0381
	iHH12	CC:GO:0044463	cell projection part	0.0381
LWK	iHS	MF:GO:0008484	sulfuric ester hydrolase activity	0.000653
	iHS	CC:GO:0005788	endoplasmic reticulum lumen	0.00617
	iHH12	CC:GO:0045211	postsynaptic membrane	0.032
	iHH12	CC:GO:1902495	transmembrane transporter complex	0.039
	iHH12	CC:GO:0030425	dendrite	0.039
	nSL	MF:GO:0004065	arylsulfatase activity	0.0169
YRI	iHS	MF:GO:0008484	sulfuric ester hydrolase activity	0.000610
	iHS	CC:GO:0005788	endoplasmic reticulum lumen	0.0116
	iHS	CC:GO:0043005	neuron projection	0.00086
	iHS	CC:GO:0045202	synapse	0.0026
	nSL	MF:GO:0042043	neurexin family protein binding	0.016

Supplementary Table 3A. Significant GO terms of the top 100 genes across all the Sub-saharan African populations in the three selection tests used in the analysis. We consider FDR < 0.05 as significant. In the table, we present the population ID, the tests where the term is reported as significant, the GO term ID, the term description and the corrected FDR value.

Population	Test	GO term	Description	FDR
CEU	iHS,nSL	BP:GO:0097105	presynaptic membrane assembly	0.025
	iHS,nSL	MF:GO:0042043	neurexin family protein binding	0.0122
	iHS,nSL	CC:GO:0098985	asymmetric, glutamatergic, excitatory synapse	0.0179
	iHS,iHH12,nSL	CC:GO:0097060	synaptic membrane	0.0053
	iHS	CC:GO:0030425	dendrite	0.0182
	iHH12,nSL	CC:GO:0044456	synapse part	0.0011
	iHH12	CC:GO:0044463	cell projection part	0.013
	nSL	BP:GO:0042391	regulation of membrane potential	0.019
FIN	iHS	CC:GO:0000775	chromosome, centromeric region	0.034
	iHS,nSL	CC:GO:0030425	dendrite	0.003
	iHS	CC:GO:0043005	neuron projection	0.003
	iHH12	CC:GO:0097060	synaptic membrane	0.0071
	iHH12	CC:GO:0044463	cell projection part	0.0085
	nSL	MF:GO:0042043	neurexin family protein binding	0.0109
	nSL	CC:GO:0098985	asymmetric, glutamatergic, excitatory synapse	0.0232
	nSL	CC:GO:0036020	endolysosome membrane	0.0437
	nSL	BP:GO:0097105	presynaptic membrane assembly	0.0211
GBR	iHS	CC:GO:0030425	dendrite	0.0016
	iHS	CC:GO:0097060	synaptic membrane	0.0183
	iHS	CC:GO:0043005	neuron projection	0.001
	iHH12	CC:GO:0043296	apical junction complex	0.01
	iHH12	CC:GO:0030426	growth cone	0.0155
	iHH12	CC:GO:1902495	transmembrane transporter complex	0.033
	nSL	BP:GO:0097105	presynaptic membrane assembly	0.049
	nSL	CC:GO:0097458	neuron part	0.0041
	nSL	MF:GO:0042043	neurexin family protein binding	0.019
IBS	iHS	CC:GO:0097060	synaptic membrane	0.03
	iHH12	CC:GO:0150034	distal axon	0.012
	iHH12	CC:GO:1902495	transmembrane transporter complex	0.024
	iHH12	CC:GO:0098794	postsynapse	0.0006
TSI	iHS	CC:GO:0030425	dendrite	0.0007
	iHS	CC:GO:0044456	synapse part	0.002
	iHS	CC:GO:0043005	neuron projection	0.0007
	iHH12	CC:GO:1902495	transmembrane transporter complex	0.0177
	iHH12	CC:GO:0030425	dendrite	0.0047
	iHH12	CC:GO:0045211	postsynaptic membrane	0.002
	nSL	CC:GO:0097458	neuron part	0.016

Supplementary Table 3B. Significant GO terms of the top 100 genes across all the European populations in the three selection tests used in the analysis. We consider FDR < 0.05 as significant. In the table, we present the population ID, the tests where the term is reported as significant, the GO term ID, the term description and the corrected FDR value.

Population	Test	GO term	Description	FDR
CDX	iHS,iHH12,nSL	BP:GO:0097105	presynaptic membrane assembly	0.03
	iHS,iHH12,nSL	CC:GO:0005587	collagen type IV trimer	0.029
	iHS,iHH12,nSL	CC:GO:0098985	asymmetric, glutamatergic, excitatory synapse	0.029
	iHS	CC:GO:0098794	postsynapse	0.026
	iHH12,nSL	CC:GO:0045202	synapse	0.025
CHB	iHH12,nSL	BP:GO:0097105	presynaptic membrane assembly	0.032
	iHH12	CC:GO:0098982	GABA-ergic synapse	0.034
	iHH12	CC:GO:0005794	Golgi apparatus	0.034
CHS	iHS,iHH12,nSL	BP:GO:0097105	presynaptic membrane assembly	0.0497
	iHS,nSL	CC:GO:0098794	postsynapse	0.0296
	iHS	CC:GO:0106018	phosphatidylinositol-3,5-bisphosphate phosphatase activity	0.0296
	iHH12	CC:GO:0098985	asymmetric, glutamatergic, excitatory synapse	0.0197
	iHH12	CC:GO:0030425	dendrite	0.028
	iHH12	CC:GO:0044456	synapse part	0.011
JPT	iHS,iHH12	CC:GO:0005587	collagen type IV trimer	0.049
	iHS	CC:GO:0099086	synaptonemal structure	0.049
	iHS,iHH12	CC:GO:0005794	Golgi apparatus	0.049
	iHH12,nSL	BP:GO:0097105	presynaptic membrane assembly	0.0342
	iHH12	CC:GO:0098985	asymmetric, glutamatergic, excitatory synapse	0.023
	iHH12	CC:GO:0098794	postsynapse	0.021
	nSL	MF:GO:0042043	neurexin family protein binding	0.0188
KHV	iHS,iHH12,nSL	BP:GO:0097105	presynaptic membrane assembly	0.036
	iHS,nSL	CC:GO:0005587	collagen type IV trimer	0.036
	iHS,nSL	CC:GO:0098985	asymmetric, glutamatergic, excitatory synapse	0.036
	iHS	CC:GO:0045211	postsynaptic membrane	0.036
	nSL	CC:GO:0098794	postsynapse	0.029

Supplementary Table 3C. Significant GO terms of the top 100 genes across all the Asian populations in the three selection tests used in the analysis. We consider FDR < 0.05 as significant. In the table, we present the population ID, the tests where the term is reported as significant, the GO term ID, the term description and the corrected FDR value.

		iHS			iHH12			nSL		
		Selected	Not Sel.	Total	Selected	Not Sel.	Total	Selected	Not Sel.	Total
95th	Escape	31	28	59	17	42	59	18	41	59
	Inactive	133	248	381	68	313	381	136	245	381
	Total	164	276	440	85	355	440	154	286	440
99th	Escape	18	41	59	10	49	59	9	50	59
	Inactive	44	337	381	34	347	381	50	331	381
	Total	62	378	440	44	396	440	59	381	440
99.9th	Escape	8	51	59	3	56	59	4	55	59
	Inactive	14	367	381	10	371	381	12	369	381
	Total	22	418	440	13	427	440	16	424	440

Supplementary Table 4A. Contingency tables of escape genes under selection reported by the three selection statistics across three extreme percentiles (95th, 99th and 99.9th). Two categories were used: escape/inactive and selected/non-selected.

	iHS			iHH12			nSL		
	Fisher's p	O.R	C.I.(0.95)	Fisher's p	O.R.	C.I.(0.95)	Fisher's p	O.R.	C.I.(0.95)
95th	0.01	2.06	1.14-3.73	0.05	1.86	0.93-3.57	0.46	0.79	0.41-1.47
99th	0.0003	3.35	1.66-6.59	0.06	2.07	0.86-4.64	0.68	1.19	0.48-2.64
99.9th	0.004	4.09	1.41-11.07	0.40	1.98	0.34-8.02	0.24	2.23	0.50-7.70

Supplementary Table 4B. Fisher's tests applied to the contingency tables. iHS reports significant p-values across the three extreme percentiles with increasing odds ratios (OR). iHH12 and nSL do not show significant enrichment in escape genes, however the odds are in line with those in iHS in five out of the six comparisons, suggesting the presence of selection but with lack of significance probably due to a sample effect.

