ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN

SCIENZE DELLA TERRA, DELLA VITA E DELL'AMBIENTE (STVA)

Ciclo 36

Settore concorsuale: 05/B1 - Zoologia e Antropologia
Settore scientifico disciplinare: BIO/08 - Antropologia

# RETROTRANSPOSONS IN HUMAN EVOLUTION AND BRAIN DISEASES: INVESTIGATING MODERN AND ANCIENT VARIABILITY THROUGH THE LENS OF MOBILE ELEMENTS

**Presentata da:** Giorgia Modenini

**Coordinatore dottorato**

Barbara Cavalazzi

**Supervisore**

Alessio Boattini

**Co-supervisore**

Davide Pettener

Esame finale anno 2024

# Index

“If you know you are on the right track,
if you have this inner knowledge,
then nobody can turn you off…
no matter what they say…”

Barbara McClintock

## Abstract

Transposable Elements (TEs) are mobile genetic elements with the ability to replicate themselves and insert their copies in a new genomic location; notably, sequences derived from TEs make up to 53-60% of the human genome. While the vast majority of TEs are not transpositionally active, they can still play functional roles such as modulating the expression or alternative splicing of nearby genes. Moreover, they can generate potentially inheritable insertions, thereby creating human-specific polymorphisms, that in the last years have been used to study human evolution, variability and susceptibility to medical disorders. Here we present a large-scale *in silico* study on polymorphic TEs, used as genetic variability markers in modern human populations and ancient genomes. Furthermore, we investigated the possible role of TEs in conferring a risk of developing brain disorders, such as addiction, schizophrenia and Alzheimer's disease. Our results point toward an important role of polymorphic TEs in: 1) shaping human genome evolution, especially the neural genome and human-specific cognitive abilities; 2) influencing phenotypic variability and adaptation; 3) being the target of recent instances of positive selection or more ancient event of adaptive introgression; 4) conferring risk of developing neuro-psychiatric disorders.

# 1. Introduction

### 1a. Transposable Elements: structure and mobilization mechanisms

Since their discovery in the late 1940's by Barbara McClintock, Transposable Elements (TEs) experienced ups and downs. First dismissed as "junk DNA" or "parasites", in the last decades they have been recognized both as deleterious (they have been linked to numerous diseases such as brain disorders and cancers), and beneficial, for example in the case of mammalian genomes evolution (Platt et al., 2018).

In humans, 53-60% of the genome is made of repetitive sequences derived from TEs (deKoning et al., 2011; Hoyth et al., 2022). Retrotransposons, a particular type of TEs that belong to mobile elements Class I, move via an RNA intermediate that is then reverse-transcribed and use a *copy-and-paste* mechanism that allows these elements to increase the number of their copies. They are classified into Long Terminal Repeats retrotransposons (LTRs), such as Human Endogenous Retroviruses or HERVs, and non-LTRs, such as SINEs (Short Interspersed Nuclear Elements, for example: Alus), LINEs (Long Interspersed Nuclear Elements) and SVAs (SINE-VNTR-Alu).

Full-length human LINE-1s (L1s) are ~6 kilobases (kb) long (Scott et al., 1987; Dombroski et al., 1991) with a ~900 bp long 5' untranslated region (5' UTR) with internal promoter activity (Swergold, 1990), a ~150 bp long 3' UTR and a poly(A) tail (Scott et al., 1987). L1s also contain two Open Reading Frames (ORF1 and ORF2), which encode, respectively, for a ~40 kDa protein with RNA binding and chaperone activities (Hohjoh et al., 1997; Kulpa et al., 2005) and for a ~150 kDa protein with reverse transcriptase (RT) and endonuclease (EN) activities (Mathias et al., 1991; Feng et al., 1996). Both ORFs are required for L1s mobilization in the human genome (Craig, 2002).

The main family of SINEs in the human genome is represented by Alus. Alu elements are ~300 bp long and have a dimeric structure determined by the fusion of two 7SL-RNA-derived monomers, separated from each other by an A-rich linker region (Ullu & Tschudi, 1984). The 5' region carries an internal RNA III polymerase promoter, and at the end of the element there is an oligo dA-rich tail of variable length.

SVAs are primate-specific retrotransposons that terminate with a poly-A tail (similarly to L1s). Their name synthesizes the three components of their sequence: the 3' LTR region of the endogenous retrovirus HERV-K10 (SINE-R), a Variable Number Tandem Repeats (VNTR) region and an antisense Alu-like region. Because of the polymorphism of their VNTR region copy number (48–2,306 bp), SVAs may vary in size; however, more than a half are ~2 kb long (Wang et al., 2005).

Finally, HERVs are endogenous viral elements that resemble and are derived from infectious retroviruses, however they are typically not infectious. HERVs are composed of group-associated antigen (gag), polymerase (pol) and envelope (env) genes, along with two LTRs at the 3' and 5' regions (Löwer et al., 1996; Griffiths, 2001).

L1s are the only known autonomously active TEs in humans (Goodier, 2016; Rishishwar et al., 2017a and 2017b); on the other hand, retrotransposition in Alus and SVAs is still made possible thanks to the L1s' enzymatic machinery (Guio & Gonzalez, 2019).

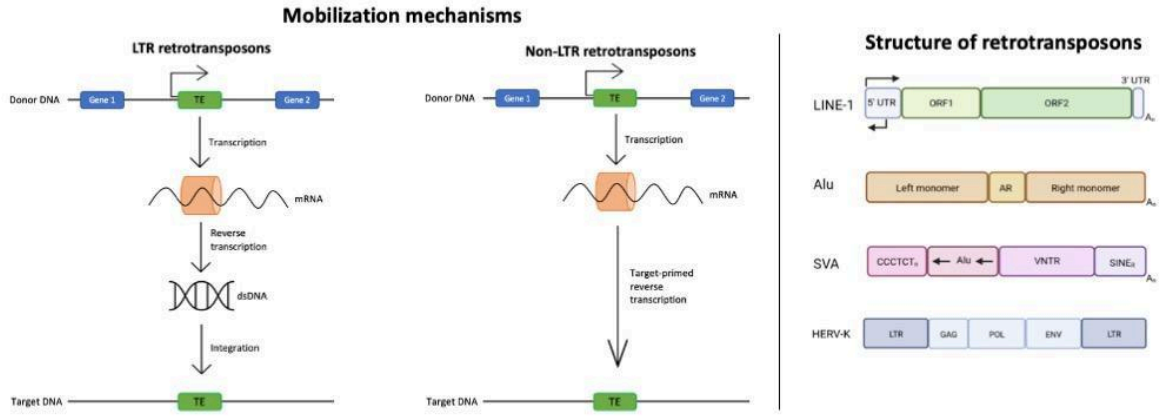



**Figure 1a.1**. Structure and mobilization mechanisms of retrotransposons.

**1b. Transposable Elements as genomic factors of variation and diseases in the human brain**

In the last decades, the impact of TEs on the host's genome has been the focus of intense research, and it has been shown that TE insertions can generate diversity in various ways, being both positive and detrimental players in human genome evolution (reviewed in Reilly et al., 2013; Bourque et al., 2018; Gebrie 2023).

While the vast majority of TEs are no longer transpositionally active, they can still play a functional role as exapted enhancers or transcriptional start sites (Rangwala et al., 2009; Deininger, 2011; Su et al., 2022; Babaian & Mager, 2016), by inserting transcription factor binding sites (TFBS) (Emera & Wagner, 2012; Lynch et al., 2015) or by acting as novel RNA genes such as long non-coding RNAs (lnc-RNAs) (Hezroni et al., 2015). Therefore, TEs participate in regulating the expression of nearby genes, at transcriptional and post-transcriptional levels, providing a crucial role as both *cis*- and *trans*-regulatory RNA sequences (Ali et al., 2021).

Growing evidence points toward an important role of TEs in shaping evolutionary and adaptive processes, such as (but not limited to): generation and transcriptional regulation of genes and pseudogenes (Moran et al., 1999; Chuong et al., 2017), somatic mosaicism (Muotri et al., 2005; Baillie et al., 2011; Evrony et al., 2012), increase in complexity and evolution of gene regulatory networks (Feschotte, 2008) and alteration of epigenetic mechanisms (Fedoroff, 2012). For instance, the insurgence of the V(D)J system of acquired immunity (Kapitonov & Jurka, 2005; Koonin & Krupovic, 2014; Huang et al., 2016) is one of the most notable biological processes associated with the domestication of TE-derived sequences, but TEs also play essential roles in embryogenesis (Friedli & Trono, 2015; Gerdes et al., 2016; Percharde et al., 2018) and neurogenesis (Muotri et al., 2005; Evrony et al., 2012; Notwell et al., 2015).

On one hand, the key part of brain evolution is changes in brain development, and there is significant evidence that TE-related mechanisms participated in shaping mammalian embryonic development, including neuronal differentiation (Ferrari et al., 2021). Indeed, "both LTR and non-LTR retrotransposons appear to have contributed to mammalian brain evolution by acting as sources of

novel non-coding RNAs, proteins, enhancers, RNA regulatory sites and sites for 3D genome organization" (Ferrari et al., 2021).

On the other hand, recent data provide evidence for dysregulation of TEs in some neurological disorders (Cordaux & Batzer, 2009). Although the mechanistic details of the functional and evolutionary impact of TEs in the brain and nervous system are still unknown, a growing number of studies suggests that TEs contribute to neurological disorders, including schizophrenia (Erwin et al., 2014; Guffanti et al., 2014; Guffanti et al., 2018; Modenini et al., 2023). These findings have major implications for understanding the neuroplasticity of the brain, which probably had a remarkable impact on brain evolution in Mammals, especially in Hominids, and could contribute to vulnerability to neurological disorders (Ahmadi et al., 2020).

The discovery of active TEs in the brain has raised questions about their influence on brain development and functioning (Reilly et al., 2013): on one hand, TE insertions may exert a controlling influence on flanking genes, causing a functionally relevant impact on the diversification of neuronal cell types or on the function of differentiated neurons. On the other hand, a detrimental impact of unregulated transposon expression must be considered: for instance, altered retrotransposon expression or function appear to be associated with neurodegeneration and aging (Reilly et al., 2013; Macciardi et al., 2022).

HERVs have been the most studied TEs in neuropsychiatric disorders. In the last two decades, there has been mounting evidence that K and W families of Human Endogenous Retroviruses (HERV-K and HERV-W) are implicated in psychiatric disorders: for instance, abnormal expression of HERV-Ws and HERV-Ks has been found in blood, cerebrospinal fluid and post-mortem brain samples from patients diagnosed with schizophrenia, both at onset and in later stages of the disease. HERV activity has also been linked to other psychiatric disorders, such as bipolar disorder, major depression, autism and Attention Deficit Hyperactivity Disorder (ADHD) (Guffanti et al., 2014).

LINEs and SINEs have also received attention for neurological disorders studies. LINEs have mostly been studied with reference to early and later brain development: Baillie and colleagues (2011) found that protein-coding loci are disproportionately affected by non-LTR retrotransposons, with

overrepresentation of LINE-1s in introns and Alus in exons. Overall, retrotransposon insertions seem to predominantly affect neurogenesis and synaptic function. Abrusan (2013) proposed that a reduced expression of genes affected by L1 insertions could potentially influence the biosynthesis and metabolism of different neurotransmitters such as dopamine, serotonin and glutamate.

In conclusion, transposable elements have shaped mammalian brain evolution and have also been identified as key players in brain disorders development. Therefore, the identification of transposable elements that could have contributed to *Homo sapiens*-specific brain features is important to unveil the functional role that these elements had on human brain evolution. Moreover, the so-called "genomic trade-off", a mechanism by which "changes in the genome that are overall beneficial persist even though they also produce disease in a subset of individuals" (Sikela & Quick, 2018), must also be considered when studying the impact of transposable elements on the human brain. This evolutionary mechanism is possibly ascribable to schizophrenia and autism spectrum disorder (ASD), the first being a real biological paradox: schizophrenia is indeed characterized by "high heritability, but it is associated with decreased reproductive success" (Sikela & Quick, 2018). Therefore, it is possible that structural variants such as TEs could take part in the aforementioned disorders but at the same time contributing to the rise of human superior cognitive abilities.

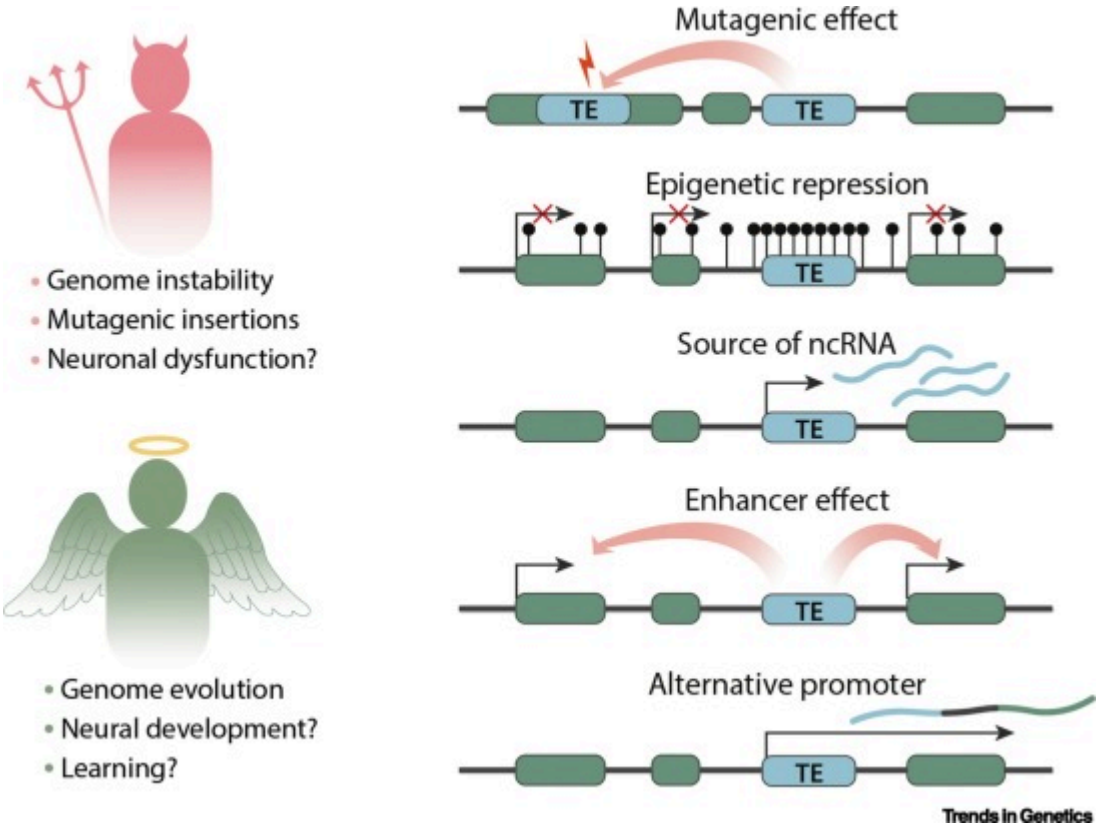



**Figure 1b.1**. Transposable Elements have both positive and negative effects on the host's genome: in particular, they can cause genome instability and neuronal dysfunction, but they are also important for genome evolution and neural development (Figure from Jönnson et al., 2020).

**1c. Polymorphic transposable elements as genetic variability markers**

TEs mobilization in germline cells can enhance diversity through the creation of potentially inheritable insertions, thereby generating human-specific polymorphisms. For instance, the analysis of the "1000 Genomes Project" (1KGP) dataset (2,504 unrelated individuals from 26 different modern populations) identified ~16,000 polymorphic TE loci, 93% of which show a worldwide allele frequency < 5%, indicating that overall polymorphic TEs are deleterious and have faced purifying selection (Rishishwar et al., 2018), but also showing promising results as valuable genetic markers for studies on human ancestry and evolution.

Even if frequently under strong negative selection, the presence of polymorphisms in mobile element loci is an exciting potential source of information, in particular when analyzing modern human diversity, as polymorphic TEs have been shown to provide several advantages when compared to other more widely used genetic markers. Firstly, polymorphic TEs shared among individuals are usually an example of identity by descent (IBD) (Batzer & Deininger, 2002): the high number of potential insertion sites in the genome and the very low transposition rates [~100-200 new Alu insertions per million years (Batzer et al., 1994)], mean that the probability of independent events of insertion in the same genomic location in two unrelated individuals is negligible. The second reason is that newly inserted TEs rarely undergo deletion, and even when they do they leave on the genome a molecular signature, making them highly stable polymorphisms (Rishishwar et al., 2015). Therefore, polymorphic TE markers are usually free of homoplasies (i.e., identical states that do not represent a shared ancestry) (Ray et al., 2006), representing a more accurate marker of relationship with respect to classical genetic markers, such as SNPs or microsatellites (Ray & Batzer 2011). Another advantage of using polymorphic TEs as variability markers is the fact that the ancestral state of every locus can be defined as the absence of the insertion (Perna et al., 1992), and this allows to draw population relationship trees with more confidence. Finally, polymorphic TEs proved to be also practically useful markers since they can be rapidly and accurately typed with PCR-based assays (Rishishwar et al., 2015).

Rishishwar's work (2015) was reinforced more recently by Watkins and colleagues (2020), who used the same approach on the Simons Genome Diversity Project (SGDP), identifying new polymorphic TEs not reported by the 1KGP and confirming the finding that polymorphic TEs are indeed exceptionally useful as variability and ancestry markers for human populations, while also potentially providing a basis for studies of medical susceptibilities and maladaptations of human populations.

Therefore, we decided to rely on polymorphic TEs to study both human variations and diseases induced by the presence or absence of these structural variants, which have been shown to be valuable genetic markers to pursue these goals.

**1d. Modern and archaic humans: close encounters**

In the early 1950's, some years after the observations of Barbara McClintock, in the journal *Nature* appeared the first reconstruction of the molecular structure of DNA (Watson & Crick, 1953). Fifty years later, in 2001, the first human sequence was released (Lander et al., 2001): it was an absolute revolution, providing scientists with the opportunity to disentangle the mysteries of genetic evolution and diseases. Then, another revolution began in 2010 when, for the first time, the sequence of a near-complete nuclear genome was obtained from the tissue of an ancient individual who lived about 4,000 years ago (Rasmussen et al., 2010): it was the beginning of the "ancient DNA revolution". Since then, thousands of ancient genomes, belonging to different human groups, have been published, allowing scientists to unravel the secrets of human history and evolution.

*Homo sapiens* appeared in Africa ~ 350 kya (thousand years ago) (Hublin et al., 2017, Schlebusch et al., 2017) and reached the Near East with the second Out of Africa, 80-130 kya. Other *Homo* species migrated out of Africa before the emergence of *Homo sapiens*: *Homo erectus* in Asia and *Homo heidelbergensis* in Europe and Middle East.

Originating from a European clade of *Homo*, Neanderthals lived throughout Europe, Middle East and Asia between 400kya and 40kya, when they went extinct, similarly as Denisovans, a sister group of Neanderthals whose remains have been found only in Siberia and East Asia. The common ancestor of Neanderthals and Denisovans split from that of *Homo sapiens* between 750kya and 550kya and the split between those archaic groups occurred 381-473kya (Prüfer et al., 2014).

Different studies (Reich et al., 2010; Green et al., 2010; Reich et al., 2011) revealed that introgression events (i.e. the acquisition of genetic material through mating and gene flow from archaic hominins into the modern human gene pool) occurred between Anatomically Modern Humans (AMH) and Neanderthals/Denisovans (Figure 1d.1): indeed, 1.5-2% of the genome of people living outside Africa is made of Neanderthal-derived sequences, and in Asian and Oceanian populations up to 5% of the DNA is derived from Denisovans (Meyer et al., 2012). Moreover, a recent study (Posth et al., 2017) suggests that more than 100 kya gene flow occurred in both directions, from Neanderthals and Denisovans to AMH and from AMH to Neanderthals.

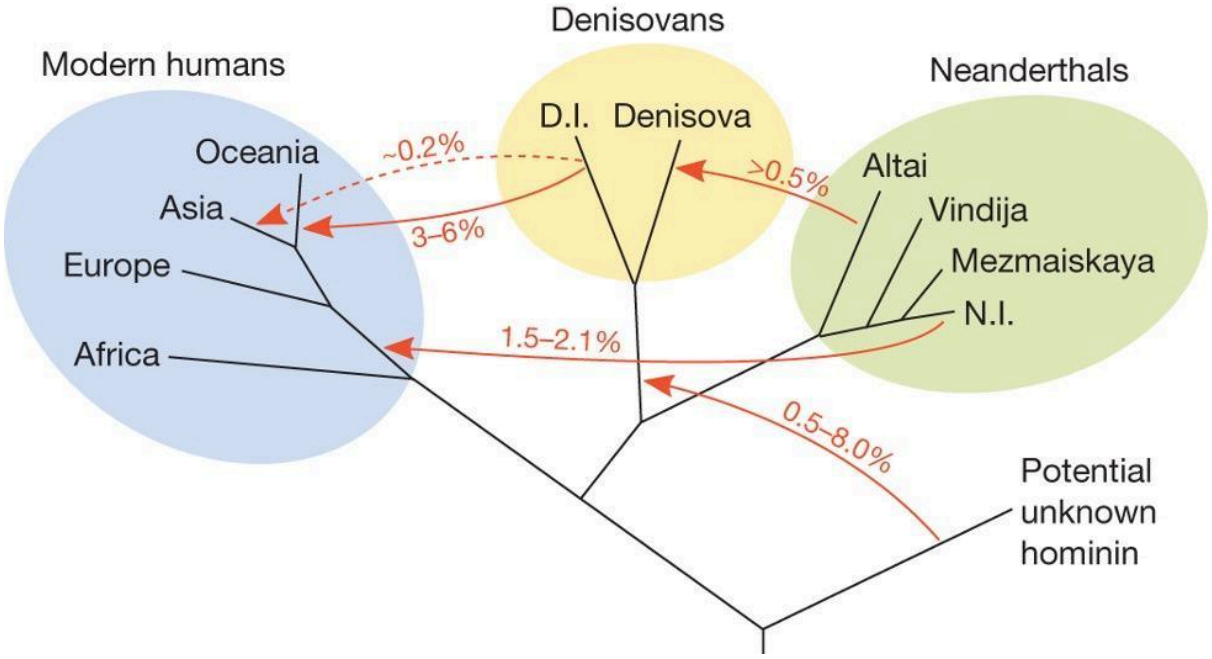



**Figure 1d.1**. Possible model of gene flows that occurred in the late Pleistocene. Modern human non-african populations carry 1.5-2% of Neanderthal-derived DNA, while populations in South-East Asia and Oceania carry up to 5% of Denisova-derived DNA. (Figure from Prüfer et al., 2014)

The best known example of adaptive introgression, a mechanism in which "beneficial variants acquired from archaic humans may have accelerated adaptation and improved survival in new environments" (Racimo et al., 2015), is the one found in Tibetan highlanders of the Tibetan Plateau in East Asia. These populations carry signals of DNA introgression from Denisovan or Denisovan-related individuals (Huerta-Sánchez et al., 2014), in particular in the EPAS1 gene, which is under strong positive selection and helped Tibetans to adapt to the extreme environment of the Tibetan Plateau (Beall et al., 2010).

Most of the studies on introgression events have relied on single nucleotide polymorphisms. However, a recent work (Hsieh et al., 2020) suggested that large copy number variants (CNVs) from Neanderthals and Denisovans contributed to local adaptation and differentiation of modern human populations. Moreover, Guichard and colleagues (2018) discovered several Neanderthal- or

Denisovan-specific TEs, along with AMH-specific insertions that possibly contributed to modern human brain evolution.

## 2. Aims of the study

In this study, we aimed at disentangling the role of polymorphic TEs in shaping various aspects of human biology, from evolution to adaptations and susceptibility to medical disorders, such as brain diseases and addiction to substances. To do so, we analyzed ancient and modern human individuals, representative of most of the worldwide variability: Europe, Asia, Africa and America.

We first inspected the content of retrotransposon insertions in ancient samples (i.e., AMH, Neanderthals and Denisovans), and compared them to modern humans.

Then, we aimed at identifying polymorphic TEs potentially related to the adaptation to an extreme environment, such as high altitude. To do so, we inspected the TEs content in modern populations from East Asia and the Tibetan Plateau (i.e., Sherpa and Tibetans compared to lowlanders such as Han Chinese).

Third, we wanted to know if some phenotypes/behavioral traits (Body Mass Index, tobacco use and alcohol consumption) are related to the presence/absence of polymorphic TEs: to achieve this goal, we analyzed 586 genomes from six isolates from Northern Italy, for which genotype and phenotype information of TEs were available.

Lastly, we focused on the role of TEs in shaping human brain evolution and diseases through the study of 20 genomes from the DorsoLateral PreFrontal Cortex (DLPFC) of ten people diagnosed with schizophrenia (plus ten controls) and the study of expression patterns of TEs in 25 individuals who developed Alzheimer's disease (AD).

In summary, our study wanted to elucidate the different roles and effects that transposable elements have on their human hosts. Through the analysis of ancient and modern human samples, we were able to reconstruct, brick after brick, the central role of retrotransposons in shaping human evolution, adaptation and phenotypes. Therefore, our study provides a vast *in silico* analysis of different aspects of TEs variation and influence on present and past human biology.

# 3. Searching for polymorphic transposable elements in ancient DNA

Giorgia Modenini[1], Alessio Boattini[1], Gabriele Scorrano[2]. *A glimpse inside human evolution using polymorphic transposable elements*. Draft version.

Affiliations:
1: BiGeA Department, University of Bologna, Bologna, Italy
2: Globe Institute, Lundbeck Foundation, University of Copenhagen, Copenhagen, Denmark

**Background**

As introduced in the previous chapters, polymorphic transposable elements are valuable genetic markers to study human variability and evolution (Rishishwar et al., 2015; Gardner et al., 2017; Watkins et al., 2020) but, to our knowledge, no research has been performed on ancient genomes, except for two high-coverage *Homo* individuals and using completely different approaches (Gardner et al., 2017; Guichard et al., 2018): Altai Neanderthal (Prüfer et al., 2014) and Denisova (Meyer et al., 2012). This is possibly due to several technological and methodological issues: 1) every method for the analysis and identification of human polymorphic TEs is designed to cope with modern genomes, which are usually paired-end sequences - while ancient DNA usually becomes available with high proportions of single-end sequences DNA; 2) due to their repetitive nature, transposable elements are often cut out from the sequence during the alignment to a reference genome; 3) ancient DNA is characterized by low percentage of endogenous ancient human DNA, high degradation and frequent sequence alterations due to cytosine deamination (Wang et al., 2021).

Ancient DNA (aDNA), due to its nature of being exposed to taphonomic processes and external agents, is subject to have a high number of base pairs substitutions which modify the original sequence. aDNA damages “include the fragmentation of DNA molecules into ultra-short DNA fragments, the conversion of the four nucleotides into various derivates and the cross-linking of DNA to other molecules” (Orlando et al., 2021). As reported by Briggs et al. (2007), cytosines are first deaminated, then converted into uracils and thereafter sequenced as thymine analogues. This process is known as “C to T misincorporations”, and includes G to A misincorporations in the case of double-stranded DNA libraries.

The work from Guichard and colleagues (2018), who studied Anatomically Modern Human (AMH), Neanderthal, Denisova and Chimpanzee genomes, identified a set of species-specific retrotransposons and established their distribution in modern human populations. Interestingly, their results strongly suggest that not only TEs impacted the differentiation and evolution of present-day humans, but also that the genes mapped by these insertions are enriched in the brain, while also taking part of networks related to neuron maturation and migration (Guichard et al., 2018).

In this work we aimed at identifying polymorphic TEs in ancient *Homo sapiens* genomes and to evaluate *in silico* their possible influence on AMH phenotypes and evolution. To do so, we: 1) analyzed with two different methods tens of ancient *Homo sapiens* samples covering ~45 thousand years, including two archaic individuals (Altai Neanderthal and Denisova); 2) compared the results from the two methods; 3) compared the presence/absence of TEs in ancient samples with their allele frequencies in modern human populations; 4) performed several *in silico* analyses to reconstruct involved gene networks and pathways and to infer a possible role for the most significant transposable elements.

**Materials and Methods**

33 high-coverage (> 6-fold) ancient genomes of *Homo sapiens* have been studied with *ngs te mapper 2* (Han et al., 2021), jointly with one Neanderthal and one Denisovan individual (Figure 3.1) to look for reference TEs in ancient samples. The individuals analyzed represent several prehistoric and historic periods from the Upper Paleolithic to historic ages (45-0.2 kya). Samples were analyzed at the GeoGenetics Centre - Lundbeck Foundation - of Copenhagen, Denmark, under the supervision of Prof. Gabriele Scorrano, PhD. A complete list of samples with age, location and related publication can be found in Supplementary Table S1.

*ngs te mapper 2* is the only software that works with single-end sequences and its output - a single *bed* file for every sample analyzed - contains only information about the presence and genomic position of the mobile element, and does not provide any information about its absence or the existence of missing data in the analyzed genome. The tool is able to identify TE insertions from short-read

next-generation sequencing (NGS) data: first, reads are "queried against a library of TE sequences to identify junction reads that span the start/end of the TE"; then, the software alignes the unmodified junction reads to the reference genome; finally, "non-TE and TE components are clustered to identify the two ends of the reference TE insertion" (Han et al., 2021).

3 of the samples used in *ngs te mapper 2* analysis (Altai Neanderthal, Denisova and UstIshim) and 5 additional ancient *Homo* individuals (Chagyrskaya Neanderthal, Kostenki14, Chan, GB1 and Ötzi) were further analyzed with the Mobile Element Locator Tool (MELT: Gardner et al., 2017) to search for both reference and non-reference polymorphic TEs. Unlike *ngs te mapper 2*, MELT looks for mobile elements only in paired-end reads and gives information about both the presence or absence of the mobile element, while also outputting possible missing data: thus, MELT was applied only on those samples for which paired-end sequences were available. MELT identifies polymorphic TEs "by searching for signatures of discordant read pairs and split reads" in whole-genome sequencing data (Gardner et al., 2017) and is designed to analyze *bam* files aligned with bwa-mem (Li & Durbin, 2009). MELT retrieves information about genomic position of the TE, its length, target site duplication (TSD), subfamily and location (i.e., intergenic or genic, the last one being further characterized with the specific location of the TE: intronic, exon, promoter, terminator, 3' UnTranslated Region - UTR - and 5' UTR); it also performs genotyping of both reference and non-reference TEs. The output is a final *vcf* file containing all the described information, in which rows are the mobile elements and columns are the genotypes of the analyzed samples.

Results from both analyses were compared to 2,504 modern human samples from the 1000 Genomes Project (1KGP) (The 1000 Genomes Project Consortium, 2015).

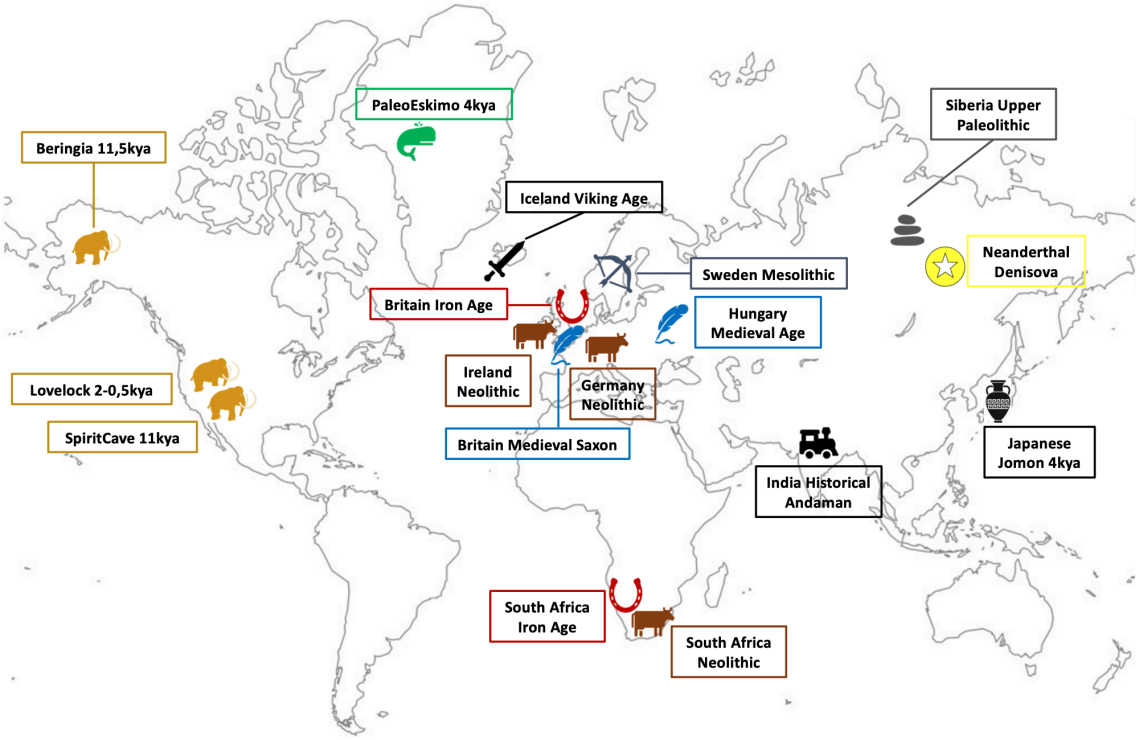



**Figure 3.1.** Temporal and geographical distribution of the 35 ancient DNA samples analyzed. Samples are worldwide distributed (except for Oceania) and cover almost every prehistoric and historic period from the Upper Paleolithic.

C to T (and G to A) damages are known to increase towards the reads' ends when mapped to a reference genome (Briggs et al., 2007) and must be determined before every aDNA analysis. Therefore, the 35 ancient genomes were first analyzed with mapDamage (Jónsson et al., 2013) to define the C to T damage.

After inferring nucleotides misincorporations, the function *trimBam* of bamUtil (Jun et al., 2015) was used to trim reads from *bam* files, which have been previously mapped to the human reference genome hs37d5 (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/). The following criteria were applied: two (2) bases were trimmed when the first two (2) C to T misincorporations were < 5% or the first C to T > 5% and second C to T << 5%; five (5) bases were trimmed when the C to T damage was > 5%.

After converting the *bam* files to the *fastq* format (as required by the software), ancient genomes were analyzed with *ngs te mapper 2* (Han et al., 2021), which is able to identify insertions in both single- and double-stranded DNA.

The individual *bed* output files provided by *ngs te mapper 2* were joined using a self-customized python script that outputted a single *vcf* file: information about the presence of the insertions were included in the final output.

The two archaic individuals (Altai Neanderthal and Denisova), plus the more ancient Chagyrskaya Neanderthal (80 kya, Paleolithic - Mafessoni et al., 2020), and five ancient *Homo sapiens* samples were further analyzed with the Mobile Element Locator Tool (MELT) v.2.2.2 (Gardner et al., 2017), in particular: "Ötzi the Iceman" (Italian Bronze Age - Keller et al., 2012), "Chan" (Iberian Mesolithic - Gonzàles-Fortes et al., 2017), "GB1" (Romanian Eneolithic - Gonzàles-Fortes et al., 2017) and "Kostenki14" (late Russian Pleistocene - Seguin-Orlando et al., 2014), along with the previously studied "UstIshim" from the Siberian Upper Paleolithic (Fu et al., 2014).

As a preliminary step, *fastq* files of the eight individuals were treated with AdapterRemoval (Lindgreen, 2012) using default parameters; then, *fastq* files were aligned to the human reference genome hs37d5 (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/) using BWA-mem (Li and Durbin, 2009) and resulting *bam* files were handled with samtools *fixmate* and samtools *sort* (Li et al., 2009; Danecek et al., 2021).

Genomes were scanned with MELT-DEL (Gardner et al., 2017) to identify and characterize reference polymorphic TEs in the eight ancient *Homo* samples; then, TypeTe (Goubert et al., 2020) was used to perform a more accurate genotyping of reference Alus. Finally, MELT-Split with adjusted parameters (-c 30 in the Individual Analysis step and -cov 5 in the Group Analysis step) was used to search for non-reference polymorphic TEs. Only "PASS" sites (i.e., variants that passed all quality controls) were retained for further investigations and collected in a single *vcf* file.

For the genomes analyzed with both methods (Altai Neanderthal, Denisova and UstIshim), results were compared to each other to evaluate the congruence of mobile elements call of the two softwares.

Results from both analyses were also compared with the list of structural variants of the 1000 Genomes Project dataset (Gardner et al., 2017), aligned to the human reference genome hs37d5, to retrieve information about the presence/absence and allele frequency (AF) of the identified variants in modern human populations and macro-areas of the world. The 1KGP dataset includes 2,504 individuals from 26 worldwide populations, divided in five macro-areas (Africa, America, Europe, East Asia and South Asia). The comparison was performed using a self-customized python script and searching for insertions in the same position or within a range of ± 10 nucleotide bases.

For *ngs te mapper 2*, using a self-customized python script and after downloading the human genes annotation in *bed* format from the UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgTables), we also determined whether a reference TE falls into a genic or a non-genic region. Instead, MELT-Split provides vcf files already including information about location of the identified non-reference TE, comprising not only gene name in RefSeq format, but also the exact position of the variant: intronic, exon, promoter, terminator, 3' UTR and 5' UTR.
Then, we converted RefSeq ID of the genes to the Official Gene Symbol with the software DAVID (Database for Annotation, Visualization and Integrated Discovery, https://david.ncifcrf.gov) (Huang et al., 2009; Sherman et al., 2022), and after the conversion, we performed an overrepresentation test with Panther (http://www.pantherdb.org).

Finally, the most significant results have been cross-checked with the lists of expression and alternative splicing quantitative trait loci (eQTL/sQTL) retrieved from the work of Cao and colleagues (2020), to search for a possible functional role of the identified TEs.

**Results**

After running *ngs te mapper 2*, a total of 146,564 reference TEs have been searched in the 35 samples and the most represented families/subfamilies are L1PA and AluS. Since we are particularly interested in polymorphic sites, we then compared our set of reference TEs with the structural variants found in the 1KGP dataset. Only 150 polymorphic TEs were in common, suggesting that the vast majority of

mobile elements retrieved by *ngs te mapper 2* are monomorphic or not detectable due to missing data: therefore, we focused on the analysis of those 150 polymorphic TEs. Interestingly, 10 TEs were located inside genes related to variations (meaning that variants of those genes influence positively or negatively the phenotype) in body height, which changed considerably during human evolution (Perkins et al., 2016; Rosenstock et al., 2019) and were present also in the 1KGP dataset (Gardner et al., 2017). Two TEs were considered as putatively more significant than others: the AluYb8 on chr10:33492650 and the AluYa5 on chr11:16223168, both identified in Altai Neanderthal by *ngs te mapper 2*. Notably, the AluYb8 is located in an intron of the gene Neuropilin1 (NRP1), whose variants are associated with changes in human height (Yengo et al., 2022): moreover, the TE acts as sQTL in various tissues, such as thyroid, adipose subcutaneous and adipose visceral omentum (Cao et al., 2020). Its allele frequencies (AF) in modern human populations from 1KGP vary across macro-areas: the Alu is more frequent in East Asia (EAS) with AF = 0.44, while it diminishes moving towards South Asia (SAS) = 0.30, Europe (EUR) = 0.20 and Africa (AFR) = 0.08. The Alu on chromosome 11 is located in an intron of the gene SOX6 (Sex Determining Region Y-Box 6), which is a "transcription factor that plays a key role in several developmental processes, including neurogenesis, chondrocytes differentiation and cartilage formation" (The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. Stelzer et al., 2016 - www.genecards.org), and is involved in skeletogenesis (Lefebvre, 2019). This Alu is present mostly in East Asia (AF = 0.25), and again its frequency diminishes moving west and south: South Asia = 0.03, Europe = 0.01 and Africa = 0.05. For both Alus, the allele frequency is higher in America than in Europe/South Asia/Africa (AluYb8 on chromosome 10 = 0.35, AluYa5 on chromosome 11 = 0.11). Furthermore, two additional TEs falling in genes related to body height variations resulted as of interest: the L1HS on chr7:134350004 in the gene BPGM (Bisphosphoglycerate Mutase) (Raghavan et al., 2022; Yengo et al., 2022; Schoeler et al., 2023) and the AluYa5 on chr9:35803105 in the gene NPR2 (Natriuretic Peptide Receptor 2) (Akiyama et al., 2019; Barton et al., 2021; Yengo et al., 2022). Both elements are widely present in modern human populations: the L1HS is more frequent in East Asia (AF = 0.69), then the allele frequencies diminish as we move to Africa (0.66), South Asia (0.64), America (0.51)

and Europe (0.43). The AluYa5 is more present in Africa (0.41), then its allele frequencies drop down to 0.34 in East Asia, 0.21 in America, 0.18 in Europe and 0.15 in South Asia.

We also looked at the other 140 polymorphic reference TEs, that are intergenic or located in genes related to hematological disorders - according to the GAD database (raw p-value = 6,1 * $E^{-3}$ and FDR = 6,8 * $E^{-2}$) - such as: AOPEP, C2, CDA, EYS, KLHL1, MARCHF1, PDE1C, PTPRD, SPATA13 and ZDHHC14. Moreover, 14 genes are involved in chem-dependency conditions: ABCA6, PRIM2, ADGRL3, ALDH1A2, AOPEP, CHN1, COMMD1, EYS, FBN1, KLHL1, PDE1C, KCNQ5, PTGFR and PTPRD, of which five are involved also in hematological disorders.

The second step was the analysis on eight high-coverage genomes with the MELT software (Gardner et al., 2017). A total of 2,030 reference polymorphic TEs (1,859 Alus, 150 LINE-1s and 21 SVAs) were detected; using the same approach described for *ngs te mapper 2*, we retrieved information about insertions' location: 1,056 (52%) are intergenic and the other 974 (48%) are located in known genes. Then, we performed an over-representation test with the software Panther (http://www.pantherdb.org) on the identified genes: interestingly, by looking at overrepresented pathways, "Neurexins and neuroligins" (raw p-value = 2.7 * $E^{-8}$ and FDR = 3.48 * $E^{-5}$) and "Protein-protein interactions at synapses" (raw p-value = 5.49 * $E^{-9}$ and FDR = 1.41 * $E^{-5}$) emerged as the most significant results. On the other hand, by looking at the over- or under- represented biological processes, "glutamate receptor signaling pathway" is overrepresented with a raw p-value = 3.55 * $E^{-6}$ and FDR = 0.00391, while "immune system process" (raw p-value = 0.000609 and FDR = 0.048) and "immune response" (raw p-value = 0.000104 and FDR = 0.016) are significantly underrepresented. Coherently, "non-motor actin binding protein" (raw p-value = 0.000703 and FDR = 0.023) is the most significant overrepresented protein class and "defense/immunity protein" is the most significant underrepresented protein class, with a raw p-value = 9.89 * $E^{-5}$ and FDR = 0.00969. In addition, the molecular function "glutamate receptor activity" is the only overrepresented significant result in this category, with a raw p-value = 1.52 * $E^{-5}$ and FDR = 0.00907.

We also compared polymorphic reference insertions identified in Altai Neanderthal, Denisova and UstIshim (the only three samples in common between the two analyses) with *ngs te mapper 2* (Han et al., 2021) and MELT-DEL (Gardner et al., 2017). A total of 198 polymorphic TEs were in common, but the only fully coherent result (meaning that the element's call is "present" in the same individual according to both softwares) was an intergenic Alu on chr13:57967126, identified in the same position by the two softwares: the Alu is present in all three samples according to both *ngs te mapper 2* and MELT.

Other 63 TEs show coherent results for the two softwares in at least one sample (insertions highlighted in bold in Supplementary Table S2), in particular: Altai Neanderthal (41), Denisova (9) and UstIshim (9), plus four elements shared between two of the samples.

Finally, using MELT-Split as described in Materials and Methods, we identified 449 non-reference polymorphic TEs in the eight ancient *Homo* samples: 399 Alus, 39 LINE-1s and 11 SVAs. Of these, 238 are intergenic, 180 are intronic, 14 in promoters, 12 in terminators, three at 3' UTRs, one at 5' UTR and one in the twelfth exon of the gene N4BP2 (NEDD4 Binding Protein 2), whose SNPs are related to variations in body height (Yengo et al., 2022). The transposable element is an AluSg located on chromosome 4:40127920 and has been identified in Altai Neanderthal.

We converted the RefSeq IDs of the genes mapped by non-reference TEs to the Official Gene Symbol with DAVID: the most significant Reactome pathways (https://reactome.org/) are "neuronal system" (raw p-value = $7.7 * E^{-4}$) and "Diseases of signal transduction by growth factor receptors and second messengers" (raw p-value = $5.8 * E^{-3}$). After the conversion, we performed an overrepresentation test with Panther: the most significant cellular components are "postsynaptic membrane" (raw p-value = $1.46 * E^{-6}$ and FDR = $3.58 * E^{-4}$) and "presynaptic membrane" (raw p-value = $3.89 * E^{-4}$ and FDR = $1.91 * E^{-2}$); the other categories (molecular function, biological process, protein classes) returned non-significant results.

**Discussion**

As we introduced in the background section, technological and methodological issues make it difficult to analyze the TE content of ancient genomes (see for example Wang et al., 2021): in fact, to our knowledge, only two studies (Gardner et al., 2017; Guichard et al., 2018) have relied on transposable elements to decipher the genetic variability of ancient humans, but they focused only on archaic individuals: Altai Neanderthal and Denisova. To fill this gap, as a first step we collected 35 samples and applied the *ngs te mapper 2* software to detect transposable elements in ancient *Homo sapiens* samples and in archaic hominins. We identified more than 146 thousand reference TEs, but even when searching for insertions within a range of ± 10 nucleotide bases, only 150 were in common between *ngs te mapper 2* and the list of reference polymorphic TEs provided by MELT in the 1000 Genomes Project (Gardner et al., 2017). This is possibly due to the following facts: 1) *ngs te mapper 2* requires *fastq* files, while MELT analyzes *bam* files aligned to a reference genome; 2) *ngs te mapper 2* was applied on single-end reads, while MELT needs paired-end sequences aligned to a reference genome with bwa-mem; 3) the two softwares apply different methods to search for TEs, thus the two tools may indeed find different insertions; 4) MELT looks only for polymorphic TEs, which are a subgroup of all the insertions in the genome.

Therefore, we decided to run also MELT v.2.2.2 (Gardner et al., 2017) on eight ancient *Homo* samples, including Altai Neanderthal, Denisova and UstIshim, that have been analyzed also with *ngs te mapper 2*. MELT identified a total of 2,030 reference TEs: of these, 198 are in common with *ngs te mapper 2* (Supplementary Table S2): thus, these elements can be considered truly polymorphic (as we discussed earlier, MELT only looks for polymorphic TEs). Then, we compared the tools outputs for the same individual, and found that 64 elements have been indeed coherently identified by the softwares, meaning that the insertion call is "present" in the same individual according to both tools. For the other 134 TEs it is not possible to compare the results, in particular because of missing data in the analyzed *bam* files (i.e., "./." according to MELT) and because *ngs te mapper 2* outputs only information about the presence of the mobile element and its genomic position, without giving information about its absence or missing data in the analyzed individual (see elements with a "?" in the "ngs" columns of Supplementary Table S2).