	AFR	EUR	ASI
95th	AP1S2, ARSD, ARSE, ARSF, ARSH, CDK16, FAM9C, HS6ST2, HTR2C, JPX, KDM6A, MAGEC3, MAOA, MED14, MSL3, MXRA5, NR0B1, OFD1, PCDH19, PNPLA4, PRKX, STS, TMEM27, UBA1, ZCCHC16, ZRSR2	ARSF, GYG2, HS6ST2, HTR2C, MAGEC3, STS, TAF7L, TMEM27, USP9X, ZCCHC16, ZFX	FUNDC1, KDM6A, STS, USP9X
99th	ARSE, ARSF, ARSH, CDK16, FAM9C, HS6ST2, HTR2C, KDM6A, MAGEC3, MAGEC3, MED14, MXRA5, OFD1, STS, TMEM27, UBA1,	HS6ST2, STS, USP9X, ZCCHC16	FUNDC1,KDM6A, STS
99th	ARSF, ARSH, CDK16, HTR2C, KDM6A, STS, UBA1	ZCCHC16	-

Supplementary Table 4C. Escape genes reported by iHS as being under positive selection in each continental group across the extreme percentiles.

		iHS				iHH12				nSL			
Group	Population	Int	I	E	D	Int	I	E	D	Int	I	E	D
AFR	ESN	18	5	0	0	30	16	1	0	17	6	0	0
	GWD	18	11	1	0	34	30	2	0	19	9	0	1
	MSL	18	19	0	0	39	34	3	0	17	9	0	0
	LWK	32	15	0	0	56	35	2	0	18	10	0	0
	YRI	14	10	1	0	38	18	1	0	18	7	0	0
EUR	CEU	3	1	0	0	12	3	0	0	0	0	0	0
	FIN	1	4	1	0	11	6	0	0	0	0	0	0
	GBR	3	1	0	0	19	2	0	0	4	1	0	0
	IBS	2	3	0	0	9	7	0	0	5	1	0	0
	TSI	3	5	0	0	6	7	0	0	0	1	0	0
ASI	CDX	5	1	0	0	17	4	0	0	1	1	0	0
	CHB	1	0	0	0	15	0	0	0	4	0	0	0
	CHS	4	0	0	0	16	0	0	0	4	0	0	0
	JPT	1	0	0	0	16	1	0	0	0	0	0	0
	KHV	3	2	0	0	17	3	0	0	0	0	0	0

Supplementary Table 5. SNPs with a selection score within the 1% extreme and with CADD score ≥ 10 in the 99.9th percentile across all populations (Intergenic (Int), Intronic (I), Exonic (E), Downstream (D).

			RegulomeDB scores						
Test	Region	n SNPs	2	3	4	5	6	ENCODE elements	OR
iHS	Intergenic	31 (0.1)	-	-	1	1	9	35%	1.08
	Genic	286 (0.9)	-	-	6	16	74	33%	
iHH12	Intergenic	35 (0.14)	-	-	1	6	11	51%	1.32
	Genic	212 (0.85)	2	2	8	25	57	44%	
nSL	Intergenic	32 (0.12)	-	-	-	3	10	40%	1.23
	Genic	230 (0.88)	1	4	4	9	64	35%	

Supplementary Table 6A. RegulomeDB annotation of the 99th percentile genic windows with intergenic overlap and the odds ratio (OR) between genic and intergenic SNPs within functional elements. All populations are considered.

Test	Region	n SNPs	RegulomeDB scores					ENCODE elements	OR
			2	3	4	5	6		
iHS	Intergenic	4 (0.05)	-	-	-	-	1	25%	0.77
	Genic	83 (0.95)	-	-	3	3	19	30%	
iHH12	Intergenic	8 (0.08)	-	-	-	2	5	87%	8.24
	Genic	98 (0.92)	1	1	2	15	26	45%	
nSL	Intergenic	10 (0.16)	-	-	-	1	5	60%	2.11
	Genic	41 (0.84)	1	-	2	1	13	41%	

Supplementary Table 6B. RegulomeDB annotation of the 99th percentile genic windows with intergenic overlap when considering extreme scoring SNPs (per-SNP 1% extreme tail). iHH12 and nSL show a significant OR increment in comparison with iHS.

Test	Group	99th				99.9th			
		Non-ovlp	Ovlp.	%	OR	Non-ovlp	Ovlp.	%	OR
iHS	AFR	194	3	1.52	0.46	47	3	6.00	1.93
	EUR	116	1	0.86	0.26	23	0	0	0
	ASI	111	7	5.93	1.93	17	0	0	0
iHH12	AFR	128	17	11.72	4.31	43	3	6.52	2.11
	EUR	77	4	4.94	1.57	17	2	10.52	3.57
	ASI	79	8	9.20	3.14	22	1	4.35	1.37
nSL	AFR	283	4	1.39	0.41	68	2	2.86	0.88
	EUR	52	1	1.89	0.58	8	1	11.11	3.77
	ASI	77	2	2.53	0.78	8	0	0	0

Supplementary Table 7. Overlapping and non-overlapping intergenic windows under putative positive selection on enhancer regions reported by HACER in any cell line (see Methods) across the three continental groups. Odds ratio (OR) of intergenic and overlapping windows shows a significant increment mainly in iHH12 across all populations.

		Observed selection on target gene		
		Y	N	Total
Observed selection on enhancer	Y	218	350	568
	N	167	395	562
	Total	385	745	1130

		Expected selection on target gene		
		Y	N	Total
Expected selection on enhancer	Y	193	375	568
	N	192	370	562
	Total	385	745	1130

Supplementary Table 8. Contingency tables of both observed and expected pairs of enhancer/target-gene in the following categories: Selected enhancer and selected gene (YY), Selected enhancer and non-selected gene (YN), Non-selected enhancer and selected gene (NY), Non-selected enhancer and non-selected gene (NN). A Chi square test is applied to study the dependency of both variables (Chi

sqr value = 9.44; p-value = 0.0021).

Chr	Start	End	Length	Test	Population	Closest gene
X	40238664	40241510	2846	iHH12	CEU,GBR,IBS	ATP6AP2
X	45179990	45196717	16727	iHH12	FIN,TSI	KDM6A(*)
X	53740262	53744843	4581	nSL	GWD	HUWE1(**)
X	73135561	73145161	9600	iHS	ESN,GWD,MSL,LWK,YRI	JPX
X	109017803	109018393	590	iHS,nSL	YRI	ACSL4(**)
X	123351438	123353650	2212	iHH12	GWD,MSL,YRI	SH2D1A(**)

Supplementary Table 9. Top enhancer regions under putative positive selection (99.9th percentile). The genes marked as "**" are found under selection in sequence in the 99.9th percentile and in the same continental group, the genes marked as "*" are found under selection as well but in a different continental group.

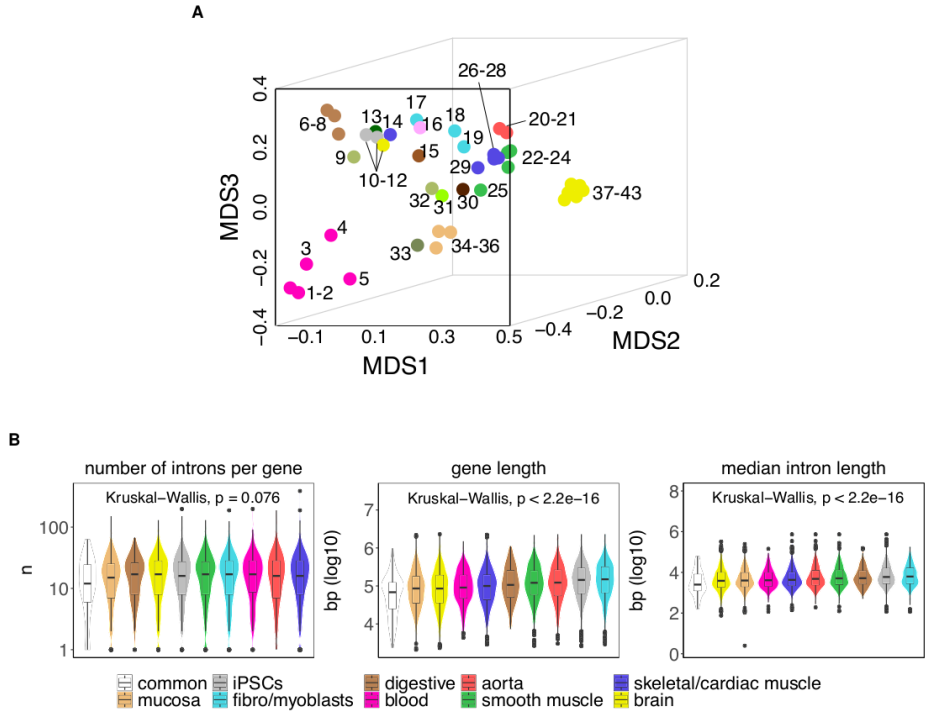
Supplementary File 1 is deposited in the Biorxiv repository associated to the preprint:

Villegas-Mirón P, Acosta S, Nye J, Bertranpetit J, Laayouni H. 2021. Chromosome X-wide analysis of positive selection in human populations: from common and private signals to selection impact on inactivated genes and enhancers-like signatures. bioRxiv doi: BIORXIV/2021/445399

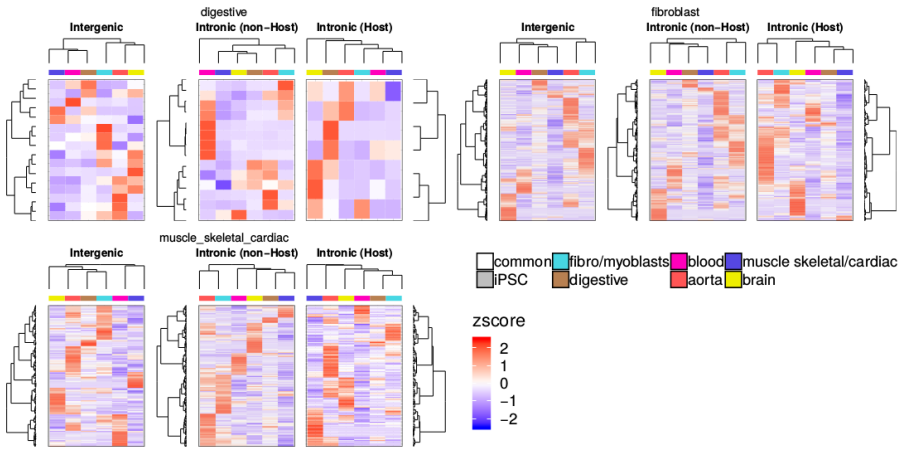
Supplementary Material:

Enrichment in intronic enhancers controlling the expression of genes involved in tissue-specific functions and homeostasis

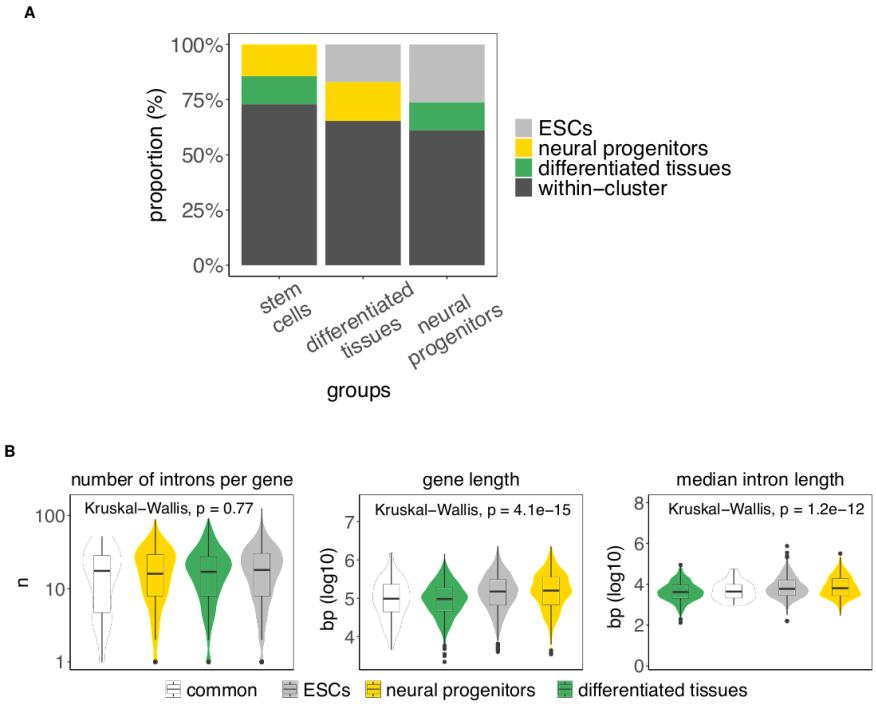
Beatrice Borsari, Pablo Villegas-Mirón, Silvia Perez-Lluch, Isabel Turpin, Hafid Laayouni, Alba Segarra-Casas, Jaume Bertranpetit, Roderic Guigo, Sandra Acosta



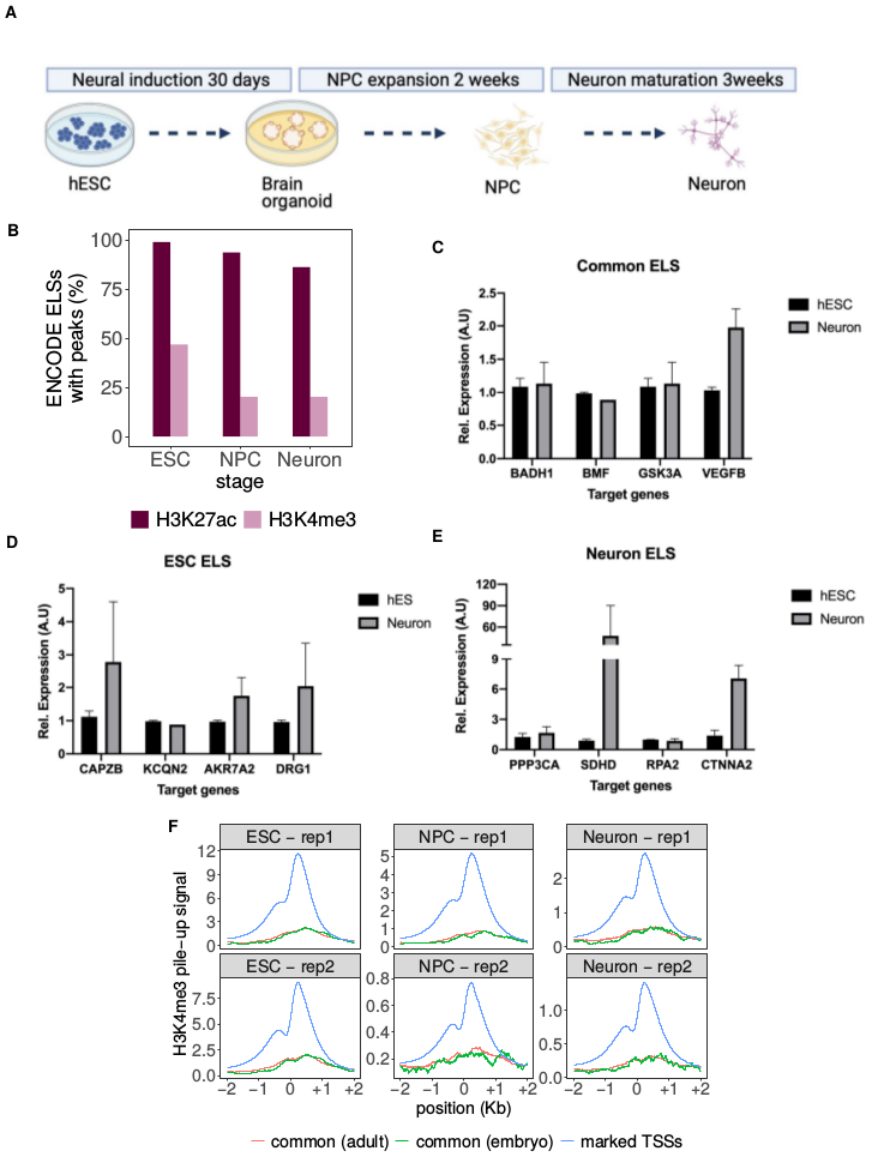
Supplementary Fig. 1. **A**: Multidimensional scaling (MDS) representation of the dissimilarities among the 43 human adult samples based on the pattern of activity of ELS-cCREs. The binary distance between a given pair of samples was computed considering presence/absence vectors of the 991,173 ELS cCREs. The correspondence between samples and numbers is reported in Supplementary Table 1 in Supplementary File.pdf. **B**: Features of genes hosting intronic ELSs in each cluster of adult samples: (1) number of introns per hosting gene, (2) length of hosting gene, (3) median intron length per hosting gene



Supplementary Fig. 2. Z-score normalized median gene expression across GTEx tissue categories of the HiC-ELSs target genes of the intergenic and intronic HiC-ELSs in digestive, fibroblast and muscle skeletal/cardiac tissues. Intronic HiC-ELSs are divided into those that target their host gene (Host) and those that target a gene outside their hosting region (non-Host). Dendrograms show the hierarchical clustering of target genes (rows) and GTEx tissue categories (columns).