MELT was used to search also for non-reference polymorphic TEs: a total of 449 elements have been identified in the eight ancient *Homo* samples analyzed. Interestingly, one is in the twelfth exon of the gene N4BP2 (NEDD4 Binding Protein 2), whose variants are related to variations in body height (Yengo et al., 2022). The transposable element is an AluSg located on chromosome 4:40127920 and has been identified in Altai Neanderthal.

Other genes act positively or negatively on human body height: for example, ten out of 150 reference TEs common between *ngs te mapper 2* and the modern human populations of the 1KGP. Of these, two can be considered noteworthy:

1) the AluYb8 on chr10:33492650 identified in Altai Neanderthal by *ngs te mapper 2*. The element is located in an intron of the gene Neuropilin1 (NRP1), which variants are associated with changes in human height (Yengo et al., 2022): moreover, the TE acts as alternative splicing Quantitative Trait Loci (sQTL) in various tissues, such as thyroid, adipose subcutaneous and adipose visceral omentum (Cao et al., 2020). The Alu is more frequent in East Asia with AF = 0.44, while it diminishes moving towards America (0.35), South Asia (0.30), Europe (0.20) and Africa (0.08).
2) The AluYa5 on chr11:16223168, identified in Altai Neanderthal and located in an intron of the gene SOX6 (Sex Determining Region Y-Box 6), which is a "transcription factor that plays a key role in several developmental processes, including neurogenesis, chondrocytes differentiation and cartilage formation" (Stelzer et al., 2016 www.genecards.org), and is involved in skeletogenesis (Lefebvre, 2019). This Alu is present mostly in East Asia (AF = 0.25) and America (AF = 0.11), and again its frequency diminishes moving west and south: South Asia = 0.03, Europe = 0.01 and Africa = 0.05.

Another TE can be considered as of interest: the AluYa5 on chr9:35803105 in the gene NPR2 (Natriuretic Peptide Receptor 2), whose variants are indicative of a short stature (Akiyama et al., 2019; Barton et al., 2021; Yengo et al., 2022). The element has been identified in three Vikings from Iceland and in one Mesolithic individual from Sweden; it is widely present in modern human populations, but with different allele frequencies with respect to the previously discussed Alus:

indeed, the AluYa5 is more present in Africa (0.41), then its allele frequencies drop down to 0.34 in East Asia, 0.21 in America, 0.18 in Europe and 0.15 in South Asia. The higher frequency observed in Africa - with respect to the other discussed Alus - could be indicative of a more ancient origin of this mobile element. When looking at the modern Icelandic population (Beyter et al., 2021), an allele frequency = 0.19 can be observed, coherent with that of modern Europeans from 1KGP. Moreover, the Alu also acts as sQTL in various tissues, including but not limited to: muscle skeletal, pituitary, thyroid, adipose subcutaneous and adipose visceral omentum (Cao et al., 2020), suggesting a functional role for this element on NPR2.

Finally, we analyzed with the softwares DAVID and Panther all the genes mapped by transposable elements identified in the three different runs (one with *ngs te mapper 2* and two with MELT v.2.2.2). Notably, reference polymorphic TEs map in genes involved in: 1) "Metabolism of lipids" (raw p-value = 7.86 * $E^{-5}$ and FDR = 8.52 * $E^{-3}$); 2) "Neurexins and neuroligins" (raw p-value = 2.7 * $E^{-8}$ and FDR = 3.48 * $E^{-5}$) and "Protein-protein interactions at synapses" (raw p-value = 5.49 * $E^{-9}$ and FDR = 1.41 * $E^{-5}$). Coherently, "neuronal system" (raw p-value = 7.7 * $E^{-4}$) and "Diseases of signal transduction by growth factor receptors and second messengers" (raw p-value = 5.8 * $E^{-3}$) are the most represented pathways (these genes are mapped by non-reference TEs). Therefore, we can speculate that structural variants such as retrotransposons during human evolution acted on phenotypes such as body height and body composition (with respect to the results "Metabolism of lipids" and the 10 TEs located in genes related to body height variations, some of which have also a functional role according to Cao et al., 2020), but also had an important role in shaping neuronal functions and central nervous system development, as suggested also by Guichard and colleagues (2018).

In conclusion, the detection of polymorphic TE in ancient DNA is only at its beginning due to severe technical issues. However, the obtained results may help to reconstruct the evolutionary history of our species, particularly when compared with TEs modern variability. Indeed, several clues point towards the fact that TEs had an important role in shaping human genome evolution and variability, and

phenotypes such as body height, which changed considerably during human history (Perkins et al., 2016; Rosenstock et al., 2019), could have been influenced by the action of structural variants.

# 4. Modern human adaptations and phenotypes explained by polymorphic transposable elements

## 4a. Human adaptation to high altitude in the Tibetan Plateau: the possible role of polymorphic transposable elements

Giorgia Modenini[1,#], Paolo Abondio[2], Marco Sazzini[1,3], Alessio Boattini[1]. *Polymorphic transposable elements provide new insights on high-altitude adaptation in the Tibetan Plateau*. Genomics. 2024 May 1;116(3):110854. doi: 10.1016/j.ygeno.2024.110854

Affiliations:
1: BiGeA Department, University of Bologna, Bologna, Italy
2: IRCCS Istituto Delle Scienze Neurologiche Di Bologna, Bologna, Italy
3: Interdepartmental Centre - Alma Mater Research Institute on Global Changes and Climate Change, University of Bologna, Italy
#: correspondence to giorgia.modenini2@unibo.it

### Background

The peopling of the Tibetan Plateau, with an average elevation of 4,000 meters, by the ancestors of Tibetan and Sherpa highlanders, is one of the most compelling examples of Anatomically Modern Humans (AMH) adapting to a new and extreme environment (Beall et al. 2010; Simonson et al. 2010; Yi et al. 2010). The establishment of stable high-altitude settlements by the ancestors of Tibetan and Sherpa populations seems to have occurred only after the Last Glacial Maximum (Aldenderfer 2011). Moreover, different studies suggested that more recent instances of migrations, admixture and geographical/cultural isolation could have further influenced the genetic variation of present-day Tibetan and Sherpa groups (Lu et al. 2016; Li et al. 2008).

In the last decades, several population genomics and genome-wide association studies (GWAS) (Beall et al. 2010; Simonson et al. 2010; Yi et al. 2010; Xu et al. 2011; Jeong et al. 2014; Hu et al. 2017; Yang et al. 2017) tried to disentangle the genetic basis of high altitude adaptation (HAA), but so far, most of the studies have relied on single nucleotide polymorphisms (SNPs) to search for evidence of natural selection in Tibetan and Sherpa populations. Two genes, related to the hypoxia-inducible transcription factor (HIF) pathway, have been identified as under positive selection in these populations: EPAS1 (endothelial PAS domain protein 1) and EGLN1 (egl-9 family hypoxia-inducible factor 1) (Beall et al. 2010; Xiang et al. 2013). Furthermore, it has been demonstrated that EPAS1

carries signals of adaptive introgression from Denisovan archaic hominins (Huerta-Sánchez et al. 2014; Zhang et al. 2021), who admixed with the ancestral population of both modern high-altitude and low-altitude East Asians.

Structural variation (SV) is an essential mutational force shaping the evolution and function of the human genome (Quan et al. 2021). However, few studies (Lou et al. 2015; Quan et al. 2021) have analyzed the link between SVs and HAA and, to our knowledge, no one has focused on retrotransposons to date.

Retrotransposons are mobile genetic elements with the ability to replicate themselves and increase the number of their copies: indeed, sequences from retrotransposons constitute at least 40% of the human genome (de Koning et al. 2011). These elements are divided into long terminal repeats (LTR) retrotransposons, to which the human endogenous retrovirus (HERV) family belongs, and non-LTR retrotransposons, represented by short interspersed nuclear elements (SINEs, such as Alu-like elements, ~300 bases long), long interspersed nuclear elements (LINEs, complete elements are ~6 kilobases long, but frequently they are shorter due to 5' truncation during insertion) and the composite family of SINE-VNTR-Alu (SVAs of variable length because of the presence of a Variable Number Tandem Repeat [VNTR] region). Among non-LTR retrotransposons, only LINE-1s are autonomously active (Goodier 2016), while Alus and SVAs rely on LINE-1's machinery to mobilize themselves (Guio and Gonzalez 2019).

Transposable Elements (TEs) have been an important source of genetic variation throughout human evolution, many of them being polymorphic and showing population-specific stratification (Wang et al. 2021). Accordingly, they are valuable genetic markers for the study of human populations variability, as shown by several recent works (Rishishwar et al. 2015; Gardner et al. 2017; Watkins et al. 2020). Therefore, we decided to focus on the study of polymorphic TEs to disentangle the genetic basis of high-altitude adaptation in the Tibetan Plateau.

Here we provide the first large-scale study on 114 high-coverage published genomes from the Tibetan Plateau and East Asia in order to assess the role of polymorphic TEs in the HAA of populations settled along the Himalayan Arc. Together with modern genomes, we analyzed four high-coverage ancient and archaic DNA samples (two ancient Tibetans and two archaic hominins, namely Altai

Neanderthal and Denisova) to provide a temporal context for the most significant results emerging from modern genomes analyses.

**Materials and Methods**

114 published modern high-coverage genomes (30x-45x) from high-altitude (HA), middle-altitude (MA) and low-altitude (LA) populations of East Asia and one African population (Yoruba) were included in this study, along with four high-coverage ancient/archaic DNA samples: one from a Denisovan individual (Meyer et al. 2012), one from a Neanderthal individual (Prüfer et al. 2014) and two "ancient Tibetans" (Jeong et al. 2016) (C1 and S10), spanning 3,150–1,250 years before present (yBP) (Supplementary Table S1 and Figure 1). The 114 modern samples are represented by: 10 Sherpa (Jeong et al. 2014; Lu et al. 2016; Mallick et al. 2016; Gnecchi-Ruscone et al. 2018), 28 Tibetans (Lu et al. 2016); 8 Tujia, 8 Yi, 6 Naxi, 29 Han Chinese from the Human Genome Diversity Project (HGDP) (Bergström et al. 2020) and 25 Yoruba from the 1000 Genomes Project (1KGP) phase 3 (1000 Genomes Project Consortium et al. 2015). Individuals were selected avoiding relatives. These samples were chosen to represent the modern genome variability of the Tibetan Plateau and East Asia. The African population was selected as an outgroup. The ancient samples were included to analyze the evolution of the HA populations of the Tibetan Plateau by searching for common variants between modern and ancient DNA samples.

Original bam files were first converted to the *fastq* format with the *bedtools* command *bamtofastq* (Quinlan and Hall 2010). All *fastq* files were then treated with AdapterRemoval (Lindgreen 2012) and subsequently aligned to the human reference genome GRCh38dh (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/GRCh38_reference_genome/) with bwa-mem (Li and Durbin 2009). The resulting *sam* files were processed with samtools fixmate to clean up read pairing information and flags and then sorted with samtools sort (Li et al. 2009). The obtained *bam* files were then indexed and processed with MarkDuplicates (see Picard: http://broadinstitute.github.io/picard).

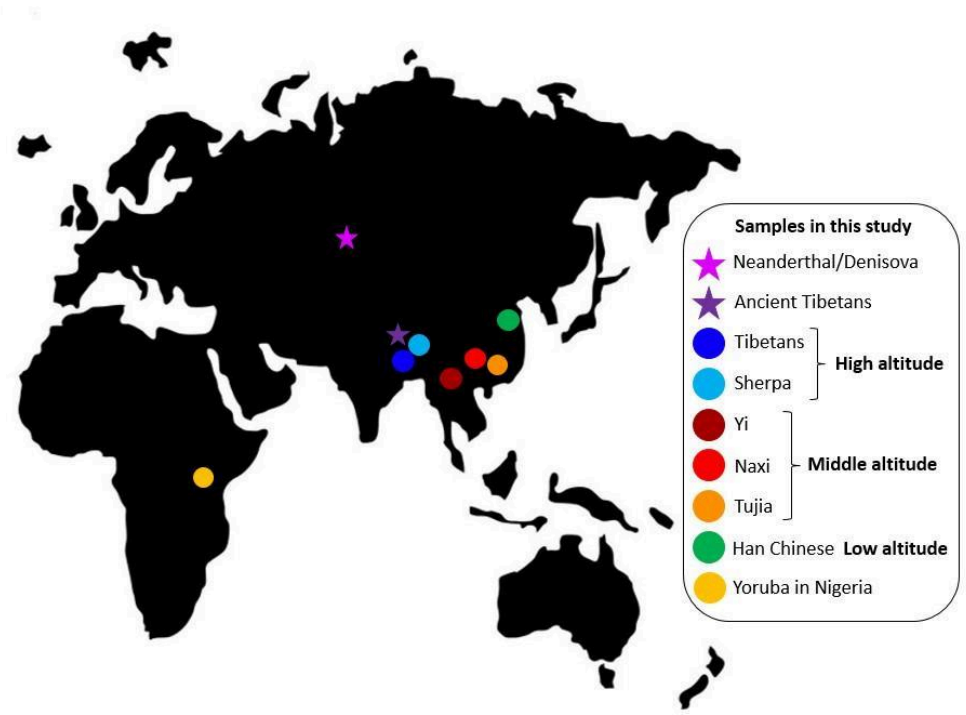



**Figure 4a.1**. Location of the modern and ancient samples analyzed in this study.

The identification of polymorphic non-reference TEs (Alu, LINE-1, SVA and HERV-K) was performed on both ancient and modern samples using the software MELT v2.2.2 with the function 'Split' (Gardner et al. 2017). MELT-DEL was applied to retrieve information about polymorphic reference TEs (Alu, LINE-1 and SVA) and TypeTE (Goubert et al. 2020) was then used to better genotype reference Alus. Only "PASS" sites were included in a single final VCF file. Fisher tests of independence with one and two degrees of freedom were performed to identify TEs with significantly different frequencies in HA compared to MA-LA populations. Tests were performed for both allelic and genotype frequencies. TEs that yielded significant results (nominal p-value < 0.01) at least for allele frequencies ("differentiated TEs") were considered as putatively contributing to Tibetan and Sherpa differentiation.

To assess the genetic relationships among the individuals included in our dataset, as well as their shared ancestry, a PCA and ADMIXTURE analysis (Alexander, Novembre and Lange, 2009) were performed on the TEs dataset. Quality control (QC) was performed with the PLINK software (Purcell

et al., 2007) on modern samples, including the removal of genetic elements belonging to sexual chromosomes, a check for the proportion of missing data (using the commands --geno 0.01 and --mind 0.01), the respect of Hardy-Weinberg equilibrium after Bonferroni correction for multiple testing (--hwe 0.01/α, where α is equal to the number of variants remaining in the dataset at this stage of the QC procedure), the removal of rare variants (--maf 0.01) and an assessment of linkage disequilibrium along the genome, using a sliding window of 500 bp, a moving step of 50 bp and a threshold value of 0.1 (--indep-pairwise 500 50 0.1). After QC, PCA was performed on the TEs dataset by applying file format conversions as provided by the *convertf* and *smartpca* tools from the EIGENSOFT package v6.0.1 (Price et al., 2006).

Similarly, the ADMIXTURE software (Alexander, Novembre and Lange, 2009) was employed to perform an estimation of shared genetic ancestry across populations. A number K of putative ancestral components between 2 and 12 was tested, and 50 iterations of each run were performed to minimize the error and maximize the log-likelihood of each ancestry estimate.

To measure population differentiation due to genetic structure, the fixation index (Fst) was computed for all TEs used to perform PCA and Admixture analyses (Weir and Cockerham 1987). For the purpose of computing their genetic distance, three groups of modern individuals with shared genetic ancestry were extracted from our samples: high-altitude Tibetans and Sherpas (HA, 38 individuals), middle-altitude Tibeto-Burman speaking (MA, 22 individuals) and low-altitude Han Chinese (LA, 29 individuals). Fst was computed for all three population pairs (HA/MA; HA/LA; MA/LA) and each distribution was independently standardized by subtracting the average Fst from each score, then dividing the obtained value by the standard deviation of the distribution. Only absolute normalized Fst scores greater than 2 were considered significant for further inquiry. Significant Fst scores distinguishing HA from both MA and LA were detected and, to obtain signals explicitly related to HA differentiation, TEs with significant Fst scores in the MA/LA comparison were removed from the previous result.

In addition, Population Branch Statistics (PBS) was computed to corroborate Fst observations and infer the directionality of differentiation, with the African Yoruba group (25 individuals) added to

explore results coming from an outgroup of different ancestry (Xin Yi et al. 2010). PBS was computed for the following trios (where the third population is the outgroup): HA/MA/Yoruba; HA/LA/Yoruba; HA/MA/LA; MA/LA/Yoruba. The top 0.1% of PBS scores for each trio was deemed significant; moreover, TEs emerging from the last trio (middle altitude/low altitude/Yoruba) were used to filter out signals otherwise confirmed by the other three, so as to highlight genetic signatures characterizing the high altitude Tibetan and Sherpa groups only.

Information about the position of all detected non-reference TEs was retrieved from the MELT output, which includes the location of the TE ("null", when the TE is in an non-genic area; "intronic", "exon", "3_UTR", "5_UTR", "Promoter", "Terminator" when the TE is in a genic or regulatory region) and the gene name, if any, in RefSeq format. On the other hand, information about reference TEs location was retrieved using a self-customized python script after downloading genes annotation in bed format from the UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgTables) and using as a reference the genomic locations identified by the following works: Mignone et al. 2002 and Dvorak et al. 2020 for 3'UTRs and 5'UTRs (1,000 bp from gene end and 210 bp from gene start, respectively); Kim et al. 2005 for promoters (500 bp from 5' UTRs); West & Proudfoot 2009 for terminators (between 250 and 1,050 bp after 3' UTRs). All RefSeq IDs were then converted into the Official Gene Symbol with the software DAVID (Database for Annotation, Visualization and Integrated Discovery; https://david.ncifcrf.gov/conversion.jsp) (Huang et al. 2009; Sherman et al. 2022) and used for further investigations. We also looked for the relevant diseases potentially related to the detected sets of genes using DAVID, which refers to the GAD database (Becker et al. 2004). The statistical overrepresentation test was performed with Panther (http://pantherdb.org/), selecting the organism Homo sapiens and calculating false discovery rate (FDR) and Fisher test. The analysis of Gene Network and Pathways was done with the Kyoto Encyclopedia of Genes and Genomes, KEGG (https://www.genome.jp/kegg/mapper/search.html), selecting the organism Homo sapiens ("hsa"). Genes mapped by significant TEs were also compared with literature lists of genes deemed under positive selection in Tibetans (Zheng et al. 2023) and candidate genes for polygenic adaptation in Tibetan and Sherpa populations (Gnecchi-Ruscone et al. 2018).

After the identification of candidate TEs that possibly contributed to HAA, we checked whether these TEs were also present in the four considered archaic (Neanderthal and Denisova) and ancient (C1 and S10) individuals, by inspecting the MELT output.

Finally, we inferred a possible function for the candidate TEs by cross-checking our results with those provided by Cao and colleagues (Cao et al. 2020), who identified a list of TEs that act as expression/alternative splicing Quantitative Trait Loci (eQTL/sQTL).

An association test using "high altitude" (1) and "low altitude" (0) as binary pseudo-phenotypes was computed with the software GEMMA (Zhou and Stephens 2012) on the whole variants dataset. Only individuals from HA (Tibetans and Sherpa) and MA+LA populations (Tibeto-Burman speaking populations and Han Chinese samples grouped together) were included in the analysis. Following instructions from the manual, we first calculated the relatedness matrix with the command -gk 2, meaning that the software calculated a standardized (rather than centered) relatedness matrix. Then, Wald's test was performed applying a linear model (-lm 1) on the previously estimated matrix. Only TEs with an adjusted p-value < 0.001 (after Benjamini-Hochberg correction) were considered as significantly associated with high altitude ("associated TEs"). Results were plotted using the R package "CMplot" (Yin et al. 2021: https://github.com/YinLiLin/CMplot).

**Results**

After using MELT-Split on 118 ancient and modern samples, we successfully identified 9,144 polymorphic non-reference TEs, of which 7,438 Alu, 1,193 LINE-1, 492 SVA, 21 HERV-K, and 3,754 polymorphic reference TEs, of which 3,492 Alu, 169 LINE-1, 93 SVA. Based on the information provided by the MELT output, 49.4% of non-reference TEs are in non-genic regions ("null"), while 50.6% are in genic regions (42.1% are in introns, 3.8% in promoters, 3.9% in terminators, 0.6% in 3' UnTranslated Regions (UTRs), 0.1% in 5' UTRs and 0.1% in exons). Moreover, when looking at reference TEs, 54.7% are intergenic, 43.4% are intronic, 0.2% in exons, 0.6% at 3' UTRs, 0.05% at 5' UTRs, 0.15% in promoters and 0.9% in terminators. 56.1% of reference TEs are into genes, while 43.9% are in intergenic regions. Finally, reference and non-reference TEs

feature systematic differences in allele frequencies, in particular reference TEs are enriched for higher allele frequencies (> 0.5) while non-reference TEs are enriched in lower allele frequencies (< 0.5). In general, this is due to the fact that ref TEs are identified based on their presence on a "single" genome (the reference) while non-ref TEs are detected based on many genomes (i.e. all the analyzed genomes).

We considered as putatively related with HAA only TEs that yielded significant Fisher tests (p-value < 0.01) based on allele frequencies. Accordingly, when comparing HA (Sherpa and Tibetan) and MA-LA (Tibeto-Burman speaking and Han Chinese) populations we detected 271 significant TEs (154 non-reference TEs and 117 reference TEs).

Obtained results from PCA and Admixture (Figures 4a.2 A and B) confirm a clear distinction between African and non-African populations, as shown by the first principal component (PC1) in the PCA and by a complete separation of ancestral components between the Yoruba group ("yellow") and the individuals of Asian ancestry ("red" and "green") at K=3, which is the number of components with the lowest CV error (0.44385). The second principal component (PC2) highlights a geographic distinction between HA and LA individuals, which are distributed vertically along a high-to-low altitude gradient including Tibetans and Sherpa at the top, then the Tibeto-Burman speaking groups (Naxi, Yi and Tujia), and finally the Han Chinese representatives at the bottom. The same gradient can be observed in the Admixture graph at K=3, where the "green" component is predominant in HA Tibetans and Sherpa, while the "red" component is characteristic of the LA Han Chinese individuals; the Tibeto-Burman speaking groups (MA) show variable combinations of "red" and "green" components, with an increasingly higher proportion of Tibetan-like ancestry for the Tujia, Yi and Naxi groups, respectively. The Ancient Tibetan samples clearly overlap with Tibetans and Sherpas representatives in the PCA, while they carry the "green" HA ancestral component in the Admixture plot at K=3, confirming their tight relationship with the modern HA groups.

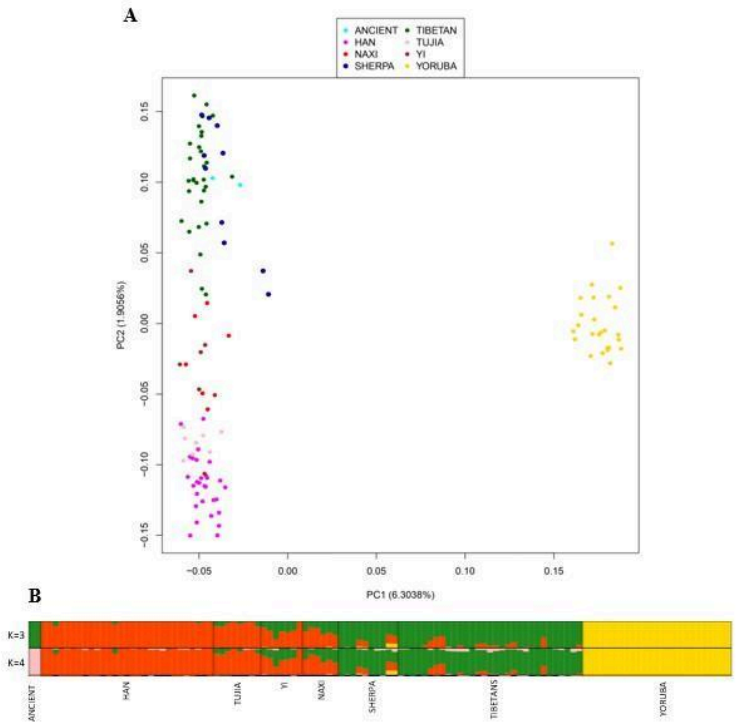



**Figure 4a.2**. PCA and Admixture plots. **A)** In the PCA, the first PC clearly discriminates between African and non-African groups, while PC2 highlights a high-to-low altitude gradient with Tibetans and Sherpas at the top and Han Chinese at the bottom. **B)** Admixture plots showing results for K=3 (lowest CV error = 0.44385) and K=4, where the two Ancient Tibetans (C1 and S10) carry their own ancestry component ("pink").

Population differentiation has been evaluated by computing the fixation index, Fst, for all TEs in our panel and by comparing the high-altitude (HA, Tibetan and Sherpa), Tibeto-Burman speaking (MA) and low-altitude (LA) groups in pairs.

Although all scores are relatively low, the HA populations (Tibetan+Sherpa) exhibit the highest differentiation (Fst=0.013) from the LA group (Han Chinese) and an intermediate level of differentiation with respect to the MA Tibeto-Burman speaking representatives (Fst=0.006). Similarly, these last two groups show the lowest mean Fst score (Fst=0.003).

After normalizing each Fst distribution, a total of 103 non-reference TEs significantly discriminate between the HA and MA groups, while 131 distinguish the HA and LA cohorts, and 115 characterize the comparison between MA and LA populations. By cross-referencing the significant scores for the three distributions, a total of 32 non-reference TEs emerge as able to discriminate between HA and

both MA and LA populations (see Supplementary Table S3). To further corroborate this finding, the resulting non-reference TEs were cross-checked with those having a significant Fisher test score for allele counts between the HA and non-HA groups: all non-reference TEs discriminating HA from both MA and LA show Fisher p-values < 0.01, confirming the significantly different presence of the transposable elements under scrutiny in the Tibetan and Sherpa groups with respect to the other two (see Supplementary Table S3).

To further define which TEs are preferentially differentiated in the HA group, rather than in the other two, three-way PBS distance analysis was carried out as described in the Materials and Methods section. Keeping into consideration the top 0.1% scores as significant among the performed tests and cross-referencing the results, a total of 62 non-reference TEs (58 Alus, 2 LINE-1s, 2 SVAs) appear to be characteristic of differentiation in the direction of the HA group (see Supplementary Table S4), with 22 falling into promoters or introns of known genes. When intersecting the PBS and the aforementioned 32 non-reference TEs with significant Fst discriminating HA from MA and LA groups, eight non-reference TEs are shared (highlighted in bold in the Supplementary Tables S3 and S4), two of them being located in the genes ASAH1-AS1 (acid ceramidase antisense RNA 1) and PHF21A (PHD finger protein 21A) on chromosomes 8 and 11, respectively.

After the identification of 271 "differentiated" TEs, we retrieved information about their location and type and found that 126 of them are located in genic regions (71 non-reference TEs + 55 reference TEs).

By analyzing the corresponding genes with the software DAVID, we observed that the most represented disease classes, according to the GAD database (Becker et al. 2004), are "cardiovascular" (p-value = $1.5*e^{-6}$), "hematological" (p-value = $1.4*e^{-6}$) and "chem-dependency" (p-value = $1.9*e^{-9}$). Indeed, the most significant disorders/affected traits (p-value < 0.001 after Bonferroni correction) are: Tobacco use disorder, Lipoproteins VLDL and Cholesterol LDL.

The association test with GEMMA on HA and MA-LA populations retrieved a total of 266 significant results ("associated" TEs: Adjusted p-value < 0.001, as highlighted by the red-dotted line in Figure

4a.3). Since 123 TEs are located in genic regions, we used DAVID to perform functional annotation clustering, and retrieved information about the relationship between those genes and diseases from the GAD database (Becker et al. 2004). Interestingly, the most significant conditions reflect the previously mentioned patterns, such as “Tobacco use disorder”, represented by 51 genes (p-value = $2.7*e^{-11}$). Other significant conditions were related to cardiovascular traits (i.e., “Blood Pressure”, “Erythrocyte Count”, “Heart Rate”, “Hemoglobin A Glycosylated”, “Glomerular Filtration Rate”), body measurements (“Body Mass Index”), “cholesterol LDL” and “Insulin”.

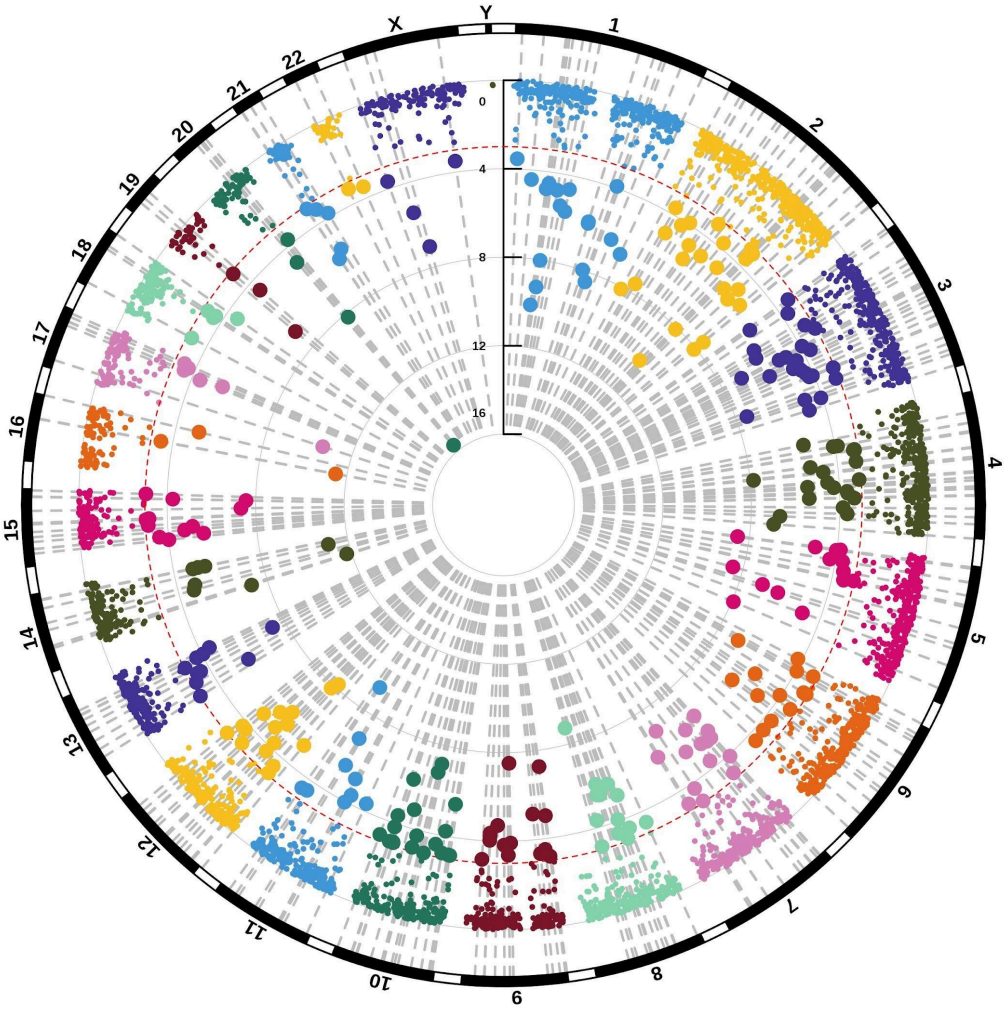


**Figure 4a.3.** Circular manhattan plot of the significant TEs associated with high altitude. TEs that yielded an adjusted p-value < 0.001 (red-dotted line) have a bigger size than the others. P-values are shown as Log10P. Plotting was performed with the "CMplot" R package (https://github.com/YinLiLin/CMplot).

After the identification of four sets of TEs potentially contributing to HAA in Tibetan and Sherpa populations (271 "differentiated" TEs; 266 "associated" TEs; 32 significant TEs for Fst and 62 for PBS), we compared these four lists: only one TE is a significant result for all tests, an AluY in a non-genic region; 11 TEs are shared among three tests out of four (Table 4a.1). Then, we verified their presence or absence in the two archaic hominins (Neanderthal and Denisova) and two ancient Tibetan individuals (C1 and S10). We also performed an in silico analysis to infer a possible function for the candidate TEs by cross-checking our results with those provided by Cao and colleagues (Cao et al. 2020). We also cross-checked the list of genes mapped by the four sets of significant TEs with those under positive selection in Zheng et al. (2023) and Deng et al. (2019) and found 10 genes in common. For instance, the gene SUPT3H is mapped by an AluYe that is significant for PBS and differential allele frequencies analyses: the gene is under positive selection in Tibetans according to Deng et al. (2019), who performed a composite of multiple signals (CMS) analysis. Moreover, by looking at the lists of candidate genes for polygenic adaptation in Tibetan and Sherpa populations (Gnecchi-Ruscone et al. 2018), it emerged that the same gene SUPT3H - and other five genes - are shared.

The most interesting TE, based on all our in silico analyses, is an AluYe on chromosome 12:44670594 in the gene NELL2 (Neural EGFL-like 2), which is involved in tobacco use disorder. The Alu has significantly different allele frequencies between HA and MA-LA populations (Fisher p-value = 0.00131) and is a significant result also for the association test performed with GEMMA (adjusted p-value = $1.81*e^{-4}$). Moreover, it is one of the significant results for PBS, with a score = 0.168776. By looking at the putative function of this Alu, it acts as eQTL in skin and as sQTL in brain cortex and putamen basal ganglia. Finally, it is present in Denisova and one ancient Tibetan (C1).

| Chr_pos | TE type | Location | Gene | Tests | eQTL | sQTL | Altai | Den | C1 | S10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 12_44670594 | INS-AluYe | intronic | NELL2 | 1, 2, 4 | X | X | 0 | 1 | 1 | NA |
| 14_64710808 | INS-AluY | intronic | PLEKHG3 | 1, 2, 4 | | | 0 | 0 | NA | 0 |
| 3_120524105 | INS-AluY | null | null | 1, 3, 4 | | | 0 | 0 | NA | 1 |
| 11_45940658 | INS-AluY | intronic | PHF21A | 1, 3, 4 | | | 0 | 0 | 1 | 0 |
| 13_28811284 | INS-AluY | null | null | 1, 3, 4 | | | 0 | 0 | 0 | 0 |
| 2_137868315 | INS-AluYg | null | null | 1, 3, 4 | | | 0 | 0 | NA | 0 |
| 5_144951450 | INS-L1Ta | null | null | 1, 3, 4 | | | 0 | 0 | 0 | 1 |
| 6_81458432 | INS-AluY | null | null | 1, 3, 4 | | | 0 | 0 | 1 | 1 |
| 8_18085627 | INS-AluYb8 | intronic | ASAH1-AS1 | 1, 3, 4 | X | X | NA | 0 | 0 | 0 |
| 9_100940618 | INS-AluYb3a1 | null | null | 1, 3, 4 | | | NA | 0 | 0 | 0 |
| 6_16950725 | INS-L1Ambig | null | null | 1, 2, 3 | X | | 0 | 0 | NA | 0 |

**Table 4a.1**. The 11 TEs shared among the four different tests (differential allele frequencies; association test with GEMMA; Fst; PBS). Chr_pos=chromosome + position; TE type=TE subfamily (INS=non-reference TE); Location=location of the TE ("null" = intergenic); Gene=gene in which is located the TE; Tests=tests for which the TE is significant (1=differential allele frequencies; 2=association test with GEMMA; 3=Fst; 4=PBS); eQTL/sQTL=if the polymorphic TE acts as eQTL and/or sQTL; Altai/Den/C1/S10=if the TE was found also in ancient or archaic individuals (Altai Neanderthal, Denisova, C1 and S10) (1=presence of the TE; 0=absence; NA=missing data).

**Discussion**

Different studies used TEs as genetic markers to study modern human variability and differentiation (Rishishwar et al. 2015; Gardner et al. 2017; Watkins et al. 2020). Therefore, we decided to investigate the possible influence of polymorphic TEs on the ad of East Asian populations exposed and adapted to an extreme environment, as represented by the high-altitude villages of the Tibetan Plateau inhabited by Tibetan and Sherpa ethnic groups.

As a first step, we performed two population genetics analyses, PCA and Admixture, on 118 ancient and modern individuals to contextualize their variability. Both of them confirm that retrotransposons are valuable genetic markers to study population differentiation: indeed, as shown in Figure 4a.2-A, there is a clear distinction between the African group (represented by Yorubas) and the East Asian populations in terms of principal components and ancestry components. Moreover, the PCA plot shows an altitudinal gradient between low/middle-landers (Han Chinese, Yi, Tujia and Naxi) and

high-landers (Tibetan and Sherpa). The Admixture graph (Figure 4a.2-B) identifies two main ancestry components: “green” for Tibetans and Sherpa, and “red” for Han Chinese, with the Tibeto-Burman speaking groups (Yi, Tujia and Naxi) showing a mixture of the two. Previous works based on SNPs (Jeong et al. 2014, 2016; Lu et al. 2016; Gnecchi-Ruscone et al. 2018) detected a similar pattern of variability in the same populations.

Fixation index (Fst) and PBS statistics computed using TEs confirm a subtle but tangible differentiation among the three analyzed groups, with an increase in average Fst following an altitudinal gradient: the HA cohort shows a higher Fst value when compared to the LA group (0.013) than the MA subjects (0.006). This is once again in line with what has emerged in previous works based on SNPs (Jeong et al. 2014, 2016; Lu et al. 2016; Gnecchi-Ruscone et al. 2018). Only non-reference TEs were deemed significant by Fst, while for PBS there is a mixture of both reference and non-reference significant TEs. Such discrepancies are likely to emerge as a consequence of the systematic differences in allele frequencies between the two TEs groups, with reference TEs enriched for higher allelic frequencies (> 0.5) and non-reference TEs enriched for lower allelic frequencies (< 0.5), combined with the analytical approach of each method.

PBS analysis has been useful to detect two polymorphic TEs falling in genetic elements, ASAH1-AS1 (acid ceramidase antisense RNA 1) and PHF21A (PHD finger protein 21A). Limited literature exists on ASAH1-AS1, which codes for the antisense sequence of the proximally located ASAH1 gene. A recently published analysis of long noncoding RNAs (lncRNAs) in head and neck squamous cell carcinoma does associate its methylation status with predictive prognostic power (Zheng et al. 2023). The gene itself has never been studied in individuals of Asian ancestry, but several genome-wide association studies (GWAS) in cohorts of European and African descent in America and the Caribbean detected several SNPs associated with traits indicative of obesity (Wang et al. 2011), bone mineral density and serum urate levels in females (Yao et al. 2021), and coronary artery calcification in type 2 diabetes (Divers et al. 2017) among others, although few signals reach the p-value threshold of significance for GWAS ($10^{-8}$). Indeed, phenotypes similar to those above listed have been described by their relationship with high altitude in cohorts of Tibetan and Chinese ancestry, highlighting

significant differences with low altitude groups (Lin et al. 2019; Zuo et al. 2022; Wang et al. 2023). Moreover, the ASAH1 gene regulated by its antisense RNA encodes a ceramidase enzyme, which drives the degradation of the waxy lipid ceramide in its two components, sphingosine and a fatty acid. ASAH1 has been variably described in the context of senescent cell permanence, neuron survival and neurite structure maintenance (Kyriakou et al. 2020; Munk et al. 2021), possibly influencing the composition of lipid rafts in the cellular membrane and, therefore, its stability and the functionality of the attached proteins. Interestingly, the ASAH1-coded acid ceramidase is a lysosomal protein that also protects the cell from oxidative stress, as shown both in retinal cells overexpressing ASAH1 (Sugano et al. 2019) and, indirectly, in a cellular model of Parkinson's disease through inhibition of ceramide synthesis (Mingione et al. 2021). Furthermore, recent studies have highlighted a role of ASAH1 genetic variants in regulating the outcome of physical activity and exercise interventions (Lewis et al. 2018; Sies and Jones 2020). In this context, ASAH1 may serve a crucial role, as it is known that hypobaric hypoxia due to high altitude exposure induces inflammation by increasing the circulating levels of reactive oxygen species, which are pervasive, unstable molecules with high oxidizing power, and this is exacerbated by physical activity (Sies and Jones 2020), as well as being responsible for premature cell senescence and apoptosis (Liguori et al. 2018; Sies and Jones 2020; Pena et al. 2022; Faraonio 2022).

The PHF21A gene has been identified in the context of RNA switches induced by hypoxia that regulate oncogenic genes, as already verified for example with the vascular endothelial growth factor alpha (VEGFA) (Subbiah et al. 2020). Indeed, it appears as though PHF21A mRNA is regulated at a translational level by hypoxia-induced switches and that, together with other genes, it composes a vast translational regulon modulating hypoxia resistance and cell survival (Subbiah et al. 2020). The expression of PHF21A is also found to be significantly reduced, and the gene possibly downregulated by miRNAs, in human fibroblast-like synoviocytes, which produce pro-inflammatory mediators in osteoarthritis (OA) and are the cause of synovial pathology associated with OA (Chen et al. 2019). Again, this may be in line with epidemiological observations pointing to an increased incidence of inflammatory-mediated musculoskeletal pathologies at high altitude and in colder environments (Vega-Hinojosa et al. 2018; Farbu et al. 2022), as well as generalized hypoxia-induced osteopenia

(Brent et al. 2022), although direct dissection of the involved molecular mechanisms are still unknown in this case.

By calculating differential allele frequencies on the 11,192 identified TEs, we detected 271 TEs that discriminate between high-altitude and low/middle-altitude groups (Fisher p-value < 0.01): 126 are in genic regions and, interestingly, the most represented disease classes (according to DAVID) (Huang et al. 2009; Sherman et al. 2022) are “cardiovascular”, “hematological” and “chem-dependency”.

To strengthen our results, we also applied an association test using GEMMA (Zhou and Stephens 2012) that identified 266 TEs associated with the “altitude” context (Adjusted p-value < 0.001): of these, four are shared between the two analyses. The condition “tobacco use disorder” (represented by 51 genes mapped by as many TEs, p-value = $2.7*e^{-11}$) was deemed significant by both differential allele frequencies analysis and association test. As for tobacco use, some studies investigated the influence of smoking at high-altitude, but the results have been so far controversial: some groups identified a correlation between smoking and a lower incidence of Acute Mountain Sickness (AMS) (Wu et al. 2012; You et al. 2012; Song et al. 2014), while others point towards an absence of such correlation (Gaillard et al. 2004; Vinnikov et al. 2015, 2016), with Wu and colleagues (Wu et al. 2012) suggesting that smoking “slightly decreases the risk of AMS but impairs long-term altitude acclimatization and lung function during a prolonged stay at high altitude”. A previous work by Ramirez and colleagues (Ramirez et al. 1991), who studied the populations of the Tuquerres Plateau in southern Colombian Andes (3,000 meters above sea level), reported that in smokers there is “an increase in hemoglobin and hematocrit and a higher mean corpuscular volume and mean corpuscular hemoglobin concentration than in non-smoking high altitude subjects”.

However, this hypothesis seems inconsistent for the Tibetan populations case. In fact tobacco, whose origins can be traced back to the American continent (Duke et al. 2021), has been introduced in Europe and Asia in the last few centuries, making it an unlikely source of selective pressures in the Tibetan Plateau; on the contrary, it may have had a more relevant influence in Andean populations, which have been exposed to this substance for millennia. Therefore, we speculate that the detected

condition acts as a proxy for other physiological traits or that the involved genes, such as NELL2, fulfill other pleiotropic roles in metabolisms influenced by the high altitude condition.

Indeed, it is well known that hypoxia (i.e., reduction of oxygen intake with increasing altitude) induces physiological and morphological variations in the human brain (Zhang et al. 2022, 2023), even though the genetic underpinnings of these changes remain largely unknown. Accordingly, we hypothesize that NELL2, which is specifically expressed in neural tissues (Matsuhashi et al. 1995; Oyasu et al. 2000) and stimulates neuronal polarization as well as axon growth (Kim et al. 2020), could contribute to altitude-driven functional changes in the brain, even if at the moment there is no experimental evidence of it.

On the whole, according to all our in silico analyses, the AluYe on chromosome 12:44670594, located in the gene NELL2, arose as the most significant result (Table 4a.1). The Alu discriminates between HA and MA-LA groups (based on allele frequencies, Fisher p-value = 0.00131); more precisely, we observe that the Alu is characterized by a steady reduction of both presence and homozygosity following a decreasing altitudinal cline. In addition, it is also associated with the "altitude" condition (according to GEMMA: adjusted p-value = $1.81*e^{-4}$), suggesting a meaningful relationship between the presence of the polymorphic TE in this gene and physiological responses to high altitude. Moreover, it is one of the significant results according to PBS (score = 0.168776), pinpointing a relevant role for this variant in the process of local differentiation between high- and low- altitude populations. This Alu also acts as eQTL in skin and as sQTL in brain cortex and putamen basal ganglia (Cao et al. 2020), indicating a possible regulatory role for this element on NELL2. The AluYe is present in Denisova and in the ancient Tibetan C1 (3,150 yBP), suggesting that this element could have anciently emerged.

Finally, we cross-checked genes mapped by significant TEs for the four applied tests (differential allele frequencies analysis, Fst, PBS and association test with GEMMA) with the lists of genes under positive selection in Tibetan populations as reported by Zheng et al. (2023) and Deng et al. (2019), and candidate gene for polygenic adaptation in Tibetan and Sherpa groups as suggested by

Gnecchi-Ruscone et al. (2018). This way, we found 15 genes mapped by significant TEs which experienced instances of positive selection/polygenic adaptation in high-altitude populations of the Tibetan Plateau. For instance, the gene SUPT3H is mapped by an AluYe significant for PBS and differential allele frequencies analysis and the gene has been found under positive selection/polygenic adaptation in Tibetan groups by Zheng et al. (2023), Deng et al. (2019) and Gnecchi-Ruscone et al. (2018). More precisely, SUPT3H is a probable transcriptional activator (Martinez et al. 2001). Interestingly, the locus containing this protein-coding gene was reported by a previous study to be associated with bone and cartilage phenotypes (Boer et al. 2015) and another work found variants in the SUPT3H-RUNX2 locus to be involved in craniofacial phenotypes (Feng et al. 2021).