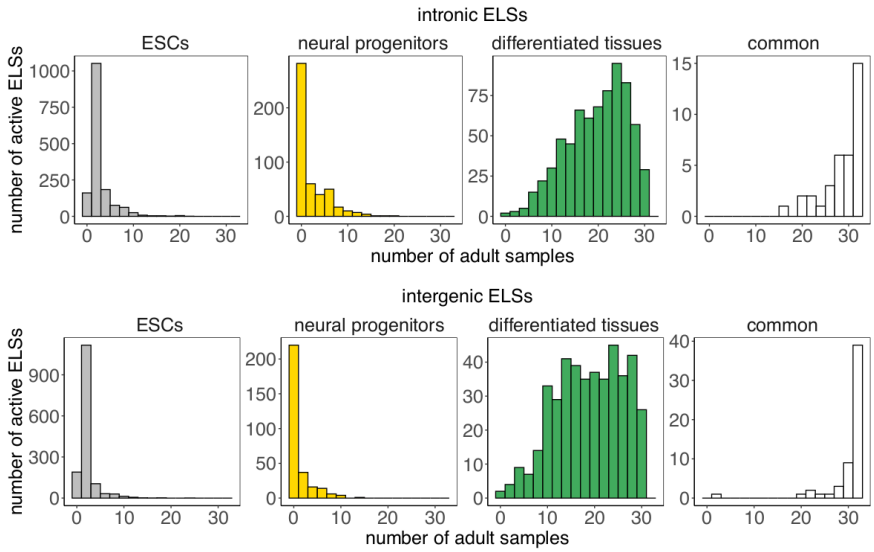


Supplementary Fig. 3. **A**: Group-specific ELSs in embryonic samples (analogous to Fig 1D). The barplot represents the type of outer samples observed within sets of ESCs-, differentiated tissues- and neural progenitors-specific ELSs. **B**: Features of genes hosting either common or specific intronic ELSs identified in embryonic samples (analogous to Supplementary Fig. 1B in Supplementary File.pdf): (1) number of introns per hosting gene, (2) length of hosting gene, (3) median intron length per hosting gene.



Supplementary Fig. 4. **A**: Scheme depicting the differentiation protocol of the ESC into NPC and neurons. **B**: Overlap between ENCODE common ELSs and H3K27ac, H3K4me3 common ELS from ESC, NPC and neurons ChIP-seq. **C-E**: Gene expression analysis in hESC and ESC-derived neurons of genes targeted by ENCODE and ChIP-seq overlapped ELSs in common (C), ESC (D) and neurons (E). Relative quantification was performed against hESC gene expression values and the reference gene was ACTB. This analysis was performed in triplicates. **F**: Pile-up signal of the H3K4me3 for common ELS in ESC, NPC and Neurons in each of the ChIP-seq replicate showing that signal is more intense in the promoters (blue), and ELSs overlapped with adult (orange) and embryo (green)

common ENCODE ELSs. In the TSS the signal is reduced beyond the 2kb distance used as a filter for the ChIP-seq samples, as reflected by the diminished signal detected in the analyzed ELSs, suggesting low promoter activity in the selected ELSs.



Supplementary Fig. 5. Overlap of intronic ELSs between embryonic and adult tissues. Stem Cell (ESCs) and Neural progenitor intronic ELSs are not present in any or very little adult samples (60 adult samples) independently if they are intronic or intergenic. Instead most of the embryonic common ELSs, especially those intergenic, are also active in adult tissues. Intermediate distribution is observed from differentiated embryonic tissues.

	Biosample Term Name	Biosample Type	Samples' Cluster	ENCODE File ID
1	natural killer cell	primary cell	blood	ENCFF529UWB
2	T cell	primary cell	blood	ENCFF098NHL
3	B cell	primary cell	blood	ENCFF379TAE
4	CD14-positive monocyte	primary cell	blood	ENCFF967MJU
5	peripheral blood mononuclear cell	primary cell	blood	ENCFF509DPX
6	pancreas	tissue	digestive	ENCFF681HOL
7	body of pancreas	tissue	digestive	ENCFF768JUC
8	stomach	tissue	digestive	ENCFF992HIZ
9	right lobe of liver	tissue	-	ENCFF476MEG
10	iPS-18a	cell line	iPSC	ENCFF920QRH
11	iPS-20b	cell line	iPSC	ENCFF231KWX
12	bipolar neuron	in vitro differentiated cells	-	ENCFF045GKW
13	thyroid gland	tissue	-	ENCFF296SZK
14	gastrocnemius medialis	tissue	-	ENCFF322RAX
15	endocrine pancreas	tissue	-	ENCFF055CJM
16	ovary	tissue	-	ENCFF586NXH
17	myotube	in vitro differentiated cells	fibro/myoblasts	ENCFF120MMC
18	skeletal muscle myoblast	primary cell	fibro/myoblasts	ENCFF037UZZ
19	fibroblast of lung	primary cell	fibro/myoblasts	ENCFF495RTY
20	aorta	tissue	aorta	ENCFF178GDW
21	thoracic aorta	tissue	aorta	ENCFF257XAQ
22	stomach smooth muscle	tissue	sm muscle	ENCFF726JTT
23	rectal smooth muscle tissue	tissue	sm muscle	ENCFF093MDL
24	vagina	tissue	sm muscle	ENCFF904XYE
25	muscle layer of duodenum	tissue	sm muscle	ENCFF862BGI
26	gastrocnemius medialis	tissue	sk/c muscle	ENCFF863OGG
27	right cardiac atrium	tissue	sk/c muscle	ENCFF278RUJ
28	skeletal muscle tissue	tissue	sk/c muscle	ENCFF311MNY
29	subcutaneous abdominal adipose tissue	tissue	sk/c muscle	ENCFF725QLM
30	esophagus	tissue	-	ENCFF442HYL
31	lung	tissue	-	ENCFF598QTT
32	liver	tissue	-	ENCFF645PQQ
33	spleen	tissue	-	ENCFF821ESA
34	mucosa of rectum	tissue	mucosa	ENCFF759YFL
35	mucosa of rectum	tissue	mucosa	ENCFF403IPC
36	colonic mucosa	tissue	mucosa	ENCFF867TJN
37	middle frontal area 46	tissue	brain	ENCFF070EXF
38	caudate nucleus	tissue	brain	ENCFF508GKP
39	angular gyrus	tissue	brain	ENCFF942KAC
40	layer of hippocampus	tissue	brain	ENCFF159NZA
41	substantia nigra	tissue	brain	ENCFF233VRB
42	temporal lobe	tissue	brain	ENCFF810IQU
43	cingulate gyrus	tissue	brain	ENCFF494WCN

Supplementary Table 1. ENCODE catalogues of cell type-specific candidate cis-Regulatory Elements (cCREs) for 43 human adult samples. The accession number (ENCODE File ID) allows to uniquely identify the catalogue on the ENCODE portal (https://www.encodeproject.org/). The color palette was inspired by the Genotype Tissue Expression (GTEx) Project.

Samples	Tissue-specific ELSs
mucosa	6,205
blood	750
iPSCs	10,966
fibro/myoblasts	2,207
digestive	302
aorta	6,231
smooth muscle	2,825
skeletal/cardiac muscle	5,467
brain	13,054

Samples	Common ELSs
all	555

Supplementary Table 2. [upper panel] Number of ELSs that are specific to each of the 9 clusters of 33 selected human adult samples. Tissue-specific ELSs are those active in 100% (iPSC, fibro/myoblasts, digestive, mucosa and aorta) or ≥ 80% (all other clusters) of the samples within a cluster. In addition, they are active in 0 (iPSC, fibro/myoblasts, digestive, mucosa and aorta) or at most 1 (all other clusters) outer sample (i.e. a sample that does not belong to the considered cluster). [lower panel] Number of ELSs active in ≥ 95% (i.e. n = 31) of the 33 selected human adult samples (common ELSs).

Genomic location	Tissue cluster	FDR	Odds ratio	Confidence interval
intronic	mucosa	2.0e-07	1.62	1.35-1.94
	iPSCs	7.6e-14	1.96	1.64-2.35
	fibro/myoblasts	3.2e-13	2.05	1.68-2.49
	digestive	4.7e-07	2.11	1.58-2.84
	blood	1.7e-14	2.44	1.93-3.07
	aorta	4.7e-29	2.76	2.3-3.32
	sm muscle	3.7e-37	3.38	2.79-4.11
	sk/c muscle	1.5e-49	3.89	3.23-4.69
	brain	1.8e-66	4.66	3.9-5.58
exonic	mucosa	9.9e-06	0.30	0.19-0.51
	iPSCs	2.0e-06	0.28	0.18-0.46
	fibro/myoblasts	4.0e-08	0.15	0.07-0.3
	digestive	2.7e-03	0.15	0.02-0.63
	blood	1.8e-02	0.44	0.21-0.9
	aorta	2.1e-04	0.38	0.24-0.63
	sm muscle	3.5e-04	0.37	0.21-0.64
	sk/c muscle	2.5e-04	0.39	0.24-0.65
	brain	3.0e-02	0.61	0.4-0.99
intergenic	mucosa	8.3e-05	0.70	0.58-0.84
	iPSCs	9.0e-10	0.58	0.48-0.69
	fibro/myoblasts	6.2e-09	0.57	0.47-0.69
	digestive	4.3e-05	0.55	0.41-0.73
	blood	6.6e-12	0.45	0.36-0.57
	aorta	6.0e-24	0.40	0.33-0.48
	sm muscle	1.1e-31	0.33	0.27-0.4
	sk/c muscle	2.7e-43	0.28	0.23-0.34
	brain	1.6e-63	0.22	0.19-0.26

Supplementary Table 3. For each cluster of samples we assessed, with Fisher's exact test, significant differences in the proportions of common vs tissue-specific ELSs that overlap intronic, exonic and intergenic regions. P value (FDR-corrected), odds ratio and confidence interval are reported for each test.

Group	Genes ∩ ELSs			
	Introns	Exons	Both	Total
mucosa	1245 (82.56%)	51 (3.38%)	212 (14.06%)	1508
blood	335 (85.24%)	14 (3.56%)	44 (11.2%)	393
iPSCs	1910 (84.03%)	59 (2.6%)	304 (13.37%)	2273
fibro/myoblasts	749 (86.89%)	15 (1.74%)	98 (11.37%)	862
digestive	129 (90.21%)	3 (2.1%)	11 (7.69%)	143
aorta	1058 (79.31%)	47 (3.52%)	229 (17.17%)	1334
smooth muscle	656 (81.59%)	29 (3.61%)	119 (14.8%)	804
skeletal/cardiac muscle	1298 (80.82%)	49 (3.05%)	259 (16.13%)	1606
brain	1523 (64.51%)	145 (6.14%)	693 (29.35%)	2361
common	144 (83.24%)	14 (8.09%)	15 (8.67%)	173

Supplementary Table 4. Number of genes whose introns and/or exons intersect tissue-specific and common ELSs identified in adult samples.

Tissue	GO term	Description
Aorta	GO:0031589	Cell-substrate adhesion
	GO:0043062	Extracellular structure organization
	GO:2000147	Positive regulation of cell motility
	GO:0043087	Regulation of GTPase activity
	GO:0061564	Axon development
Blood	GO:0042110	T cell activation
	GO:0051056	Regulation of small GTPase mediated signal transduction
	GO:0002764	Immune response-regulating signaling pathway
	GO:0002521	Leukocyte differentiation
	GO:0050900	Leukocyte migration
Brain	GO:0061564	Axon development
	GO:0050808	Synapse organization
	GO:0022604	Regulation of cell morphogenesis
	GO:0099177	Regulation of trans-synaptic signaling
	GO:0098742	Cell-cell adhesion via plasma-membrane adhesion molecules
Muscle skeletal/cardiac	GO:0003012	Muscle system process
	GO:0042692	Muscle cell differentiation
	GO:0007517	Muscle organ development
	GO:0051056	Regulation of small GTPase mediated signal transduction
	GO:0034330	Cell junction organization
Muscle smooth	GO:0043062	Extracellular structure organization
	GO:0003012	Muscle system process
	GO:0019932	Second-messenger-mediated signaling
	GO:0003013	Circulatory system process
	GO:0099177	Regulation of trans-synaptic signaling
Mucosa	GO:0051056	Regulation of small GTPase mediated signal transduction
	GO:0038127	ERBB signaling pathway
	GO:0034330	Cell junction organization
	GO:0043087	Regulation of GTPase activity
	GO:0032970	Regulation of actin filament-based process
Digestive	-	-
Fibro/myoblasts	GO:0043087	Regulation of GTPase activity
	GO:0010975	Regulation of neuron projection development
	GO:0051056	Regulation of small GTPase mediated signal transduction
	GO:0090130	Tissue migration
	GO:2000147	Positive regulation of cell motility
iPSCs	GO:0010975	Regulation of neuron projection development
	GO:0098742	Cell-cell adhesion via plasma-membrane adhesion molecules
	GO:0022604	Regulation of cell morphogenesis
	GO:0061564	Axon development
	GO:0050808	Synapse organization
Common	GO:0034330	Cell junction organization
	GO:1903706	Regulation of hemopoiesis
	GO:1901652	Response to peptide
	GO:0002521	Leukocyte differentiation
	GO:0035264	Multicellular organism growth

Supplementary Table 5. Significantly enriched GO terms (Biological Process) associated with genes hosting intronic ELSs identified in adult samples. Only the top five enriched terms are shown.

Tissue	Hosting	GO term	Description
Aorta	Intronic	-	-
	Intergenic	MF:0004499	N,N-dimethylaniline monooxygenase activity
		MF:0004024	alcohol dehydrogenase activity, zinc-dependent
		MF:0004022	alcohol dehydrogenase (NAD) activity
Blood	Intronic	-	-
	Intergenic	BP:0019886	antigen processing and presentation of exogenous peptide antigen via MHC class II
		BP:0060333	interferon-gamma-mediated signaling pathway
		BP:0050852	T cell receptor signaling pathway
Brain	Intronic	BP:0000226	microtubule cytoskeleton organization
		BP:0030030	cell projection organization
		BP:0120036	plasma membrane bounded cell projection organization
	Intergenic	CC:0033267	axon part
		CC:0005815	microtubule organizing center
		CC:0015630	microtubule cytoskeleton
Fibro/myoblasts	Intronic	CC:0015629	actin cytoskeleton
	Intergenic	-	-
Digestive	Intronic	-	-
	Intergenic	MF:0035591	signaling adaptor activity
Mucosa	Intronic	BP:0044281	small molecule metabolic process
		MF:0016289	CoA hydrolase activity
		MF:0008395	steroid hydroxylase activity
	Intergenic	MF:0016620	oxidoreductase activity, acting on the aldehyde or oxo group of donors, NAD or NADP as acceptor
Muscle skeletal/cardiac	Intronic	BP:0006085	acetyl-CoA biosynthetic process
		BP:0006520	cellular amino acid metabolic process
		BP:0019752	carboxylic acid metabolic process
	Intergenic	CC:0071556	integral component of lumenal side endoplasmic reticulum membrane
		MF:0042605	peptide antigen binding
		CC:0098553	lumenal side of endoplasmic reticulum membrane
Muscle smooth	Intronic	MF:0016408	C-acyltransferase activity
		MF:0019842	vitamin binding
		MF:0050662	coenzyme binding
	Intergenic	-	-
iPSCs	Intronic	-	-
	Intergenic	-	-
Common	Intronic	-	-
	Intergenic	-	-

Supplementary Table 6. Significantly enriched GO terms associated with the intergenic and intronic eQTL-ELSs' target genes. Only the three top enriched Biological Process (BP) terms are shown for each analysis, when no BP terms are found Molecular Function (MF) and Cellular Component (CC) terms are shown instead.

Tissue	Hosting	GO term	Description
Aorta	Intronic	BP:0014910	regulation of smooth muscle cell migration
		BP:0048660	regulation of smooth muscle cell proliferation
		BP:0003205	cardiac chamber development
	Intergenic	MF:0004722	protein serine/threonine phosphatase activity
		CC:0031012	extracellular matrix
Blood	Intronic	BP:0150079	negative regulation of neuroinflammatory response
		BP:0042093	T-helper cell differentiation
		BP:0002294	CD4-positive, alpha-beta T cell differentiation in immune response
	Intergenic	BP:0048006	antigen processing and presentation, endogenous lipid antigen via MHC class Ib
		BP:0061737	leukotriene signaling pathway
		BP:0048007	antigen processing and presentation, exogenous lipid antigen via MHC class Ib
Brain	Intronic	BP:1990709	presynaptic active zone organization
		BP:0098698	postsynaptic specialization assembly
		BP:0099068	postsynapse assembly
	Intergenic	BP:0042426	choline catabolic process
		BP:0055070	copper ion homeostasis
		BP:1902003	regulation of amyloid-beta formation
Fibro/myoblasts	Intronic	BP:0072273	metanephric nephron morphogenesis
		BP:0061383	trabecula morphogenesis
		BP:0030010	establishment of cell polarity
	Intergenic	BP:0007442	hindgut morphogenesis
		BP:1902260	negative regulation of delayed rectifier potassium channel activity
		BP:0001946	lymphangiogenesis
iPSCs	Intronic	BP:0045636	positive regulation of melanocyte differentiation
		BP:0061550	cranial ganglion development
		BP:0045986	negative regulation of smooth muscle contraction
	Intergenic	BP:0021825	substrate-dependent cerebral cortex tangential migration
		BP:0043383	negative T cell selection
		BP:0030318	melanocyte differentiation
Muscle skeletal/cardiac	Intronic	CC:0071005	U2-type precatalytic spliceosome
		CC:0031674	I band
		CC:0030018	Z disc
	Intergenic	-	-
Muscle smooth	Intronic	-	-
	Intergenic	-	-
Digestive	Intronic	-	-
	Intergenic	-	-
Mucosa	Intronic	-	-
	Intergenic	-	-
Common	Intronic	-	-
	Intergenic	BP:0007156	homophilic cell adhesion via plasma membrane adhesion molecules
		BP:0098742	cell-cell adhesion via plasma-membrane adhesion molecules
		BP:0034728	nucleosome organization