The present study highlights the putative contribution of the non-coding genome in high-altitude adaptation in a cohort of Tibetan and Sherpa individuals, when compared to mid-altitude and low-altitude populations of similar ancestry and geographic origins. Indeed, several transposable elements show a significant differentiation between the inhabitants of the Tibetan plateau and the other analyzed groups (as detected by Fst and PBS statistics), revealing a possible role in the control of peculiar genes involved in oxidative stress and hypoxia-induced inflammation, which themselves modulate the expression or translation of other genes. Furthermore, the prevalence of TEs with significantly different frequencies between HA and MA-LA groups and associated to the “altitude” context (according to GEMMA) similarly highlights genes involved in cardiovascular, hematological, chem-dependent and respiratory conditions, indicating that metabolic and signaling pathways taking part in these functions are disproportionately impacted by the effect of environmental stressors in HA individuals through both coding and regulatory elements. This extensive nested modulation, also pointed out by the fact that some of the detected TEs are quantitative trait loci influencing expression and/or alternative splicing, may be suggestive of a wider network of relationships between coding and non-coding elements, which intervenes in fine-tuning the physiological responses to high altitude environments.

## 4b. Susceptibility to medical conditions: insights from six isolates of North-Eastern Italy

Giorgia Modenini[*,1,#], Giacomo Mercuri[*,2], Paolo Abondio[3], Giuseppe Giovanni Nardone[4], Aurora Santin[4], Paola Tesolin[4], Beatrice Spedicati[4,5], Alessandro Pecori[5], Giulia Pianigiani[5], Maria Pina Concas[4], Giorgia Girotto[4,5], Paolo Gasparini[4,5], Massimo Mezzavilla[6], Alessio Boattini[1]. *Relationship between transposable elements and behavioral traits: insights from six genetic isolates from North-Eastern Italy.* Under review at Mobile DNA (2nd round of revision).

Affiliations:
1: BiGeA Department, University of Bologna, Bologna, Italy
2: Department of Chemistry, Life Sciences and Environmental Sustainability, University of Parma, Parma, Italy
3: IRCCS Istituto Delle Scienze Neurologiche Di Bologna, Bologna, Italy
4: Department of Medicine, Surgery and Health Sciences, University of Trieste, Trieste, Italy
5: Institute for Maternal and Child Health – IRCCS, Burlo Garofolo, Trieste, Italy
6: Department of Biology, University of Padua, Padua, Italy
*: these authors contributed equally to the study
#: correspondence to Giorgia Modenini (giorgiamodenini@gmail.com)

### Background

The Italian peninsula, due to its complex population structure, could play an important role in the understanding of the genetic diversity of current populations, being the natural crossroad for human migrations across the Mediterranean since prehistoric periods. These migration patterns left a tangible mark on present-day Italians, revealing a heterogeneous network of genomic landscapes across the peninsula, with North Italian groups being more closely related to Western/Eastern European populations and a progressively increasing genetic connection with Northern African and Middle Eastern populations as we move southwards (Sazzini et al., 2016). On top of this clinal variation across the peninsula, the natural variety of environments (Pesaresi et al., 2014) provoked a series of local adaptive events that determined, among other factors, a differential disease susceptibility of Italian subpopulations (Sazzini et al., 2016). A refined understanding of these local events would improve our knowledge of human diversity as a whole, and on a more practical level allow us to provide more ad hoc medical care and measures to particularly susceptible subpopulations.

A refined understanding of these local events would improve our knowledge of human diversity as a whole, and on a more practical level allow us to provide more ad hoc medical care and measures to particularly susceptible subpopulations. The underlying genetic variability of Italy remains under-sampled and underrepresented, with available human genome reference datasets such as the 1KGP, HGDP, and SGDP only sampling three populations for the whole peninsula: Tuscans (TSI), individuals from Bergamo and Sardinians, a notion that only worsens when considering that Italy

presents an increased number of isolated villages and subpopulations when compared to other European groups (Esko et al., 2013; Cocca et al., 2020), most of which remain uncharacterized.

These still isolated groups provide a desirable study subject to understand the Italian genetic variability: population isolates are characterized by small effective population sizes (Ne), which result in a decreased variability and stronger genetic drift effects, potentially increasing the frequency of variants that are rare or absent elsewhere and aiding at the discovery of novel rare variant signals underpinning complex traits such as medical risks and susceptibilities (Xue et al., 2017). Population isolates tend not only to be genetically homogenous but are also characterized by an elevated diversity when compared to neighboring populations and their source population (Esko et al., 2013), because of geographical and/or cultural barriers that are necessary for the formation of the isolate in the first place. For these reasons, the isolates provided useful tools for genome-wide association studies (Southam et al., 2017).

However, most of the available research on these populations is based almost exclusively on SNP data, while little work was done using other types of genetic markers. For instance, information about the variability of Transposable Elements (TE), despite them being a primary component of the human genome, has become accessible only in recent years, thanks to the availability of whole-genome sequencing data and in particular to the development of new tools for their detection and genotyping (Rishishwar et al., 2015; Gardner et al., 2017; Watkins et al., 2020). When TEs transpose in the germline, they can create novel inheritable insertions, thereby generating human-specific polymorphisms (Rishishwar et al., 2015). One of the most useful features of polymorphic TEs is that the ancestral state of these markers is known to be the absence of the insertion (Perna et al., 1992; Batzer et al., 1994). Interestingly, such markers have never been used to study the genetic underpinnings of human isolated communities; therefore, this study is the first of a kind.

In the last decades, we have come to know much more about the impact of these elements on the genome and gene networks, and it has been shown that TE insertions can generate diversity in a variety of ways. For example, transposable elements have been linked to providing polyadenylation signals inducing the termination of transcripts (Lee et al., 2008), modifying splicing patterns, and

providing new splicing sites (Belancio et al., 2008), epigenetically affecting nearby genes (Hata & Sakaki, 1997; Enriquez-Gasca et al., 2020), acting as novel promoters, enhancers, and transcription factor building sites (Kim & Hahn, 2011; Pontis et al., 2019), and often carrying their enhancers and promoters (Cordaux & Batzer, 2009). With their innate ability to act as disruptors and deregulators of gene expression, TE insertions have been associated with a variety of human diseases: for example, several cancer types (Anwar et al., 2017; Chénais 2022), hemophilia A and B (Kazazian et al., 1988; Nakamura et al., 2015), some inheritable genetic diseases such as Dent's disease or Duchenne muscular dystrophy (Payer & Burns, 2019), metabolic diseases (Jelassi et al., 2012), substance abuse, and central nervous systems diseases (Reilly et al., 2013).

In particular, much interest has been given in recent years to the impact of transposable elements on the central nervous system (Manolio et al., 2009; Reilly et al., 2013; Erwin et al., 2015). Genome-wide approaches allowed researchers to study the role of transposable elements in stress-related learning mechanisms in rats (Rau & Fanselow, 2009), which have been used as a model for PTSD in humans (Ponomarev et al., 2010). Likewise, transposable elements have also been associated with alcoholism in humans using the same genome-wide approach (Reilly et al., 2013).

In this study, we aim to reconstruct the TE variability of six isolates from Friuli-Venezia Giulia (North-Eastern Italy) thanks to the availability of whole sequencing data from 589 individuals (Esko et al., 2013). Firstly, after determining the position and the genotypes of polymorphic TEs in these populations, we use them to evaluate the isolates' structure, in the context of European and worldwide reference populations. Then, leveraging on the advantages offered by genetic and geographic isolates, we focused on exploring the potential association between non-reference polymorphic TEs, Body Mass Index (BMI) variations and behavioral traits of health and social relevance such as tobacco use and alcohol consumption. In fact, these traits could lead to an increased risk of developing addiction-related or metabolic diseases, such as tobacco use disorder, alcoholism, and obesity.

**Materials and Methods**

The dataset used in this study was generated in 2008 (Esko et al., 2013; Xue et al., 2017; Cocca et al., 2020) from the sampling of 611 individuals from six geographically and historically isolated villages in the Friuli-Venezia Giulia region of North-Eastern Italy, namely Sauris, Illegio, Resia, Erto, Clauzetto and San Martino del Carso (Figure 4b.1). Since a few of the individuals present in the dataset were duplicates (specifically, 22 individuals from Resia), and three missed village information, they were removed, leading the total number of analyzed individuals to 586. During the sampling, subjects were asked to fill out an anamnesis form to acquire more data on their general health and lifestyle habits. Phenotypic data on more than 70 phenotypes was collected, also including food preferences, olfactory perception, gustatory perception and anthropometric measures. Since the form was administered at individual discretion, missing rates vary wildly between phenotypes and individuals. We chose to focus solely on phenotypes exhibiting a missing rate of less than 10% in our dataset, thus the traits included in our analysis were sex, age, alcohol consumption, smoking, as well as height and weight, from which we calculated the corresponding BMI (weight/height$^2$). Phenotypes linked to specific diseases or health conditions, such as the occurrence of diabetes, displayed a missing rate exceeding 40%, and as such we chose to not include them in our analyses.


**Figure 4b.1.** Location of the six isolates in Friuli-Venezia Giulia, north-east of Italy (SMC = San Martino del Carso).

Genomes were scanned in search of non-reference polymorphic TEs (Alus, LINE1s, and SVAs), using the Mobile Element Locator Tool (MELT) v2.2.2 (Gardner et al., 2017). The WGS data was aligned with *bwa* to the human reference HumanG1Kv38, and the aligned reads were used as input for the Mobile Element Locator Tool (MELT). For the calling process, we used the MELT mobile element reference sequences and the collection of insertion sites discovered in Phase III of the 1000 Genomes Project (1KGP) as analysis priors. After the identification of these TEs, a self-customized Python script was applied to the resulting *vcf* files to calculate both allele and genotype frequencies of each TE for all the isolated villages. Allele frequencies were then analyzed for significant differences between villages with Fisher's exact test, using a significant threshold of nominal p-value < 0.01 ("differentiated" TEs).

MELT provides gene names in RefSeq format: therefore, RefSeq accession numbers were converted to their respective Official Gene Symbol using the Database for Annotation, Visualization and Integrated Discovery (DAVID) (Sherman et al., 2022; https://david.ncifcrf.gov/), taking into consideration the specific gene region TEs were inserted in (Intron, Exon, Promoter, Terminator, 5' UTR and 3' UTR).

To compare TE diversity of the isolates with other human populations, we built a new dataset consisting of polymorphic TEs (identified with MELT; Gardner et al., 2017) that were present both in the six isolates and in the populations of the 1000 Genomes Project. This newly merged dataset contained a total of 2,814 genetic loci for 3,090 individuals from 32 populations. The populations were divided into 6 groups based on geographic macro areas, consistent with the super populations of 1KGP (The 1000 Genomes Project Consortium, 2015), i.e. Africa (AFR), America (AMR), East Asia (EAS), Europe (EUR), South Asia (SAS), plus the isolates from Friuli-Venezia Giulia (FVG).

TEs were coded as single nucleotide variants, substituting the insertion with a nucleotide base that was non-complementary to that of the other allele in the genotype file. Information about the true nature of each insertion was kept in the original *vcf* file. Variants were then filtered with PLINK v1.9 (Purcell et al., 2007) as follows: 1) Removal of insertions located on sexual chromosomes or mitochondrial genome insertions, to retain only autosomal variability and removal of duplicates, using the --*exclude* option. 2) Exclusion of individuals and variants with > 1% missing data with the commands --geno 0.01 (for variants) and --mind 0.01 (for individuals). 3) Removal of variants that did not respect the Hardy-Weinberg Equilibrium (HWE) with the option --hwe, setting a significant threshold of 0.01 using a Bonferroni Correction for multiple testing (threshold= 0.01/number of variants). 4) Removal of variants with a minor allele frequency < 0.01 (--maf 0.01). 5) Removal of closely related individuals with an Identity by Descent (IBD) estimate higher than 0.25, using the --genome option to calculate the pairwise IBD estimates between every couple of individuals and --remove to exclude one of the two related individuals. Therefore, the final filtered dataset was made of 1,703 variants shared among 3,087 individuals.

The generated dataset was then used to perform a series of analyses on TE insertions from the six isolates when compared to 1KGP groups. Both a Principal Component Analysis (PCA) and Admixture analysis were applied: PCA was performed after the conversion from the PLINK format (*bed, bim, fam*) with the *convertf* and *smartpca* tools of the EIGENSTRAT v6.0.1 package (Price et al., 2006). Admixture was implemented with the ADMIXTURE tool (Alexander & Lange, 2011), testing between 2 and 23 potential ancestry components (K) and performing 50 iterations of each run to minimize the estimation error and maximize the *log-likelihood* of each ancestry estimate.

We then compared FVG isolates with other European populations, subsampling the original 1KGP dataset as follows: Utah residents with North-Western European ancestry (CEU), Finnish in Finland (FIN), British in England and Scotland (GBR), Iberian populations in Spain (IBS), Tuscans in Italy (TSI). PCA and Admixture analyses were implemented using the above approach, the only difference being that we tested a number K of putative ancestry components between 2 and 12.

As introduced in the "background" section, individuals were asked to fill out an anamnesis form, including information on their health status and lifestyle habits. Three phenotypes were selected to perform association studies between polymorphic TEs and the considered traits: tobacco use, alcohol consumption, and body mass index (the latter was calculated as weight/height$^2$). The association studies were performed with the software GEMMA (Zhou & Stephens, 2012; Zhou & Stephens, 2014), using a genome-wide (GWAS) like approach. In particular, we applied for all the considered phenotypes a univariate linear mixed model (uvLMM) for association tests between a marker, a chosen phenotype, and any chosen covariates, while also correcting for the potential presence of population stratification (indeed a typical feature of isolates), and estimating genetic correlation among phenotypes (Zhou & Stephens, 2014). GEMMA was applied to the full FVG dataset (12,709 TEs and 586 individuals) and three separate uvLMM association analyses were performed, using sex and age as covariates: 1) BMI; 2) a binary alcohol drinker/non-drinker variable (set as "1" for drinker individuals and "0" for non-drinkers); 3) a binary smoker/non-smoker variable (using "1" for smokers and "0" for non-smokers). A fourth association test on the smoker individuals was performed to

evaluate the possible association between polymorphic TEs and the number of cigarettes smoked per day/number of years smoking. Variables were tested using Wald's test with a significant threshold of p-value = 0.001; Manhattan plots with significant results have been obtained with the *CMplot* R package (https://github.com/YinLiLin/CMplot; Yin et al., 2021). To better assess the importance of the associated TEs, we also performed a haplotype reconstruction/association test procedure on the significant variants from the alcohol and smoking tests detected with GEMMA using Beagle (S.L. Browning and B.L. Browning, 2007; B.L. Browning, Y. Zhou and S.R. Browning, 2018) (this software performs association tests only on binary traits). First, we selected regions of interest (10kb upstream and downstream the significant TE, for a total of 20kb) with VCFtools (Danececk et al., 2011) and phased those regions with the software Beagle v5.1. The obtained *vcf* files were converted into the Beagle format with vcf2beagle (https://faculty.washington.edu/browning/beagle_utilities/) and the case status "smoking" or "alcohol" was included in the second row of the *bgl* files. Lastly, the association test on the reconstructed haplotypes was performed with Beagle v3.3.2 and the significant results were checked with the cluster2haps utility.

In order to investigate a possible function for the identified TEs, we then cross-checked the significant results with the lists of polymorphic TEs acting as expression/alternative splicing quantitative trait loci produced by Cao and colleagues (2020). For each gene analyzed we collected measures of genetic constraints such as pLI (probability of loss of function intolerance) (*Lek et al., 2016*), RVIS (Genic Intolerance) (Petrovski et al., 2013) and SSC score (Singletons Score) (Mezzavilla et al., 2020) for prioritization. We considered as constrained those genes with pLI > 0.9 or RVIS < -0.43 or SSC score < -2.

**Results**

After the analysis of polymorphic non-reference TEs with MELT v.2.2.2 (Gardner et al., 2017), a total of 9,525 Alus, 2,283 LINE1s, and 901 SVAs were retrieved.

Then, allele frequencies were scanned for significant differences among the isolates: this way, a total of 3,987 TEs (31.37%) were identified as "differentiated", of which 3,195 Alus (33.54%), 636 LINE1s (27.86%), and 156 SVAs (17.31%). When considering all comparison European populations,

the corresponding rates of “differentiated” TEs are 53.45% (Alus), 58.63% (LINE1s) and 51.24% (SVAs).

Of these insertions, we also considered their location (Table 4b.1).

| | **Alu** | **LINE-1** | **SVA** |
|---|---|---|---|
| INTRONIC | 1,281 (40,1%) | 242 (38%) | 65 (41,7%) |
| PROMOTER | 138 (4,3%) | 23 (3,6%) | 10 (6,4%) |
| TERMINATOR | 106 (3,3%) | 28 (4,4%) | 11 (7%) |
| EXON | 38 (1,2%) | 6 (1%) | 6 (3,8%) |
| 3’-UTR | 39 (1,2%) | 6 (1%) | 3 (1,9%) |
| 5’-UTR | 18 (0,6%) | 2 (0,3%) | 0 |
| INTERGENIC | 1,575 (49,3%) | 329 (51,7%) | 61 (39,1%) |
| TOTAL | 3,195 | 636 | 156 |

**Table 4b.1**. Significantly different polymorphic TEs between the six villages, divided by insertion location relative to gene region (with percentages) and TE superclass.

As expected, most polymorphic TEs insertions are located in intronic and intergenic regions and only a negligible fraction are located in exonic regions. However, it is interesting to note that SVAs, which can be up to 3kb long (Wang et al., 2005), are overall less frequent in intergenic sequences while they appear more often located in “functional” regions (regulators or exons) when compared to Alus and LINE1s. This finding corroborates the notion that SVA insertions have the innate potential to regulate gene expression through their location insertion and their sequence characteristics (Gianfrancesco et al., 2019; Barnada et al., 2022).

Both TE-based PCA and Admixture show that, while closely related to other European populations, our isolates tend to cluster amongst themselves and are dominated by drift-induced ancestry components (Supplementary Figure S1). In particular, the first PC discriminates between African and non-African populations, while the second PC highlights a West-to-East geographical pattern

including individuals from Friuli-Venezia Giulia, Europeans, Americans, South Asians, and East Asians.

The PCA between European and FVG populations divides the two groups along the first PC, while the second component highlights the variability between the isolates, separating Resia and some individuals from Clauzetto and Sauris from the rest (Figure 4b.2 A). As expected considering their geographical proximity and historical relatedness, Tuscans (TSI) and Central Europeans (CEU) are the closest groups to the FVG isolates. This PCA is similar to the one resulting in Esko et al. (2013) based on SNPs. Looking at the second and third PCs, it is interesting to note that PC2 separates Resia from Clauzetto, while the third component highlights the differentiation between Sauris and Illegio. Instead, Erto, San Martino and most individuals from Clauzetto cluster together with the other European populations (Figure 4b.2 B), hence suggesting a lower degree of isolation for these groups. Finally, looking at the Admixture graph (Figure 4b.2 C), the "tidiest" model is for K = 7, as K = 9 presents excessive noise, especially in the African outgroup. However, at the best fitting K = 9 (CV error = 0.31088), Illegio, Resia, San Martino and Sauris are all dominated by their own ancestry components, which are present only marginally in Clauzetto, Erto, and the other European populations.

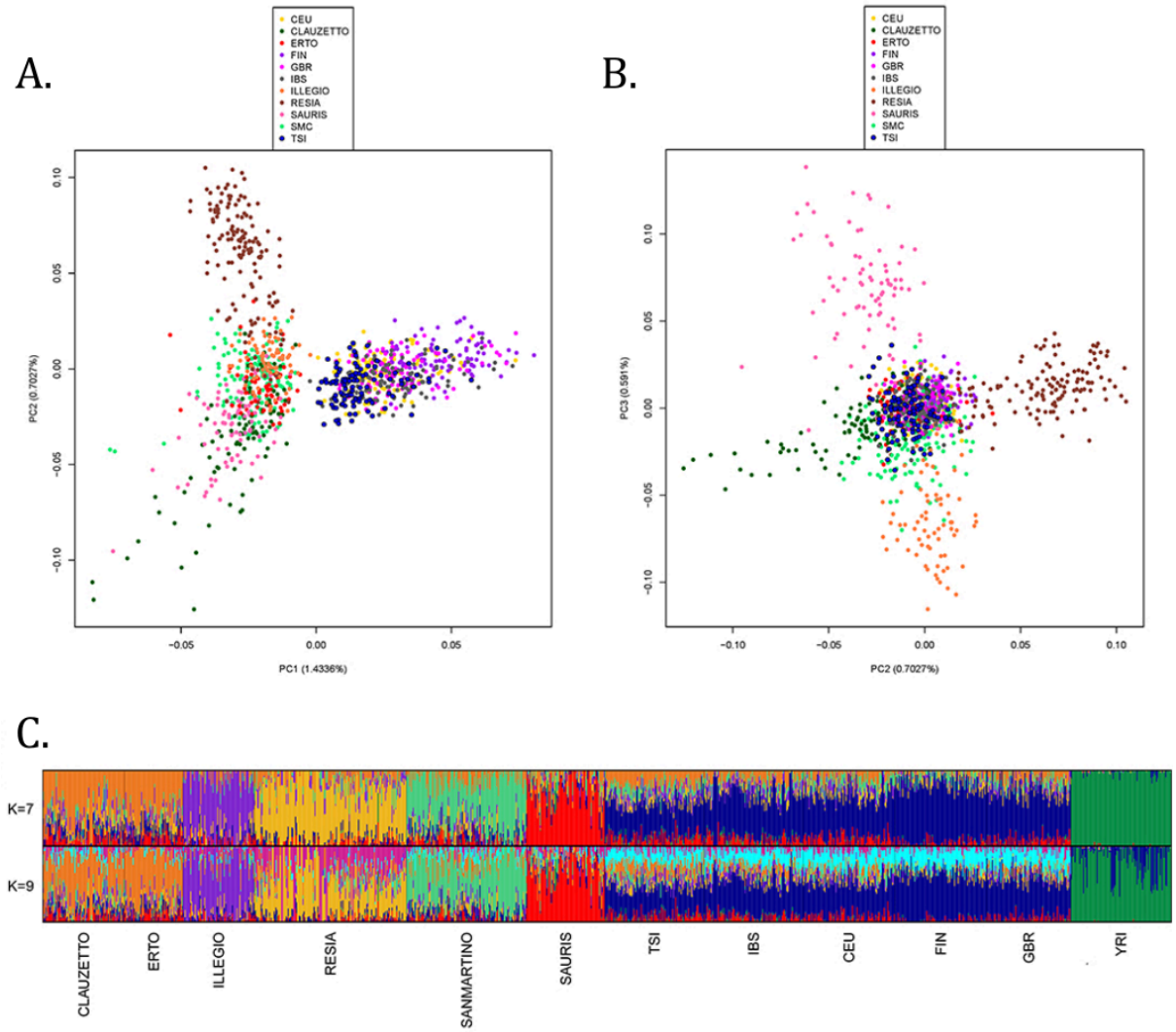



**Figure 4b.2**. **A and B)** PCA plots of European populations from 1KGP and FVG isolates, first against second component (A) and second against third component (B). **C)** Admixture barplots for K = 7 and K = 9.

Several polymorphic TEs were identified by the association tests with GEMMA (Zhou & Stephens 2012, 2014) as possibly associated with the conditions detailed in Materials and Methods, and some of them also act as eQTLs/sQTLs:

1) Variations in Body Mass Index: three insertions were deemed significant, namely twoAlus and one SVA (Figure 4b.3, second Manhattan plot from the outside to the inside). Notably, the SVA on chr17:49150166 is located in the gene SPAG9 (Sperm Associated Antigen 9) (Table 4b.2). The other two significant results are two intergenic Alus located on chr1:241980544 and chr14:65796449.

Interestingly, both act as eQTLs and sQTLs in several tissues, such as pituitary, blood, heart, testis, and ovary.

2) Alcohol consumption: six Alus were found to be significant (Figure 4b.3, inner Manhattan plot), with only one in a genic region, the Alu on chr12:14020945 in the gene GRIN2B (Glutamate Ionotropic Receptor NMDA Type Subunit 2B). This TE was previously identified as "differentiated" among the isolates and is generally widespread in our six villages (Table 4b.2). The other five intergenic elements are all Alus and are located on chr6:1257163, chr6:161283170, chr12:58367298, chr13:112866653 (eQTL in testis, skin and colon sigmoid) and chr18:26214257 (sQTL in testis, adipose visceral omentum, thyroid and breast mammary tissue).

3) Tobacco use (smoking): seven TEs were deemed significant (Figure 4b.3, outer Manhattan plot), three of which are located in genes, namely the Alu on chr3:42856928, which acts as eQTL and sQTL in different tissues (including brain and lung) and is located in the gene ACKR2 (Atypical Chemokine Receptor 2); the Alu on chr11:102654750 in WTAPP1 (Wilms Tumor 1 Associated Protein Pseudogene 1), acting as eQTL in adrenal gland; and the Alu on chr12:129970510 in TMEM132D (Transmembrane Protein 132D). These last two Alus are mostly widespread in the six villages (Table 4b.2) and were both identified as "differentiated" between the isolates (Table 4b.2). The other four intergenic insertions are located on chr2:174296971 (Alu), chr2:188989149 (Alu, acting as sQTL in brain hippocampus and cerebellum), chr14:65796449 (Alu, acting as eQTL/sQTL in several tissues), and a LINE-1 on chr22:26454699.

4) A further association test with GEMMA was performed on the "smoking" condition by taking into account the amount of cigarettes smoked per day and the number of years smoking. In total, six Alus were significantly associated with this phenotype, four of which were inserted in gene introns. The Alu on chr2:155232845, located in the gene GALNT13 (Polypeptide N-Acetylgalactosaminyltransferase 13); the Alu on chr12:123580101, located in the gene PITPNM2 (Phosphatidylinositol Transfer Protein Membrane Associated 2); the Alu on chr18:29519986, located in the gene TRAPPC8 (Trafficking Protein Particle Complex Subunit 8); and the Alu on chr19:19350607, located in the gene NCAN (Neurocan).

Among these genes, TMEM132D, and GRIN2B show evidence of genetic constraints using either pLI Score, RVIS, or SSC Score. Absolute genotype frequencies of these insertions are reported in Table 4b.2. The Alus in the genes SPAG9, PITPNM2, TRAPPC8 and NCAN, despite being significant, appear to be rare, therefore we did not reported the genotype frequencies in Table 4b.2 (the percentage of individuals who carry the insertion is 1.9% for SPAG9, 0.18% for PITPNM2, 0.26% for TRAPPC8 and 0.19% for NCAN).

We finally reconstructed haplotypes around the above mentioned TEs and performed haplotype-based association tests as described in Methods. We obtained two significant results, both for the alcohol phenotype, namely the two intergenic Alus on chr6:1257163 and chr6:161283170. In both cases, the associated haplotype is characterized by the presence of the mobile element. The first haplotype included 19 SNPs and one Alu (p-value = 0.00164); the second is an haplotype with 7 SNPs and the TE (p-value = 0.000335).


**Figure 4b.3**. Circular manhattan plots of the first three association tests (BMI, alcohol, smoking). The plot is read from the outside to the inside, in this order: smoking, BMI and alcohol. Greater dots over the red-dotted line (Wald's p-value < 0.001) are the significant TEs for the association tests. Only autosomal chromosomes were represented as shown in the black circle outside the plots.

| | Resia | | | Erto | | | Illegio | | | Sauris | | | San Martino | | | Clauzetto | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **0\0** | **0\1** | **1\1** | **0\0** | **0\1** | **1\1** | **0\0** | **0\1** | **1\1** | **0\0** | **0\1** | **1\1** | **0\0** | **0\1** | **1\1** | **0\0** | **0\1** | **1\1** |
| **GRIN2B** | 53 | 60 | 30 | 27 | 30 | 6 | 23 | 37 | 17 | 35 | 39 | 11 | 52 | 70 | 8 | 32 | 51 | 5 |
| **ACKR2** | 56 | 66 | 20 | 35 | 23 | 5 | 33 | 32 | 12 | 31 | 41 | 12 | 58 | 58 | 14 | 43 | 39 | 6 |
| **WTAPP1** | 104 | 37 | 1 | 59 | 4 | 0 | 65 | 11 | 1 | 76 | 9 | 0 | 119 | 11 | 0 | 77 | 11 | 0 |
| **TMEM132D** | 75 | 57 | 11 | 22 | 30 | 11 | 42 | 31 | 4 | 45 | 38 | 2 | 46 | 67 | 17 | 38 | 38 | 12 |
| **GALNT13** | 118 | 24 | 0 | 60 | 3 | 0 | 77 | 0 | 0 | 77 | 7 | 0 | 118 | 12 | 0 | 86 | 2 | 0 |

**Table 4b.2**. Absolute genotype frequencies of the five Alus located in the genes GRIN2B, WTAPP1, ACKR2, TMEM132D, and GALNT13.

**Discussion**

The study of isolated communities is at the basis of population genetics research (Charlesworth, 2009; Hatzikotoulas et al., 2014). In fact, isolates yield genomes that show high homogeneity and are subject to similar environmental and cultural pressures, such as lifestyle habits, diet, sanitary conditions, and disease vectors. These populations are also an ideal subject to study the phenotypic effects of variants that were otherwise only marginally present in larger populations (Hatzikotoulas et al., 2014). In this picture, Italian isolates are particularly important, mainly because of the peninsula's central role in human migrations since prehistoric times and of the high number of genetically distinct isolated communities that have been established throughout history (Destro Bisol et al., 2008). Polymorphic TEs, which have previously been used as both variability and susceptibility markers only in "general" populations (Reilly et al., 2013; Rishishwar et al., 2015; Gardner et al., 2017), are here applied for the first time to human isolates. Using the Mobile Element Locator Tool (Gardner et al., 2017) more than 12,000 polymorphic TEs were identified in the six villages of Friuli-Venezia Giulia. These TEs were used as genetic markers to obtain a first overview of their potential impact on diversity and disease susceptibility in isolated populations, in particular: 1) to study communities' differentiation; 2) to explore the genetic variability of the isolates; 3) and to analyze their possible role

as genetic variants underlying susceptibility to different behavioral traits or medical conditions (tobacco use, alcohol consumption, and BMI variations).

Firstly, after calculating allele and genotype frequencies of the identified TEs, we found that of 12,709 TEs, 3,987 (31.37%) have significantly different allele frequencies between the six isolates (Fisher's exact test, p-value < 0.01), while the corresponding rate in European comparison populations is 53.78%. Considering the much lower geographic dimensions of FVG compared to Europe, these values suggest the presence of genomic structure among the isolates.

Then, TEs were used as markers for exploratory population analyses, such as PCA and Admixture, to look at the general diversity and ancestry of FVG isolates in the context of European genetic variability, as represented by the polymorphic TE content of European populations from 1KGP (Gardner et al., 2017). Our results show that FVG isolates tend to cluster amongst themselves (PC1 in Figure 4b.2 A, Figure 4b.2 C), compared to European populations; however some differentiation between the isolates is evident, particularly for Resia and some individuals from Clauzetto (PC 2 in Figure 4b.2 B, Figure 4b.2 C), as well as Sauris and Illegio (PC3 in Figure 4b.2 B, Figure 4b.2 C). Instead, Erto, San Martino and most individuals from Clauzetto overlap with the other European populations. These results agree with previous SNP-based studies, according to which, Clauzetto is the least isolated village among the six FVG isolates (Esko et al., 2013); at the same time Clauzetto, Erto and San Martino overlap to the considered European populations and have the lowest inbreeding coefficients among the villages (Cocca et al., 2020). The observed patterns of genetic variability and ancestry components could be explained by population structuring and genetic drift, a suggestion made also by previous works on the same dataset (Esko et al., 2013; Xue et al., 2017; Cocca et al., 2020). The observation of a strong correlation between SNP-driven results and TE-driven results in terms of population structure further highlights that the variability of polymorphic TE is mainly the result of demographic events.

To sum up, population structure analyses confirmed that on the whole, our populations show the typical marks of isolates also from the TEs point of view. As previously mentioned, due to their internal homogeneity both at genetic and social levels, isolates are ideal populations for performing

genome-wide association studies. On the other hand, their relatively low census size implies a moderate number of available samples. More importantly, the presence of population structure is well known to induce false positives in association studies. However, the impact of the observed structure is probably moderate or at least not higher than in association studies at a country level, as suggested by exploratory population analyses (PCA, Admixture) and proportions of "differentiated" TEs. In addition, the usage of GEMMA should overcome distortions due to population structure, as confirmed by the fact that only a minority (3/22) of the associated variants show significant differentiation among isolates.

In this context, polymorphic TE insertions are particularly worthy of investigation, being potential risk variants for several medically relevant phenotypes, because of their innate ability to act as deregulators of gene networks (Enriquez-Gasca et al., 2020). Notably, the link between transposable elements and the health of the Central Nervous System is not new (Manolio et al., 2009; Erwin et al., 2015), with the effects of TEs being associated with stress, neurodegeneration, ageing, and drug abuse (Reilly et al., 2013). As such, TE markers can allow us to perform a first exploration of the medical susceptibility of individuals from the studied villages, by testing for association between TEs and phenotypes linked to behavioral and anthropometric traits.

Accordingly, we performed four association tests with GEMMA (Zhou & Stephens, 2012; Zhou & Stephens, 2014) in order to obtain a first overview of the polymorphic TEs that could underpin the variability of selected phenotypes, i.e. tobacco use, alcohol consumption, height and weight, from which we calculated body mass index (weight/height$^2$). For tobacco use, two separate analyses were run, the first comparing smokers with non-smokers, the second only on smoker individuals, testing for the association between the number of cigarettes smoked per day and the number of years smoking. The GEMMA algorithm was selected due to its ability to take into account population structure, which is a typical feature of isolated populations. In addition, sex and age were introduced in the models as covariates. Manhattan plots are shown in Figure 4b.3. Several TEs were deemed significant, some of which are located in known genes: an SVA (chr17:49150166) in the gene SPAG9 (BMI variations); the Alu on chr3:42856928 in the gene ACKR2, the Alu on chr11:102654750 in

WTAPP1, the Alu on chr12:129970510 in TMEM132D (tobacco use/smoking) and the Alu on chr12:14020945 in the gene GRIN2B (alcohol consumption). Interestingly, two of the results for the alcohol consumption phenotype were deemed significant also for the haplotype-based association test performed with Beagle (S.L. Browning and B.L. Browning, 2007; B.L. Browning, Y. Zhou and S.R. Browning, 2018). In both cases, haplotypes including polymorphic TEs appear as significantly associated with the status "alcohol drinker". As for the number of cigarettes/number of years smoking, it resulted of interest the Alu insertion on chr2:155232845 in the gene GALNT13. Additionally, the insertions in WTAPP1, TMEM132D, and GRIN2B were also identified as "differentiated" when looking at genotype and allele frequencies between the isolates. Two of these genes (TMEM132D and GRIN2B) also show evidence of genetic constraints and thus should be prioritized in further investigations, as genes showing evidence of purifying selection in healthy individuals may be judged more likely to cause certain kinds of disease. For instance, the gene GRIN2B encodes a member of the ionotropic glutamate receptor superfamily and plays a major role in brain development and synaptic plasticity, with mutations in this gene often associated with neurodevelopmental disorders (Platzer & Lemke, 2018). Moreover, variants of this gene have been associated with alcohol and tobacco consumption (Saunders et al., 2022), general risk-taking behaviors (Karlsson Linnér et al., 2019), opioid dependence (Sherva et al., 2021), and several neurological disorders such as schizophrenia (Goes et al., 2015) and Alzheimer's disease (Kulminski et al., 2022). Regarding tobacco use, the insertion in ACKR2 (also known as D6) emerged as one of the most promising results, in fact the Alu acts as eQTL/sQTL in brain and lung tissues. The gene (Nibbs et al., 1997) controls chemokine levels and localization and is known to be involved in inflammatory responses (Cancellieri et al., 2013). Moreover, a work by Bazzan and colleagues (2013) on chronic obstructive pulmonary disease (COPD) "demonstrates an increased expression of the atypical chemokine receptor D6 in peripheral lung from smokers with COPD but not in smoking subjects who did not develop the disease and nonsmoker control subjects". Furthermore, TMEM132D, encoding for a transmembrane protein, has already been associated with many neurological disorders such as anxiety and panic disorders (Otowa et al., 2014) and general behavioral disinhibition, including alcohol consumption and dependence, illicit drug use, and nicotine use (McGue et al., 2013). Lastly, an Alu inserted in the gene GALNT13

was found to be highly associated with the prolonged use of tobacco (number of cigarettes smoked daily/number of years of smoking). Mutations in GALNT13, which normally encodes for Polypeptide N-Acetylgalactosaminyltransferase 13, a transferase linked with the metabolism of proteins and the glycosylation of mucins (Festari et al., 2017) also expressed in the Purkinje cells of the developing brain (Zhang et al., 2003), have been associated with nicotine use (Saunders et al., 2022), severe comorbidity between nicotine dependence and major depression (Zhou et al., 2018). Furthermore, GALNT13 expression has been found to be increased in lung cancers (Nogimori et al., 2016), suggesting yet another deep link between the identified gene insertions and genes involved in brain development, possible neurological/physical diseases and addiction-seeking behaviors.

Polymorphic transposable elements emerge as a compelling avenue for elucidating human genetic diversity. The innovative utilization of polymorphic TEs as markers for genetic variability within isolated communities represents an unprecedented methodological advancement. This study demonstrates the utility of polymorphic TEs in effectively encapsulating genetic variability and historical contexts among isolates, substantiated by congruent outcomes with prior investigations relying on single nucleotide variants (Esko et al., 2013; Xue et al., 2017; Cocca et al., 2020). While progress has been made, the comprehensive impact of transposable elements on the human genome remains incompletely understood, as does the cascade of effects on diverse phenotypes. This investigation identifies numerous TE insertions correlated with specific phenotypes, such as substance use and metabolic disorders. It is imperative to underscore the exploratory nature of our analyses, necessitating further empirical validation to establish definitive causal links between these insertions and medical susceptibility. Nevertheless, the identified insertions stand as pivotal points of interest, providing a foundational platform for subsequent research. In the context of isolated communities, these populations serve as invaluable "laboratories", affording unique insights into the influence of transposable elements on physical, psychological, and behavioral traits. Consequently, prospective studies should prioritize the validation of identified variants and engage in selection analyses to discern potential instances of natural selection within these isolated populations. This forward-looking

research agenda holds significant promise for advancing our understanding of the intricate interplay between transposable elements and human phenotypic traits.

# 5. Brain diseases as evolutionary trade-offs

## 5a. Polymorphic and evolutionarily recent transposable elements contribute to schizophrenia

Giorgia Modenini[1,*], Paolo Abondio[1,2,*], Guia Guffanti[3], Alessio Boattini[1,#], Fabio Macciardi[4,#]. *Evolutionarily recent retrotransposons contribute to schizophrenia*. Transl Psychiatry. 2023 May 27;13(1):181. doi: 10.1038/s41398-023-02472-9

Affiliations:
1: BiGeA Department, University of Bologna, Bologna, Italy
2: Department of Cultural Heritage, University of Bologna, Ravenna, Italy
3: Department of Psychiatry, McLean Hospital-Harvard Medical School, Belmont, MA, USA
4: Department of Medical Education (Neuroscience), CUSM, Colton, CA, USA
*: These authors contributed equally to the study
#: Correspondence to alessio.boattini2@unibo.it, fmacciar@hs.uci.edu

### Background

Recently, TEs have been found to be active in several different regions of the human brain: this suggests that TEs play a role in normal brain development and, possibly, in psychiatric disorders and neurological diseases (Ferrari et al., 2021; Ahn et al., 2023). TEs activity in both germline and somatic cells is highly regulated and usually repressed through different mechanisms, such as post-transcriptional (Nishikura, 2006; Goodier et al., 2012) and epigenetic processes (Slotkin and Martienssen, 2007). However, despite cells' machinery to regulate transposition, some TEs are able to escape repression and generate new insertions in germline cells during embryonic development. Indeed, numerous deleterious effects caused by TEs activity have been suggested for schizophrenia (Guffanti et al., 2014; Guffanti et al., 2018; Modenini et al., 2023). Moreover, a recent study suggests a potential key role for TEs in rewiring the local functional architecture of Human Accelerated Regions (HARs) in Schizophrenia and Bipolar Disorder (Erady et al., 2022). Indeed, HARs have been implicated in neurodevelopmental and neuropsychiatric disorders (Cheung et al., 2022; Doan et al., 2016; Hubisz & Pollard, 2014), and most HARs are known to act as developmental enhancers that are involved in controlling and regulating human cognition (Boyd et al., 2015; Franchini & Pollard, 2017; Girskise et al., 2021; Levchenko et al., 2018; Ziffra et al., 2021).

In this study, we aimed at identifying polymorphic non-reference TEs (TEs) that can potentially contribute to schizophrenia. Therefore, we looked at the non-reference TE content of Dorso-Lateral Pre-Frontal Cortex (DLPFC) genomes of schizophrenic individuals (SCZ) and psychiatrically healthy controls (CTRL) to investigate the brain tissue-specific presence of TEs, verify their somatic rather than germ-line origin and investigate their possible contribution to this cognitive disorder. To accomplish this task, we: 1) compared SCZ and CTRL genomes; 2) checked for the presence of TEs and population-specific/geographic distribution of the identified variants in the 1000 Genomes data; 3) performed haplotype-based association tests; 4) explored *in silico* the possible functional roles of TEs as *cis*-regulatory elements of protein-coding genes and as putative modifiers of known HARs.

**Subjects and Methods**

To disentangle the relationship between polymorphic non-reference TEs and schizophrenia, as well as to investigate the possible influence of these elements on cognitive evolution, twenty (20) high-coverage genomes from the DLPFC of ten SCZ and ten CTRL were analyzed jointly with 125 samples from five different populations of the 1KGP: Europeans (CEU), Han Chinese (CHB), Indians (ITU), Yoruba (YRI) and Luhya (LWK). This study was supported by the University of California, Irvine (USA) and by the collaboration with Prof. Fabio Macciardi, MD/PhD.

DNA from the DLPFC of ten SCZ and ten CTRL individuals has been obtained from the UCI Brain Bank. Donors or their first-degree relatives signed an informed consent to the UCI Brain Bank to have their tissues donated for scientific research, under an UCI-IRB approved protocol. Our sample includes 14 men and 6 women, whose ages at death ranged from 31 to 68 (average = 46.1 ± 11.4 (of which CTRL: 48 ± 13, SCZ: 44.3 ± 10, p-value = ns)). Cases and controls were matched for sex and age. Brain tissues have been collected within a mean postmortem interval (PMI) of 19 ± 4 hours. All samples presented a pH from 6.0 to 7.1 (average 6.4 ± 0.3). Specimens were checked for the presence of other potential disease states as described in Guffanti et al. (2018). Following dissection, samples were flash frozen. We extracted DNA from 80-100 ng DLPFC frozen samples using the Qiagen DNA kit. DNA concentration was assessed using a NanoDrop spectrophotometer.

The twenty high-coverage (~30-fold) genomes have been sequenced by Illumina and *fasta* files were retrieved from the Illumina sequences. AdapterRemoval (Lindgren, 2012) was used to remove adapters from the *fastq* files. Alignment to the human reference genome hs37d5 (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/) was performed with BWA-mem (Li and Durbin, 2009). After sorting and merging with Samtools (Li et al., 2009), obtained *bam* files were indexed and processed with MarkDuplicates (http://broadinstitute.github.io/picard/). Finally, the GATK best practices were applied to generate VCF files that include SNPs and TEs (https://gatk.broadinstitute.org/hc/en-us/articles/360035894711-About-the-GATK-Best-Practices).
The Mobile Element Locator Tool (MELT) v.2.1.5 (Gardner et al., 2017) with the function *Split* was used to detect the polymorphic non-reference TE content of the 20 DLPFC samples jointly with 125 1KGP individuals.

To assess the genetic relationships among the individuals included in our dataset (20 DLPFC + 125 1KGP samples), as well as their ancestry, a principal component analysis (PCA) and ADMIXTURE analysis were implemented both on the whole variant dataset (SNPs + TEs) and on the TE-based only. Quality control (QC) was performed with the PLINK software (Purcell et al., 2007) on both datasets, including the removal of genetic elements belonging to the sex chromosomes, a check for the proportion of missing data (using the commands --geno and --mind with a threshold of 0.01), the respect of Hardy-Weinberg equilibrium after Bonferroni correction for multiple testing (--hwe 0.01/α, where α is equal to the number of variants remaining in the dataset at this stage of the QC procedure), the removal of rare variants (--maf 0.01) and an assessment of linkage disequilibrium along the genome, using a sliding window of 50 bp, a moving step of 5 bp and a threshold value of 0.1 (--indep-pairwise 50 5 0.1).

Fisher tests of independence with one and two degrees of freedom were performed to identify TEs with different allele and/or genotype frequencies between SCZ and CTRL. No correction for multiple testing was done due to the low number of individuals analyzed. Only TEs that yielded a nominal

p-value for allele counts (< 0.05) were considered as putatively related with an increased risk of developing schizophrenia ("differentiated" TEs).

Information about the location of the "differentiated" TEs were retrieved from MELT output; gene annotation in "RefSeq" format was converted into the official gene symbol ID using the software DAVID (Database for Annotation, Visualization and Integrated Discovery, https://david.ncifcrf.gov/). Then, genes were scanned for potential known relationships with schizophrenia and/or neural development to infer a possible link between the presence/absence of the TEs and the pathological condition.

To strengthen our findings, a haplotype reconstruction on the whole variant dataset (SNPs + TEs) was used to contextualize the genotyped TEs into their local genetic environment and to evaluate the frequency of the corresponding haplotypes within the DLPFC cohort. After checking for missing data and Hardy-Weinberg equilibrium expectations, ambiguous SNPs (carrying an A/T or C/G combination of alleles, for which the maternal and paternal chromosome, as well as the strand cannot be unequivocally defined) were removed. Information about the ancestral or derived nature of each SNP was deduced by using a reconstructed reference human genome sequence. Briefly, ancestral/derived states of each allele were previously assigned by aligning the human reference genome hs37d5 with the five available Ensembl Compara primates EPO reference genome sequences (Herrero et al., 2016): bonobo (*Pan paniscus*), chimpanzee (*Pan troglodytes*), gorilla (*Gorilla gorilla*), macaque (*Macaca fascicularis*) and orangutan (*Pongo abelii*). Only alleles conserved in all the compared genomes are considered as ancestral. In this framework, for TEs, the derived allele corresponds by definition to the presence of the element. Haplotype estimation was finally performed with the SHAPEIT software version 1.9 (Delaneau, Zagury and Marchini, 2013) on a dataset of 8,331,932 variants (SNPs and TEs). Haplotype estimation also included haplotype phasing.

Then, an association test with Beagle v.3.3.2 (S.L. Browning and B.L. Browning, 2007) was performed on the "differentiated" TEs. First, we selected regions of interest using VCFtools (Danecek et al., 2011) and phased those regions with Beagle v5.1 (S.L. Browning and B.L.

Browning, 2007; B.L. Browning, Y. Zhou and S.R. Browning, 2018) as described in the manual. Then, we converted the phased VCF files into the Beagle format with the Beagle utility *vcf2beagle* (https://faculty.washington.edu/browning/beagle_utilities/) and included the case status ("schizophrenic") in the first row of the *.bgl* file. Lastly, we performed the association test with Beagle v3.3.2 as described in the manual and then checked for significant results with the *cluster2haps* option.