Supplementary Table 7. Significantly enriched GO terms associated with the intergenic and intronic HiC-ELSs' target genes. Only the three top enriched Biological Process (BP) terms are shown for each analysis, when no BP terms are found Molecular Function (MF) and Cellular Component (CC) terms are shown instead.

Tissue	Hosting	GO term	Description
Brain	Host	GO:2000809	positive regulation of synaptic vesicle clustering
		GO:2000807	regulation of synaptic vesicle clustering
		GO:1990709	presynaptic active zone organization
		GO:0048790	maintenance of presynaptic active zone structure
		GO:1905274	regulation of modification of postsynaptic actin cytoskeleton
	Non-Host	GO:1902527	positive regulation of protein monoubiquitination
		GO:0030150	protein import into mitochondrial matrix
		GO:0030033	microvillus assembly
		GO:1901976	regulation of cell cycle checkpoint
		GO:0044743	protein transmembrane import into intracellular organelle
Muscle skeletal/cardiac	Host	GO:0055003	cardiac myofibril assembly
		GO:0046580	negative regulation of Ras protein signal transduction
		GO:0048747	muscle fiber development
		GO:0051058	negative regulation of small GTPase mediated signal transduction
		GO:0055013	cardiac muscle cell development
	Non-Host	-	
Blood	Host	GO:1905449	regulation of Fc-gamma receptor signaling pathway in phagocytosis
		GO:1903613	regulation of protein tyrosine phosphatase activity
		GO:0050855	regulation of B cell receptor signaling pathway
		GO:0045589	regulation of regulatory T cell differentiation
		GO:0050853	B cell receptor signaling pathway
	Non-Host	-	
Aorta	Host	GO:0014910	regulation of smooth muscle cell migration
		GO:0014909	smooth muscle cell migration
		GO:0034446	substrate adhesion-dependent cell spreading
		GO:0014812	muscle cell migration
		GO:0007179	transforming growth factor beta receptor signaling pathway
	Non-Host	GO:0001503	ossification
		GO:0071495	cellular response to endogenous stimulus
		GO:0009719	response to endogenous stimulus
Fibro/myoblasts	Host	GO:0051781	positive regulation of cell division
		GO:0030010	establishment of cell polarity
		GO:0007163	establishment or maintenance of cell polarity
		GO:0046578	regulation of Ras protein signal transduction
		GO:0010631	epithelial cell migration
	Non-Host	GO:0010463	mesenchymal cell proliferation
		GO:0030879	mammary gland development
		GO:0061448	connective tissue development
		GO:0007178	transmembrane receptor protein serine/threonine kinase signaling
		GO:0008284	positive regulation of cell proliferation
iPSCs	Host	GO:0003056	regulation of vascular smooth muscle contraction
		GO:0071625	vocalization behavior
		GO:0060292	long-term synaptic depression
		GO:0097106	postsynaptic density organization
		GO:0009187	cyclic nucleotide metabolic process
	Non-Host	-	-
Common	Host	-	
	Non-Host	GO:0035278	miRNA mediated inhibition of translation
		GO:0040033	negative regulation of translation, ncRNA-mediated
		GO:0045974	regulation of translation, ncRNA-mediated
		GO:0034644	cellular response to UV
		GO:0045047	protein targeting to ER

Supplementary Table 8. Significantly enriched GO terms associated with the target genes of HiC-ELSs' that are host and non-host of these ELSs. Only the five top enriched Biological Process (BP) terms are shown.

Tissue	Region	Transcription factors
Brain	Intronic	Mef2a, HINFP, Sox8, Zbtb3, ZBTB26, HAND2, Sox4, Sox2
	Intergenic	SOX13
Blood	Intronic	ELF3, RUNX1
	Intergenic	ELF3, RUNX2
Muscle skeletal/cardiac	Intronic	Mef2a, GSC, Tgif2, NR2F2
	Intergenic	-
Muscle smooth	Intronic	-
	Intergenic	-
Fibro/myoblasts	Intronic	RUNX1, ZNF263, BATF
	Intergenic	RUNX2, HAND2, TEAD3, BATF
iPSCs	Intronic	POU5F1, Sox3, ZEB1, TEAD2, VEZF1, TBX1, Foxj1, ZNF384
	Intergenic	POU5F1, ZEB1, Glis2, SRY, Sox3, Gcm1, ZNF519, Cux2, MYB, GATA3, Zfx, GATA4
Mucosa	Intronic	ETV4, CDX2, RARa, HNF4G, Klf1
	Intergenic	ZEB1, THAP1, ZNF416, ZNF384, OTX1, RARa, KLF10
Digestive	Intronic	TFCP2
	Intergenic	-
Aorta	Intronic	-
	Intergenic	NFIX
Common	Intronic	Elk4(ETS)
	Intergenic	HOXA9, TFDP1, Atf1

Supplementary Table 9. Transcription factors corresponding to the significantly enriched transcription factor binding sites (TFBSs) reported by HOMER in each group of ELSs and genomic location.

	Biosample Term Name	Biosample Type	Samples' Group	ENCODE File ID
1	HUES6	cell line	stem cells (ESC)	ENCFF205SDB
2	HUES64	cell line	stem cells (ESC)	ENCFF180QLH
3	HUES48	cell line	stem cells (ESC)	ENCFF086FKD
4	mesendoderm	in vitro differentiated cells	-	ENCFF620BVM
5	H9	cell line	stem cells (ESC)	ENCFF021HBJ
6	H9	cell line	stem cells (ESC)	ENCFF505OUS
7	H1	cell line	stem cells (ESC)	ENCFF051OUV
8	mesodermal cell	in vitro differentiated cells	-	ENCFF250CGY
9	endodermal cell	in vitro differentiated cells	-	ENCFF138DOQ
10	neuroepithelial stem cell	in vitro differentiated cells	neural progenitors	ENCFF138OGZ
11	ectodermal cell	in vitro differentiated cells	-	ENCFF332EYK
12	radial glial cell	in vitro differentiated cells	neural progenitors	ENCFF593TNG
13	neural progenitor cell	in vitro differentiated cells	neural progenitors	ENCFF112ZGF
14	mid-neurogenesis radial glial cells	in vitro differentiated cells	neural progenitors	ENCFF376XBS
15	neural stem progenitor cell	in vitro differentiated cells	neural progenitors	ENCFF455CQW
16	neural cell	in vitro differentiated cells	neural progenitors	ENCFF477EUQ
17	smooth muscle cell	in vitro differentiated cells	differentiated tissues	ENCFF281QON
18	thymus	tissue	differentiated tissues	ENCFF059PHA
19	adrenal gland	tissue	differentiated tissues	ENCFF840ANN
20	IMR-90	cell line	-	ENCFF469PXS
21	fibroblast of lung	primary cell	differentiated tissues	ENCFF292NZP
22	muscle of trunk	tissue	differentiated tissues	ENCFF800YES
23	muscle of leg	tissue	differentiated tissues	ENCFF941JIE
24	stomach	tissue	differentiated tissues	ENCFF198WHL
25	hepatocyte	in vitro differentiated cells	differentiated tissues	ENCFF093BQM
26	large intestine	tissue	differentiated tissues	ENCFF903RGX
27	small intestine	tissue	differentiated tissues	ENCFF543DVJ

Supplementary Table 10. ENCODE catalogues of cell type-specific candidate cis-Regulatory Elements (cCREs) for 27 human embryonic samples. The accession number (ENCODE File ID) allows to uniquely identify the catalogue on the ENCODE portal (https://www.encodeproject.org/).