**Results**

The Mobile Element Locator Tool (MELT) v.2.1.5 (Gardner et al., 2017) with the function *Split* was used to detect the non-reference TE content of the 20 DLPFC samples jointly with 125 1KGP individuals. Only sites: 1) on autosomal chromosomes; 2) that passed every quality control (i.e., "PASS") and 3) that were deemed as "genic" (which means that they fall into genes or in regulatory regions such as promoters or terminators) were used for further analyses, which included a total of 7,952 TEs: 6,542 Alus, 1,065 LINE-1s and 345 SVAs.

To check for a potential underlying genetic structure of cases and controls and to contextualize the DLPFC samples in the worldwide genomic landscape, as represented by the five 1KGP populations, a total of 3,211 TEs (i.e., TEs that passed QC) were used for PCA and Admixture analyses (Figure 5a.1, A and B). PCA was performed by applying a series of file format conversions and computations as required by the *convertf* and *smartpca* tools from the EIGENSOFT package v6.0.1 (Price et al., 2006). Similarly, the ADMIXTURE software (Alexander, Novembre and Lange, 2009) was employed to estimate the shared genetic ancestry across populations. We tested between 2 and 7 putative ancestral components (K), performing 50 iterations of each run to minimize the estimation error and maximize the log-likelihood of each ancestry estimate, and found that the best K estimate was 3, with CV error = 0.37350. PCA and Admixture plots show that TEs can be useful predictors of the genomic structure of human populations, as previously highlighted by other Authors (Rishishwar et al., 2015; Gardner et al., 2017; Mallick et al., 2020), and that the results are coherent with those obtained using SNPs (1000 Genomes Project Consortium, 2015; Mallick et al., 2016; Byrska-Bishop et al., 2022).

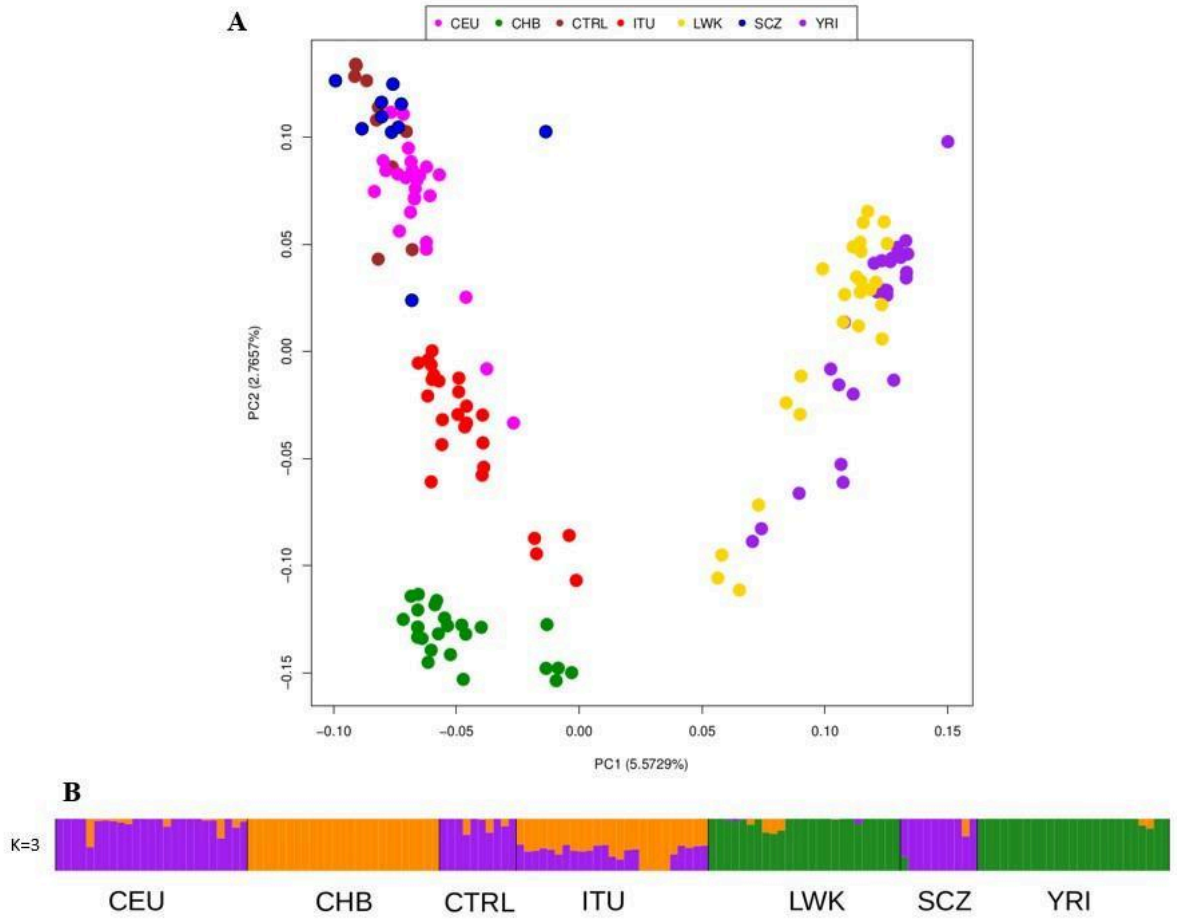



**Figure 5a.1.** PCA and Admixture plots based on TEs. DLPFC samples (blue and brown in the PCA) clearly cluster with European samples (as represented by CEU) and have the same ancestral components (violet, in Admixture plot), with the single exception of one SCZ sample that shows signs of admixture with a sub-saharan African source (represented by LWK and YRI).

7,338 TEs (93%) identified in the 20 DLPFC individuals are also present in the 125 samples from 1KGP, suggesting their germline rather than somatic origin; they are also useful predictors of the genomic structure of different human populations, as confirmed by the intermediate position of Indians (ITU) between Europeans (CEU) and Chinese (CHB), as well as by the clear differentiation between Eurasian and African samples. The remaining 7.3% of the TEs we detected as singletons only in the DLPFC samples (n=534): 35 SVAs (10% of total non-reference SVAs; 1.75/subject), 56 LINE1s (5%; 2.8/subject) and 473 Alus (7%; 23.6/subject). These TEs are unique and not shared across other samples nor are they listed in known reference databases, like euL1db (Mir et al., 2015)

and gnomAD (Collins et al., Nat 2020). They may be regarded as somatic insertions or they may still be germline TEs with low frequencies, since we cannot distinguish between the two possible origins. Even in case they are somatic rather than germline events, they still represent a minority of our observed TEs.

We also checked for allele frequencies (AFs) of TEs identified in DLPFC in the five populations of the 1KGP: TEs show systematic differences in AFs across populations: 3,131 of 7,952 TEs (2,711 Alu (41.4%), 332 LINE-1 (31.1%) and 88 SVA (25.5%)), have a significant geographic stratification (Fisher p-value < 0.05), with 2,263 (28%) presenting an allele frequency > 5%. Among these, 1,501 TEs are found only in African populations (Luhya and Yoruba), 833 are exclusive of non-African populations (Europeans, Indian Telugus, and Chinese) and 955 are common to all five groups (Figure 5a.2).

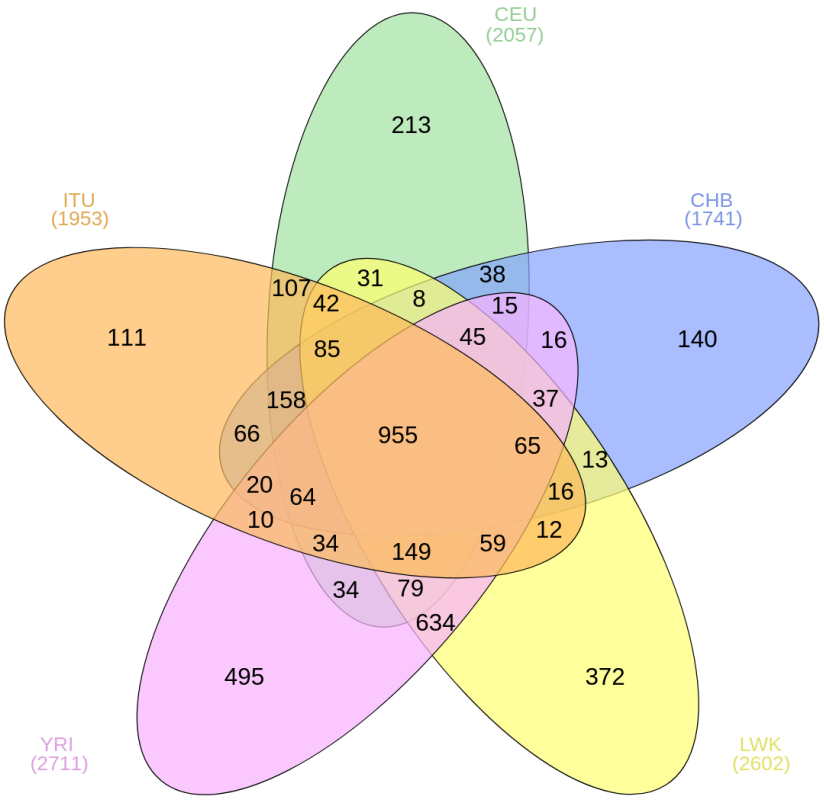


**Figure 5a.2**. Venn diagram showing the number of TEs with an allele frequency > 5% found across the five considered populations.

38 TEs were identified as potentially contributing to schizophrenia: 3 LINE-1s, 3 SVAs and 32 Alus (Table 5a.1). Every significant TE belongs to evolutionarily recent families (L1Ta and AluY), with two exceptions: the L1 on chr12:126802943 (undetermined subfamily) and the Alu on chr7:141748320 (which belongs to the subfamily Sz, older than Y).

Interestingly, most of the 38 "differentiated" TEs are located into or nearby genes with neurodevelopmental functions, and notably, variants of at least seven genes have been already associated with an increased risk of developing schizophrenia: LRRC4C (Li et al., 2018), LRRC7 (Hathy et al., 2021; Hathy et al., 2020; Carlisle et al., 2011), ST8SIA4 (Curto et al., 2019; Volk et al., 2016; Krocher et al., 2014), MGAM (Wu et al., 2021), ADAMTS1 (Pantazopoulos et al., 2021), MIR548AJ2 (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014) and SCN5A (Spellmann et al., 2018; Roden, 2014; Roberts, 2006).

| Type | Family | Chr. | Bp | Gene | Location | SCZ(-) | SCZ(+) | CTRL(-) | CTRL(+) | Fisher | SCZ(--) | SCZ(-+) | SCZ(++) | CTRL(--) | CTRL(-+) | CTRL(++) | Fisher |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SVA-INS | SVA | 9 | 33130559 | B4GALT1 | Intronic | 12 | 8 | 16 | 4 | 0,301 | 2 | 8 | 0 | 7 | 2 | 1 | 0,023 |
| SVA-INS | SVA | 11 | 73527418 | MRPL48 | Intronic | 20 | 0 | 15 | 5 | 0,047 | 10 | 0 | 0 | 5 | 5 | 0 | 0,033 |
| SVA-INS | SVA | 20 | 5268423 | PROKR2 | Terminator | 7 | 13 | 16 | 4 | 0,010 | 1 | 5 | 4 | 6 | 4 | 0 | 0,017 |
| LINE1-INS | L1Ta1d | 3 | 38626082 | SCN5A | Intronic | 20 | 0 | 14 | 6 | 0,020 | 10 | 0 | 0 | 4 | 6 | 0 | 0,011 |
| LINE1-INS | L1Ta | 9 | 32463887 | DDX58 | Intronic | 4 | 16 | 11 | 9 | 0,048 | 1 | 2 | 7 | 3 | 5 | 2 | 0,166 |
| LINE1-INS | L1Ambig | 12 | 126802943 | LINC02347 | Promoter | 17 | 3 | 13 | 7 | 0,273 | 8 | 1 | 1 | 3 | 7 | 0 | 0,020 |
| ALU-INS | AluYa3 | 1 | 70091072 | LRRC7 | Promoter | 15 | 5 | 11 | 9 | 0,320 | 6 | 3 | 1 | 1 | 9 | 0 | 0,020 |
| ALU-INS | AluYa4 | 1 | 163314443 | NUF2 | Intronic | 12 | 8 | 6 | 14 | 0,111 | 5 | 2 | 3 | 0 | 6 | 4 | 0,038 |
| ALU-INS | AluYb6 | 2 | 9888801 | TAF1B | Promoter | 9 | 11 | 2 | 18 | 0,031 | 1 | 7 | 2 | 0 | 2 | 8 | 0,023 |
| ALU-INS | AluYb7 | 2 | 36476695 | CRIM1 | Terminator | 15 | 5 | 20 | 0 | 0,047 | 6 | 3 | 1 | 10 | 0 | 0 | 0,087 |
| ALU-INS | AluYa4 | 2 | 114106446 | PAX8-AS1 | Terminator | 17 | 3 | 9 | 11 | 0,019 | 7 | 3 | 0 | 2 | 5 | 3 | 0,044 |
| ALU-INS | AluYa5 | 3 | 169951024 | PRKCI | Intronic | 12 | 8 | 18 | 2 | 0,065 | 2 | 8 | 0 | 8 | 2 | 0 | 0,023 |
| ALU-INS | AluYa4 | 4 | 17150918 | QDPR | Terminator | 11 | 9 | 20 | 0 | 0,001 | 4 | 3 | 3 | 10 | 0 | 0 | 0,011 |
| ALU-INS | AluYb7 | 4 | 23511024 | MIR548AJ2 | Promoter | 13 | 7 | 20 | 0 | 0,008 | 3 | 7 | 0 | 10 | 0 | 0 | 0,003 |
| ALU-INS | AluYa4 | 4 | 154901048 | SFRP2 | Promoter | 15 | 5 | 11 | 9 | 0,320 | 5 | 5 | 0 | 5 | 1 | 4 | 0,047 |
| ALU-INS | AluYb | 4 | 183647531 | TENM3 | Intronic | 10 | 10 | 15 | 5 | 0,191 | 0 | 10 | 0 | 5 | 5 | 0 | 0,033 |
| ALU-INS | AluYb8 | 5 | 55689499 | ANKRD55 | Promoter | 14 | 6 | 8 | 12 | 0,111 | 4 | 6 | 0 | 0 | 8 | 2 | 0,043 |
| ALU-INS | AluYb8 | 5 | 84516075 | EDIL3 | Promoter | 20 | 0 | 15 | 5 | 0,047 | 10 | 0 | 0 | 5 | 5 | 0 | 0,033 |
| ALU-INS | AluYa | 5 | 86372695 | MIR4280 | Terminator | 7 | 13 | 7 | 13 | 1,000 | 3 | 1 | 6 | 0 | 7 | 3 | 0,013 |
| ALU-INS | AluYb | 5 | 100497396 | ST8SIA4 | Promoter | 13 | 7 | 5 | 15 | 0,025 | 5 | 3 | 2 | 1 | 3 | 6 | 0,137 |
| ALU-INS | AluYg5b3 | 5 | 159122155 | ADRA1B | Promoter | 10 | 10 | 18 | 2 | 0,014 | 1 | 8 | 1 | 8 | 2 | 0 | 0,005 |
| ALU-INS | AluYc1 | 6 | 75338236 | COL12A1 | Terminator | 17 | 3 | 10 | 10 | 0,041 | 7 | 3 | 0 | 2 | 6 | 2 | 0,103 |
| ALU-INS | AluYa5 | 6 | 166279907 | LINC00473 | Terminator | 11 | 9 | 18 | 2 | 0,031 | 3 | 5 | 2 | 8 | 2 | 0 | 0,095 |
| ALU-INS | AluYa5 | 7 | 141013590 | TMEM178B | Intronic | 20 | 0 | 15 | 5 | 0,047 | 10 | 0 | 0 | 6 | 3 | 1 | 0,087 |
| ALU-INS | AluSz | 7 | 141748320 | MGAM | Intronic | 10 | 10 | 13 | 7 | 0,523 | 0 | 10 | 0 | 4 | 5 | 1 | 0,033 |
| ALU-INS | AluYb6 | 8 | 56033091 | XKR4 | Intronic | 14 | 6 | 20 | 0 | 0,020 | 5 | 4 | 1 | 10 | 0 | 0 | 0,033 |
| ALU-INS | AluYb8 | 9 | 75542985 | ALDH1A1 | Intronic | 15 | 5 | 7 | 13 | 0,025 | 6 | 3 | 1 | 1 | 5 | 4 | 0,093 |
| ALU-INS | AluYb3a1 | 9 | 91099740 | SPIN1 | Terminator | 15 | 5 | 20 | 0 | 0,047 | 5 | 5 | 0 | 10 | 0 | 0 | 0,033 |
| ALU-INS | AluYb6 | 11 | 19382774 | NAV2 | Intronic | 13 | 7 | 19 | 1 | 0,044 | 4 | 5 | 1 | 9 | 1 | 0 | 0,057 |
| ALU-INS | AluYa5 | 11 | 40727097 | LRRC4C | Intronic | 20 | 0 | 12 | 8 | 0,003 | 10 | 0 | 0 | 2 | 8 | 0 | 0,001 |
| ALU-INS | AluYg6 | 11 | 76990585 | GDPD4 | Intronic | 13 | 7 | 19 | 1 | 0,044 | 4 | 5 | 1 | 9 | 1 | 0 | 0,057 |
| ALU-INS | AluYa3 | 12 | 31120751 | TSPAN11 | Intronic | 13 | 7 | 19 | 1 | 0,044 | 4 | 5 | 1 | 9 | 1 | 0 | 0,057 |
| ALU-INS | AluYa5 | 12 | 81315235 | LIN7A | Intronic | 16 | 4 | 16 | 4 | 1,000 | 6 | 4 | 0 | 8 | 0 | 2 | 0,043 |
| ALU-INS | AluYa3 | 13 | 108669030 | FAM155A | Promoter | 9 | 11 | 17 | 3 | 0,019 | 1 | 7 | 2 | 8 | 1 | 1 | 0,003 |
| ALU-INS | AluYc1 | 15 | 39691605 | C15orf54 | Terminator | 5 | 15 | 9 | 11 | 0,320 | 2 | 1 | 7 | 1 | 7 | 2 | 0,022 |
| ALU-INS | AluYe | 16 | 80010958 | MAF | Promoter | 15 | 5 | 20 | 0 | 0,047 | 6 | 3 | 1 | 10 | 0 | 0 | 0,087 |
| ALU-INS | AluYb | 17 | 60376780 | TBC1D3P2 | Promoter | 18 | 2 | 13 | 7 | 0,127 | 9 | 0 | 1 | 3 | 7 | 0 | 0,003 |
| ALU-INS | AluYe | 21 | 28221356 | ADAMTS1 | Promoter | 17 | 3 | 11 | 9 | 0,082 | 8 | 1 | 1 | 2 | 7 | 1 | 0,009 |

**Table 5a.1.** List of 38 TEs that are significantly "differentiated" between SCZ and CTRL: 3 SVAs, 3 LINE-1s and 32 Alus. Presence or absence of the TE is defined by "+" and "-", respectively. Genotypes are displayed as follows: "++" = homozygous, "+-" = heterozygous, "--" = absence.

The most significant allele-wise results (nominal p-value < 0.01) among the 38 significant TEs include three Alus and one SVA, whose insertions can be found on chr4:17150918 and chr4:23511024 (both AluY and only observed in SCZ), chr11:40727097 (AluYa5, only observed in CTRL) and chr20:5268423 (SVA, more frequent in SCZ). The three Alus show a statistically significant geographical distribution (Figure 5a.3, A-B-C), presenting variable insertion frequencies across populations, while the SVA on chr20:5268423 (Figure 5a.3, D) has similar allele frequencies in all the considered populations.

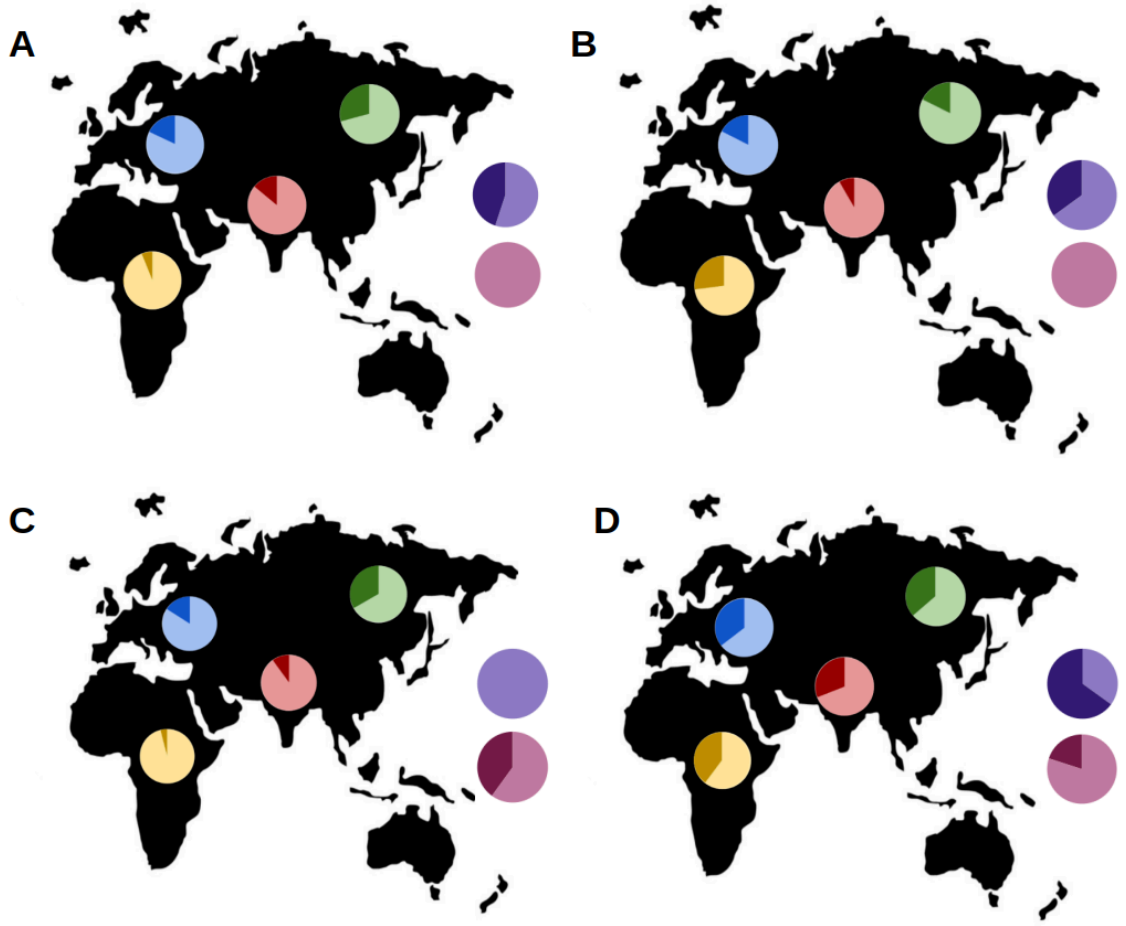



**Figure 5a.3**. Allele frequencies of Alus on chr4:17150918 (A), chr4:23511024 (B) and chr11:40727097 (C), compared to SVA on chromosome chr20:5268423 (D). Darker colors indicate the presence of the TE (+), while lighter colors indicate the absence (-). Following populations are displayed: Europeans (blue), Indian Telugus (red), Chinese (green) and Africans (yellow), represented by Luhya in Kenya and Yoruba in Nigeria. Allele frequencies for schizophrenic individuals and healthy controls are shown in violet and pink, respectively.