Samples	Group-specific ELSs
ESCs	3,112
neural progenitors	784
differentiated tissues	1,166

Samples	Common ELSs
all	94

Supplementary Table 11. [upper panel] Number of ELSs that are specific to each of the 3 groups of 22 selected human embryonic samples. Group-specific ELSs are active in ≥ 80% of the samples within a group, and in at most 1 outer sample (i.e. a sample that does not belong to the considered group). [lower panel] Number of ELSs active in 100% of the 22 selected human embryonic samples (common ELSs).

Genomic location	Samples' Group	FDR	Odds ratio	Confidence interval
intronic	ESCs	3.1e-02	1.67	1.08-2.62
	neural progenitors	1.3e-04	2.45	1.55-3.92
	differentiated tissues	9.9e-05	2.48	1.58-3.94
exonic	ESCs	7.6e-01	0.87	0.14-36.11
	neural progenitors	1.0e+00	1.57	0.23-67.33
	differentiated tissues	8.1e-01	2.04	0.33-84.53
intergenic	ESCs	3.1e-02	0.60	0.39-0.93
	neural progenitors	9.9e-05	0.40	0.25-0.63
	differentiated tissues	9.9e-05	0.39	0.24-0.6

Supplementary Table 12. For each group of samples we assessed, with Fisher's exact test, significant differences in the proportions of common vs group-specific ELSs that overlap intronic, exonic and intergenic regions. P value (FDR-corrected), odds ratio and confidence interval are reported for each test.

Group	Genes ∩ ELSs			
	Introns	Exons	Both	Total
ESCs	907 (89.27%)	21 (2.07%)	88 (8.66%)	1016
neural progenitors	359 (87.56%)	13 (3.17%)	38 (9.27%)	410
differentiated tissues	492 (86.16%)	24 (4.2%)	55 (9.63%)	571
common	33 (82.5%)	1 (2.5%)	6 (15%)	40

Supplementary Table 13. Number of genes whose introns and/or exons intersect group-specific and common ELSs identified in embryonic samples.

Group	GO term	Description
Neural progenitors	BP: 0060291	Long-term synaptic potentiation
	BP: 0050770	Regulation of axonogenesis
	BP: 0097061	Dendritic spine organization
	CC: 0008328	Ionotropic glutamate receptor complex
	CC: 0098878	Neurotransmitter receptor complex
	CC: 0014069	Postsynaptic density
	MF: 0004970	Ionotropic glutamate receptor activity
	MF: 0005089	Rho guanyl-nucleotide exchange factor activity
	MF: 0008013	Beta-catenin binding
Differentiated tissues	BP: 1900020	Positive regulation of protein kinase C activity
	BP: 1900040	Regulation of interleukin-2 secretion
	BP: 0060766	Negative regulation of androgen receptor signaling pathway
	CC: 0098651	Basement membrane collagen trimer
	CC: 0098644	Complex of collagen trimmers
	CC: 0005583	Fibrillar collagen trimer
	MF: 0044548	S100 protein binding
	MF: 0035252	UDP-xylosyltransferase activity
	MF: 0030020	Extracellular matrix structural constituent conferring tensile strength
Stem cells (ESC)	BP: 0042908	Xenobiotic transport
	BP: 0045986	Negative regulation of smooth muscle contraction
	BP: 0098698	Postsynaptic specialization assembly
	CC: 0099092	Postsynaptic density, intracellular component
	CC: 0031304	Intrinsic component of mitochondrial inner membrane
	CC: 0008328	Ionotropic glutamate receptor complex
	MF: 0008146	Sulfotransferase activity
	MF: 0005547	Phosphatidylinositol-3,4,5-triphosphate binding
	MF: 0070300	Phosphatidic acid binding
Common	CC: 0071565	nBAF complex
	CC: 0016514	SWI/SNF complex
	CC: 0070603	NI/SNF superfamily-type complex

Supplementary Table 14. Significantly enriched GO terms associated with the genes harboring intronic ELSs identified in embryonic samples. Only the top three enriched terms are shown in each analysis (BP: Biological Process; CC: Cellular Component; MF: Molecular Function).

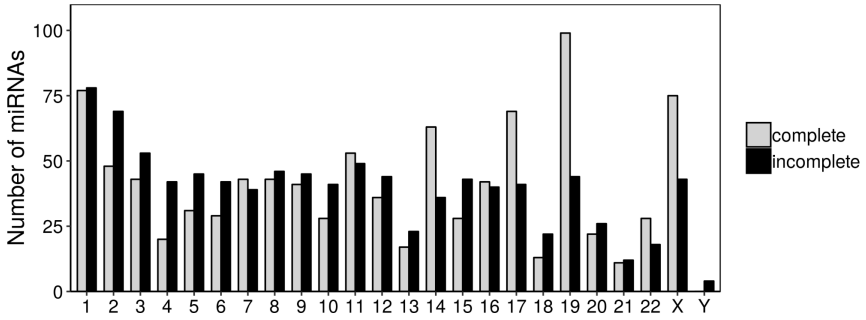
Gene	Coordinates	ELS ID	Tissue	Location	Hosting	Peak ChIP-seq k27ac	Primer
PPP3CA	4:102023034-102023302	EH37E0737564	Brain	intronic	Host	Neuron 1	F:GCCAACACTCGCTACCTCTT R:AAGGCCCACAAATACAGCAC
CAPZB	1:19670858-19672569	EH37E0073593	Brain	intronic	Host	ESC1	F:CCTGGTCCCCAGTCTATGTG R:ACCACCTTGTCTCTGGCAAT
SDHD	11:113019264-113019880	EH37E0240118	Brain	intronic	Non host	Neuron 2	F:CAGAATGGTGTGGAGTGCAG R:AGTGGAGAGATGCAGCCTTG
AKR7A2	1:19670858-19672569	EH37E0073593	Brain	intronic	Non host	ESC1	F:GCCGAGATCTGTACCCTCTG R:GAAGAGCTCCGTTTCCACCT
RPA2	1:28307519-28308214	EH37E0078769	Brain	intronic	Non host	Neuron1/2	F:CCTTCTCAAGCCGAAAAGAA R:TCATCAACCAAAGTGGCAGA
DRG1	22:31735086-31735685	EH37E0629324	Brain	intronic	Non host	ESC1	F:TTACTCCAAAGGGTGGTGGT R:CAAATCCAATTCGAGCATCA
CSPG5	3:47577542-47578617	EH37E0656905	Brain	intergenic	Non host	ESC1	F:CCACTGCTGCTGTTTCTGG R:CTGCCCTTCACCAGCTCTT
CTNNA2	2:80527433-80528886	EH37E0528734	Brain	intronic	Host	Neuron 1	F:CAGAAAGGCTGTGCTGATGA R:CTTGTCCTGCTACGCACATC
KCNQ2	20:62086208-62086923	EH37E0609018	Brain	intronic	Host	ESC1/2, Neuron 1	F:CACAGGCAGAAGCACTTTGA R:GAGAGGTTGGTGGCGTAGAA
ACTB	7:5733031-5733564	EH37E0886351	Common	intronic	Non host	All	F:ATTGGCAATGAGCGGTTC R:TGAAGGTAGTTTCGTGGATGC
BAHD1	15:40390946-40391339	EH37E0363650	Common	intronic	Non host	All	F:GATGATGAGCCTCCTGTGGT R:GCGATGCAAACACTTCATTC
BMF	15:40390946-40391339	EH37E0363650	Common	intronic	Host	All	F:CAGTGCATTGCAGACCAGTT R:AAGGTTGTGCAGGAAGAGGA
GSK3A	19:40939210-40940400	EH37E0490611	Common	intergenic	Non host	All	F:CTCATTTGGGGTCGTGTACC R:GATCTGCAGCTCTCGGTTCT
VEGFB	11:62320405-62321311	EH37E0221959	Common	intronic	Non host	All	F:CTGGCCACCAGAGGAAAGT R:CATGAGCTCCACAGTCAAGG

Supplementary Table 15. Selection of brain-specific and common ENCODE ELSs overlapping with hESC-derived neural maturation ChIP-seq. Target gene is identified by HiC interaction and only genes regulated by one ELS in our ENCODE analysis are selected. The ELSs coordinates and the ELS ID, as well as the genomic location (intronic vs intergenic) and the nature of the targeted genes (host or non-host) are shown in the 5th and 6th column. The presence of peaks on the neural maturation ChIP-seq experiment is shown and the primers used for gene expression analysis are also shown in the last column.

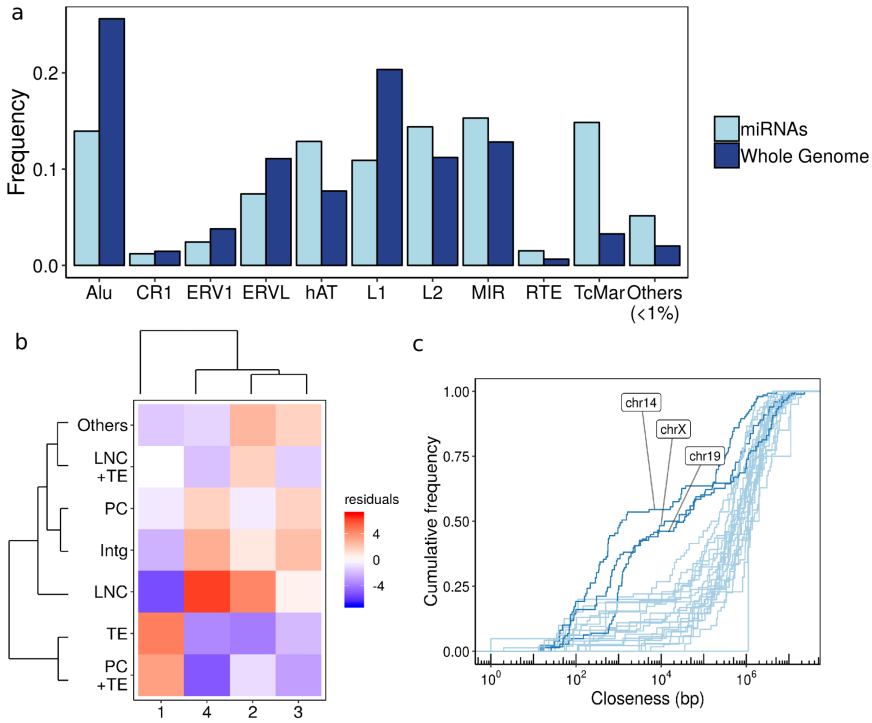
Supplementary Material:

Signatures of genetic variation in human microRNAs point to processes of positive selection related to population-specific disease risks

Pablo Villegas-Mirón, Alicia Gallego, Jaume Bertranpetit, Hafid Laayouni and Yolanda Espinosa-Parrilla



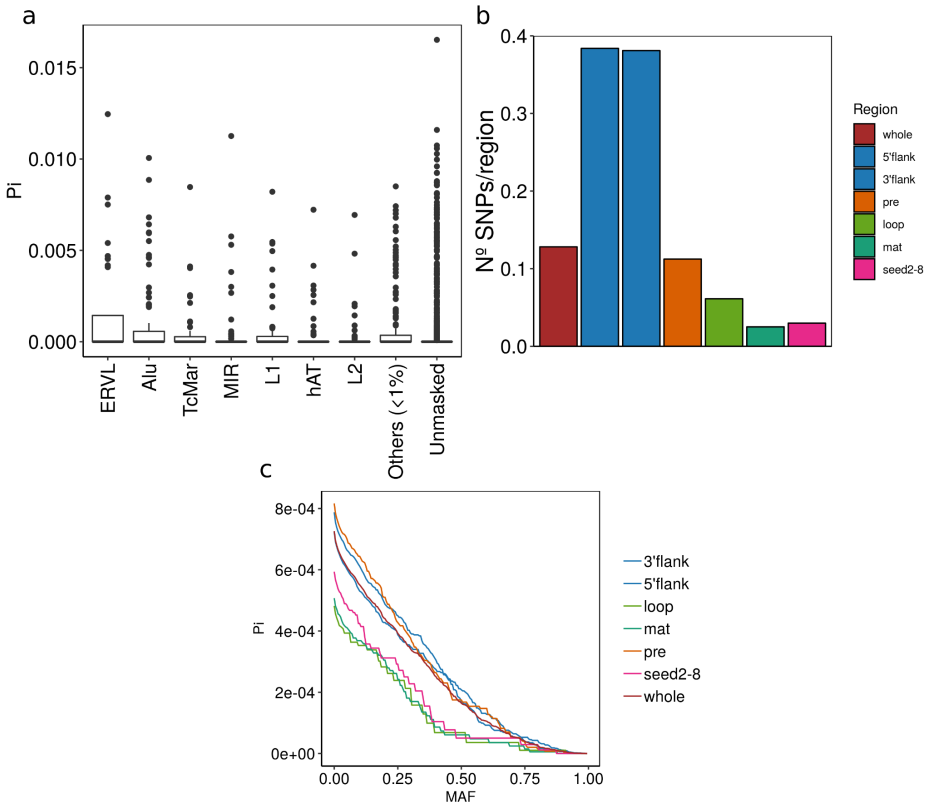
Supplementary Fig. S1 Number of miRNAs per chromosome that present both mature sequences in their hairpin (complete annotation) and only one mature sequence in one of their arms (incomplete annotation)



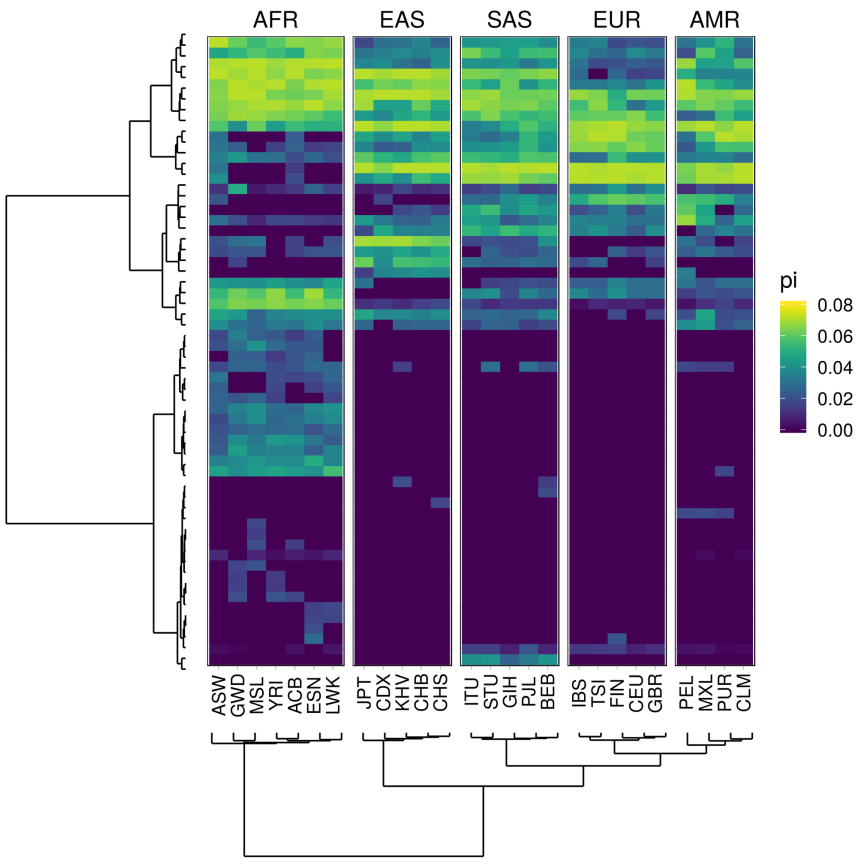
Supplementary Fig. S2 (a) Frequencies of transposable elements described by the RepeatMasker database (v4.0.5) in the whole genome and found overlapping miRNA sequences. **(b)** Chi square residuals associated with each of the genomic context categories across conservation groups. Dendrograms show the hierarchical clustering performed across rows (genomic context) and columns (conservation). **(c)** Cumulative frequency of the closeness found between miRNAs (distance to the closest miRNA) in each chromosome. The increase of

frequency in chromosomes 14, 19 and X show groups of highly close miRNAs that correspond to the main clustering hotspots in the human genome

Supplementary Fig. S3 Genomic location of human miRNA clusters



Supplementary Fig. S4 (a) Nucleotide diversity of miRNAs hosted by the different families of transposable elements. The "Others" category is made by minor categories represented by less than 1% of the total miRNAs. **(b)** SNP density per functional region calculated in the whole miRNA dataset. **(c)** Mean nucleotide diversity of the miRNA functional regions across the SNP MAF range



Supplementary Fig. S5 Heatmap showing the mean nucleotide diversity values per population of the seed regions harbouring one or more SNPs of the whole dataset. The dendrograms represent the hierarchical clustering performed on the miRNAs (rows) and populations (columns)

Supplementary Tables S1, S2 and **S3** are deposited in the Biorxiv repository associated to the preprint (Supplementary File 2):

Villegas-Mirón P, Gallego A, Bertranpetit J, Laayouni H, Espinosa-Parrilla Y. 2021. Signatures of genetic variation in human microRNAs point to processes of positive selection related to population-specific disease risks. bioRxiv doi: https://doi.org/10.1101/2021.05.24.445417