The haplotype-based analysis retrieved two significant results. The first one is a 188bp haplotype including the AluYb on chr5:100497396, in the promoter of the gene ST8SIA4. This haplotype contains four variants: T+TG (where the letters indicate SNPs and "+" stands for the presence of the TE) and is present with 2 copies in SCZ and 15 in CTRL, suggesting a strong association (p-value = 6.86 * $E^{-5}$) between the presence of the haplotype and the absence of the disease. The second significant result is a 1,172bp haplotype GC-TTI (where "-" indicates the absence of the TE and "I" is an InDel) including the AluYb7 on chr4:23511024 in the promoter of MIR548AJ2: interestingly, such haplotype was found with 19 copies in CTRL and 6 copies in SCZ (p-value = 3.93 * $E^{-5}$).

Accordingly, the Alu is completely absent in CTRL samples (Table 2) and present in 7 SCZs, only in heterozygous condition. Notably, both genes have been already associated with schizophrenia.

As highlighted in the introduction, TEs can have functional effects on nearby genes: therefore, we verified if the "differentiated" TEs were also acting as expression or alternative splicing Quantitative Trait Loci (eQTL/sQTL) by cross-checking the list of the 38 significant TEs with the lists produced by Cao et al. (2020), based on the GTEx dataset. 27 TEs (3 LINE-1s and 24 Alus) were detected as potential eQTLs, 7 of which in the brain. As for sQTLs, 13 TEs (2 LINE-1s and 11 Alus) were detected as potentially contributing to alternative splicing in different tissues, 2 of them supposedly acting in the brain (chr11:76990585 and chr2:36476695).

Lastly, we checked if some of the "differentiated" TEs are also located into Human Accelerated Regions (HARs), as originally identified by Pollard et al. (2006a, b), Prabhakar et al. (2006), Bird et al. (2007), Capra et al. (2013) and Gittelman et al. (2015). 12 of the "differentiated" TEs fall into genes that are located into as many HARs (ADAMTS1, ANKRD55, CRIM1, EDIL3, LRRC4C, LRRC7, MAF, NAV2, QDPR, TENM3, TSPAN11, XKR4). Notably, variants of three genes (ADAMTS1, LRRC4C and LRRC7) have already been associated with an increased risk of developing schizophrenia, and the three corresponding TEs were identified as significant in the differential allele frequencies analysis.

**Discussion**

Recent studies have shown that TEs can have both a positive and a detrimental role in shaping human cognitive traits (Guffanti et al., 2016) and in the development of brain and Central Nervous System (CNS) structures (Suarez et al., 2018; Ferrari et al., 2021). Moreover, several Authors, focusing on the study of reference TEs, suggested their contribution to different cognitive and neurological disorders, such as schizophrenia (Bundo et al., 2014; Guffanti et al., 2018; Misiak et al., 2019). Therefore, in this work (Modenini et al., 2023) we aimed at providing the first investigation on the contribution of non-reference (polymorphic) TEs to an increased risk of developing schizophrenia.

Firstly, we evaluated whether TEs contributed to population substructure by inspecting the distribution of these elements in our 20 DLPFC samples compared to 125 individuals from the 1KGP. PCA and Admixture analyses based on TEs (Figure 5a.2, A and B) show that SCZs and CTRLs fall within the European genetic variability (as represented by CEU) and that they contribute to the higher genomic diversity of African (YRI and LWH) compared to non-African populations. Indeed, TEs are present with significantly different allele frequencies across populations within our worldwide dataset, similarly to the well-known patterns previously detected in SNP-based studies (1000 Genomes Project Consortium, 2015; Mallick et al., 2016; Byrska-Bishop et al., 2022). Furthermore, SCZ and CTRL samples show a clear genetic homogeneity, which allows to exclude spurious associations due to hidden population substructure. Only a limited number of TEs (7.3%) are present in one individual: they may be considered either somatic or low-frequency germline retrotranspositions because they are not shared across samples and, thus, these single events should be better considered private insertions (or singletons). The total number of singletons and their proportion per individual are concordant with the estimates provided by Watkins et al. (2020) for the Caucasian/European group.

The standard allele frequency-based association methods highlighted that 38 TEs have significantly different AFs between SCZ and CTRL (Table 5a.1). Even considering the low sample size, we observe that AF differences for these TEs are similar or even higher than those observed between the most diverse "control" populations (i.e., the five 1KGP groups) for the same insertions, making it highly improbable that the observed differences emerged by chance. Admittedly, our results are based on a relatively low number of subjects, however the previously mentioned population genetic analyses (PCA and Admixture) show that our 20 DLPFC samples overlap with Europeans, indicating that spurious associations due to population substructure can be excluded. Therefore, these 38 TEs constitute our set of putative candidates for an increased risk of developing schizophrenia. Focusing only on the most significant results based on differential AF (p-value < 0.01), four candidates came to light: three Alus and one SVA on chr4:17150918, chr4:23511024, chr11:40727097 and chr20:5268423, respectively. The AluYa5 on chr11:40727097 is completely absent in SCZ (but present in 27.4% of CTRL) (Figure 5a.3, C) and is located in the second intron of the Leucine Rich

Repeat Containing 4C (LRRC4C) gene, which is highly expressed in the frontal cortex and has been associated with a positive response to antipsychotic therapy with lurasidone in SCZ patients (Li et al., 2018). In this case, the absence of the insertion is preferentially associated with the schizophrenic condition (Table 5a.1). On the contrary, the other two Alus are present only in SCZ and the SVA on chromosome 20 is more frequent in SCZ subjects.

To strengthen our results, we reconstructed haplotypes of the 20 subjects and performed an association test on the 38 candidate TEs, which retrieved two significant results. The AluYb7 on chr4:23511024, that was deemed significant for the AF test (p-value < 0.01), is located in the promoter of MIR548AJ2, one of the 108 genes that have been identified by GWAS studies as potentially contributing to schizophrenia (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). The haplotype (GC-TTI) is characterized by the absence of this Alu, which is indeed completely absent in CTRLs and present in 7 SCZs, only in heterozygous conditions. Therefore, our hypothesis is that the presence of the element is putatively related to an increased risk of developing schizophrenia. On the contrary, the haplotype with the AluYb on chr5:100497396 (T+TG), in the promoter of the gene ST8SIA4, is found with 15 copies in CTRL and 2 in SCZ, suggesting a strong association (p-value = $6.86 * E^{-5}$) between the presence of the TE and the absence of the considered trait. Accordingly, we could hypothesize that the haplotype with the TE has a protective role against the disease. Further in vitro or in vivo experiments could elucidate this potential relationship.

As highlighted in the previous paragraphs, TEs can act in *cis* by, for example, altering the expression of a gene or by having an impact on its alternative splicing; moreover, TEs could change the local functional architecture of HARs in schizophrenia and bipolar disorder (Erady et al., 2022). Therefore, we cross-checked our significant results with: 1) the lists of polymorphic TEs acting as eQTL/sQTL produced by Cao and colleagues (2020) based on the GTEx dataset; 2) the lists of known HARs potentially controlling for cognitive functions. Interestingly, 7 TEs act as eQTLs in the brain: for instance, the AluYb3a1 on chr9:91099740 is an eQTL in the frontal cortex and is located in the

terminator of SPIN1 (Spindlin 1). Two Alus have also been identified as potentially contributing to alternative splicing in brain tissues: the AluYg6 chr11:76990585 and the AluYb7 on chr2:36476695. These two Alus are located in the third intron of GDPD4 (Glycerophosphodiester Phosphodiesterase Domain-Containing Protein 4) and in the terminator of the CRIM1 gene, respectively. The AluYb7 on chr2:36476695 acts as sQTL in the frontal cortex, and CRIM1 encodes for the cysteine-rich neuron motor 1 protein, which is developmentally regulated and involved in CNS development and organogenesis (Kolle et al., 2000), other than being part of the HAR-genes that are functionally relevant in brain networks implicated in cognition (Wei et al., 2019).

In conclusion, our analysis provides the first overview of TEs as genetic variants that are possibly related to an increased risk of developing schizophrenia: therefore, these findings suggest that a neurodevelopmental genetic mechanism is at play in the etiopathogenesis of this complex disorder. Our findings are based on a relatively low number of individuals, but the identification of two TEs being further confirmed by highly significant haplotype-based analyses gives the opportunity to hypothesize that a similar framework applied to a larger cohort of subjects could confirm and possibly extend our results, and experimental validation of the identified TEs will elucidate their effective impact on this highly complex disorder.

## 5b. Searching for signatures of positive selection and adaptive introgression in human brain

Paolo Abondio[1,*], Giorgia Modenini[2,*], Alessio Boattini[2], Fabio Macciardi[3]. *Polymorphic transposable elements under positive selection participate in the recent and ongoing evolution of the human cognitive genome*. Draft version.

Affiliations:
1: IRCCS Istituto Delle Scienze Neurologiche Di Bologna, Bologna, Italy
2: BiGeA Department, University of Bologna, Bologna, Italy
3: Department of Medical Education (Neuroscience), CUSM, Colton, CA, USA
*: these authors contributed equally to the study

### Background

Signatures of positive selection in the genome are a characteristic mark of adaptation that can reveal an ongoing, recent, or ancient response to environmental change throughout the evolution of a population. New sources of food, climate conditions, and exposure to pathogens are only some of the possible sources of selective pressure, and the rise of advantageous genetic variants is a crucial determinant of survival and reproduction (Abondio et al., 2022).

There is an open debate about the influence of positive selection on genetic variants involved in the development of human cognitive functions and diseases. For instance, Srinivasan and colleagues (2016) and Polimanti & Gelernter (2017) suggested that recent instances of positive selection acted on variants that confer risk of developing schizophrenia or autism spectrum disorder (ASD). On the contrary, recent works by Yao (2020) and Gonzàlez-Peñas (2023) propose that not only there is no evidence of recent positive selection on autism/schizophrenia risk variants, but also that the protective alleles are those actually under positive selection, a condition caused by the "non-antagonistic pleiotropy" mechanism.

In the last decades, mounting evidence on the impact of TEs on human brain physiology seems to point towards a relevant role of the mobile genome in shaping the architecture of the neural tissue (Guffanti et al., 2018). Furthermore, an increasing number of studies suggest that TEs also contribute to the emergence of different neurodevelopmental and neurodegenerative disorders such as schizophrenia, bipolar disorder, autism spectrum disorder (ASD), Alzheimer's disease and Parkinson's disease (recently reviewed in DeRosa et al., 2022; see also Guffanti et al., 2016, 2018; Macciardi et al., 2022; Peze-Heidsieck et al., 2022; Ravel-Godreuil et al., 2021). In our recent work,

we identified several polymorphic TEs that possibly contribute to the development of schizophrenia and, in general, to the cognitive genome structure (Modenini et al., 2023). Therefore, here we present a follow-up study on our previously generated dataset of 20 genomes from the Dorsolateral Prefrontal Cortex (DLPFC: Guffanti et al., 2018; Modenini et al., 2023), using a completely different approach to test whether the previously identified polymorphic retrotransposons have been subject to recent instances of positive selection, as also suggested by other studies (Srinivasa et al., 2016; Polimanti & Gelernter, 2017).

Retrotransposons, and in general TEs, have been the subject of numerous studies (Langmüller et al., 2023) that highlighted how these elements have been considered as genome "parasites" for quite a long time and are usually under negative or purifying selection because of their mutational effects on the host's genome (reviewed in Bourgeois & Boissinot, 2019). Only recently, Rishishwar and colleagues (2018), found that positive selective pressures act on TEs in different human populations by identifying 169 polymorphic TEs under positive selection in European and Asian populations.
In combination with positive selection, another evolutionary force shaped human genome and phenotypes: adaptive introgression (AI), a process in which "beneficial variants acquired from archaic humans may have accelerated adaptation and improved survival in new environments" (Racimo et al., 2015). Thanks to the sequencing of archaic hominins, in the last decade a growing number of publications revealed that: 1) *Homo sapiens* genomes carry signals of AI; 2) the introgressed sequences are derived from Neanderthals, Denisovans and from another still unknown "archaic species"; 3) the introgressed variants have an impact on human fitness, phenotypes and evolution (Racimo et al., 2015; Rotival & Quintana-Murci, 2020).

In this study, we aim to test whether complex structural variants particularly associated to brain-related genes - such as retrotransposons - could have recently experienced positive selective pressures and/or have been subject to more ancient events of introgression by: 1) applying three recently developed tests for positive selection (nSL, DIND, H12) to our previously generated dataset

of ten schizophrenic individuals and ten psychiatrically healthy controls; 2) testing the presence of AI using the recently published method VolcanoFinder (Setter et al., 2020).

**Materials and Methods**

Samples used in this study have been generated and managed as described in the previous chapter and in two published works (Guffanti et al., 2018; Modenini et al., 2023). Briefly, 20 DLPFC samples, 125 individuals from 1KGP, aligned to the same hs37d5 human reference genome, were analyzed with MELT v.2.1.5. In sum, 7,952 TEs (6,542 Alus, 1,065 LINE-1s and 345 SVAs) and 8,331,932 SNPs were used for this study.

Three independent and complementary tests, the nSL, DIND and H12 statistics, were computed to detect different typologies of selective events due to relatively recent instances (ca. 10,000 years ago onwards) of positive selection on the 20 DLPFC samples. In particular, with respect to other frequency- and haplotype-based tests (Tajimas's D, Fu and Li's F*, Fay and Wu's H, and iHS), the DIND statistic provides robustness to variation in sequencing coverage and to low sample sizes (Barreiro et al., 2009), while the nSL enabled to properly account for variation in recombination rates (Ferrer-Admetlla et al., 2014). For these purposes, we first filtered out variants showing derived allele frequency lower than 0.2, as they were proved to bias DIND results (Fagny et al., 2014), and we calculated DIND scores for each variant by using self-customized Python scripts and a window size of 100kb surrounding the variant. The selscan v1.1.0b package (DeGiorgio et al., 2014) was instead used to compute nSL scores for each variant by considering the decay of homozygosity along windows of maximum 4,500 consecutive loci, a gap scale of 20kb and a maximum gap size of 200kb. Moreover, a third test based on haplotype frequency, H12 (Garud et al., 2015), was performed on the same dataset, because of its ability to detect recent selective sweeps of varying softness. For this test, the analytical procedure was followed as described in the original publication: statistical scores were computed using a Python script provided by the Authors, considering a window size of 400 variants with a shift of 50, and associated to the variant in the center of each window. The Authors suggested using the median of the distribution as a threshold of significance; however, in this case the median value also

corresponds to the lowest possible H12 score. Therefore, a score threshold of 0.02875, corresponding to the next highest H12 score was considered. Then, a second Python script was applied in this context to detect H12 peaks according to significant score distribution across the genome, resulting in a collection of segments of variable length surrounding each peak. The segments associated with the top 1% peak scores were scanned for the presence of TEs and cross-checked with the results of the other selection tests.

Significant results and distribution of nSL, DIND and H12 scores (both on SNPs and TEs) were plotted with the CMplot R package (https://github.com/YinLiLin/CMplot; Yin et al., 2021).

A gene network analysis was carried out with the algorithm implemented in the *signet* R package (Gouy et al., 2017) to test a realistic approximation of polygenic adaptation on known pathways. Statistical values from the genome-wide distribution of DIND, nSL and H12 scores, obtained for the DLPFC dataset, were associated to each gene mapped by variants under selection. For each gene, the highest absolute values for DIND, nSL and H12 among those computed for all its variants were picked as the reference scores of that locus. The list of genes and associated scores was used as input for the *signet* algorithm based on results from each of the three tests, as each can inform the gene network analysis across different times and intensities of selection, as mentioned previously. The KEGG database (https://www.genome.jp/kegg/mapper/search.html) was used as input to provide fully annotated functional pathways to the algorithm. A process of simulated annealing, to discover the highest scoring subnetwork (HSS) for each pathway, was set for 20,000 iterations, and the same number was used for the production of the null distribution with which to determine the probability of finding these HSS. Gene networks were considered significantly targeted by selection at multiple loci contributing to the same biological function if they showed p-values $< 0.05$ after controlling for false discovery rate (FDR). Since simulated annealing is a probabilistic algorithm that approximates the best answer to an optimization problem (in this case, discovering the HSS), the procedure was repeated five times for each run, and only gene networks showing significant scores for at least four runs were considered noteworthy. Similarly, only genes belonging to these HSS, that appear at least

three times out of four runs, or four times out of five runs, are deemed relevant for their contribution to polygenic adaptation towards a specific biological function.
Once the results of the three tests performed on the DLPFC samples were obtained, a probabilistic network analysis was carried out to detect instances of polygenic adaptation on specific pathways as driven by the contribution of multiple genes belonging to significant HSS.

While the aforementioned tests look for recent and ongoing signatures of positive selection, a new statistical method detecting putative genomic regions under adaptive introgression (AI) was applied to account for the possibility that ancient genetic contributions from archaic admixture events could have provided novel TEs to the modern populations. VolcanoFinder v1.0 (Setter et al., 2020) detects AI sweeps by looking at the pattern of excess intermediate-frequency polymorphism they produce in the flanking region of the variant of interest. It considers a model in which part of the population under scrutiny has received genomic segment carrying an advantageous allele from a donor source, followed by a selection event that raised the overall frequency of both the haplotype around the introgressed allele and those around the same variant, if it was already present in the recipient population. We applied the VolcanoFinder algorithm on the DLPFC dataset by providing the proportion of ancestral and derived alleles for each polymorphic variant, the unnormalised site frequency spectrum and the grid of variants to be tested (i.e., all the polymorphic sites). We ran the program using the Model1 algorithm on polarized data and estimated, for each single variant, the genetic distance D (between donor and recipient populations) that would maximize the likelihood ratio.

Significant results for the three tests (nSL, DIND and H12) and from the VolcanoFinder analysis were compared with the lists of eQTL/sQTL identified by Cao and colleagues (2020) to look for TEs with a putative role in the brain. For this purpose, we selected only TEs carrying signals of positive selection and/or adaptive introgression that act as eQTL/sQTL in the following brain tissues: cerebellum, nucleus accumbens, caudate and putamen basal ganglia, anterior cingulate cortex (BA24), cerebellar

hemisphere, hypothalamus, frontal cortex (BA9), amygdala, substantia nigra, cortex and hippocampus (Cao et al., 2020).

A previous work by Rishishwar and colleagues (2018) identified 169 TEs under positive selection in European and Asian populations. Therefore, we compared DIND, nSL and H12 results with those provided by Rishishwar and colleagues (2018) to strengthen our findings.
Furthermore, we cross-checked the significant results from Modenini et al. (2023), this work and the lists of HAR-BRAIN genes identified by Wei et al. (2019).

**Results**

We searched for TEs putatively subjected to positive selection in our DLPFC dataset (20 genomes: 10 SCZ + 10 CTRL: Guffanti et al., 2018; Modenini et al., 2023). Analyses were based on haplotype profiles (as defined by SNPs) associated with each of the detected TEs. According to the nSL test, out of 7,313 total TEs for which the statistic was effectively computed, 306 (4.2%) showed putative signatures of selection in the DLPFC dataset. When considering the DIND statistics for positive selection, 36 (0.5%) TEs resulted under putative positive selection in DLPFC. However, when the frequency of the TE is taken into consideration and a lower threshold of 0.2 is applied, only 17 TEs (0.2%) identified through the DIND statistic are still showing a positive result. The H12 test found 160 TEs (2.2%) putatively under positive selection in the DLPFC dataset.
Considering all three tests together, we found 410 Alu, 54 LINE-1 and 38 SVA, for a total of 502 TEs under positive selection in our post-mortem brain samples: three TEs resulted under positive selection according to both nSL and DIND statistics, while 8 TEs gave signatures of selection according to nSL and H12. One element, the AluY on chr2:133058082 in the promoter of MIR663B, was simultaneously found under positive selection according to all the three tests.
By looking at scores distribution of the nSL test, it can be highlighted that almost every significant TE (Figure 5b.1, black dots) has a negative score and in general their distribution is shifted onto a negative score; this means that the selective pressures acted on the derived allele, which in the case of

non-reference polymorphic TEs is, by definition, the presence of the element. Instead, single nucleotide variant's scores (colors in Figure 5b.1) have a 0-centered distribution, as expected.

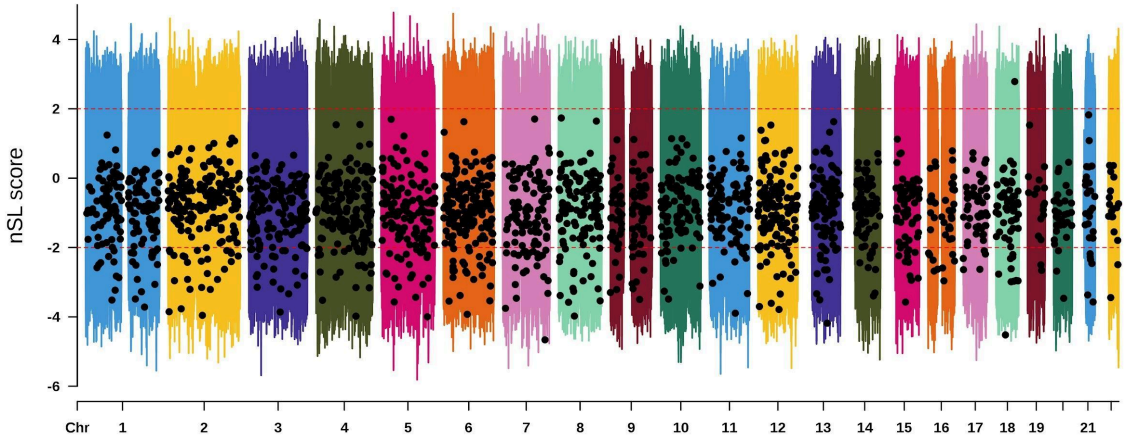



**Figure 5b.1**. Manhattan plot of the normalized nSL scores computed on TEs (black dots) and SNPs (background colors). Significant variants have a score greater/equal than 2 or lower/equal than -2.

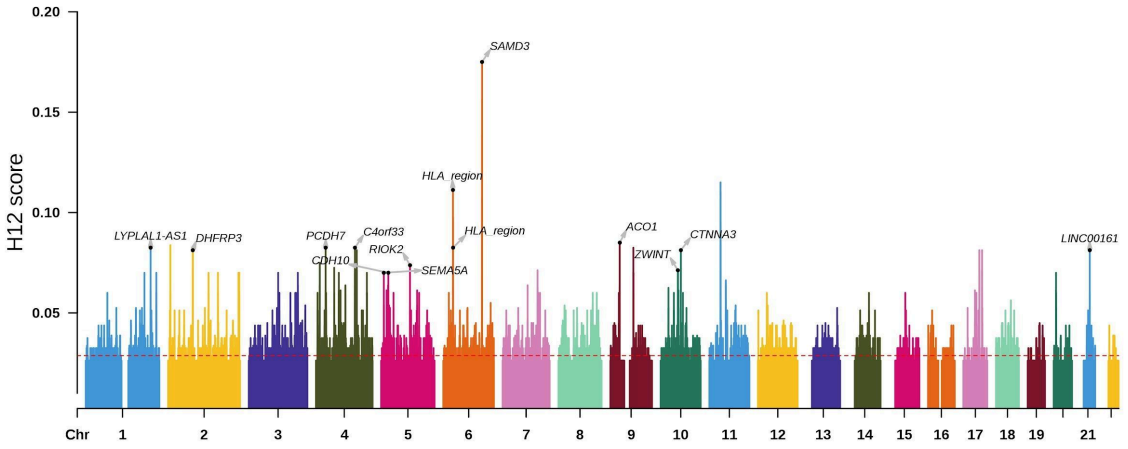



**Figure 5b.2.** Manhattan plot of the non-normalized H12 scores. Significant windows have a score ≥ 0.02875 (red-dotted line). Black dots represent TEs in significant windows: genes mapped by those TEs are highlighted.

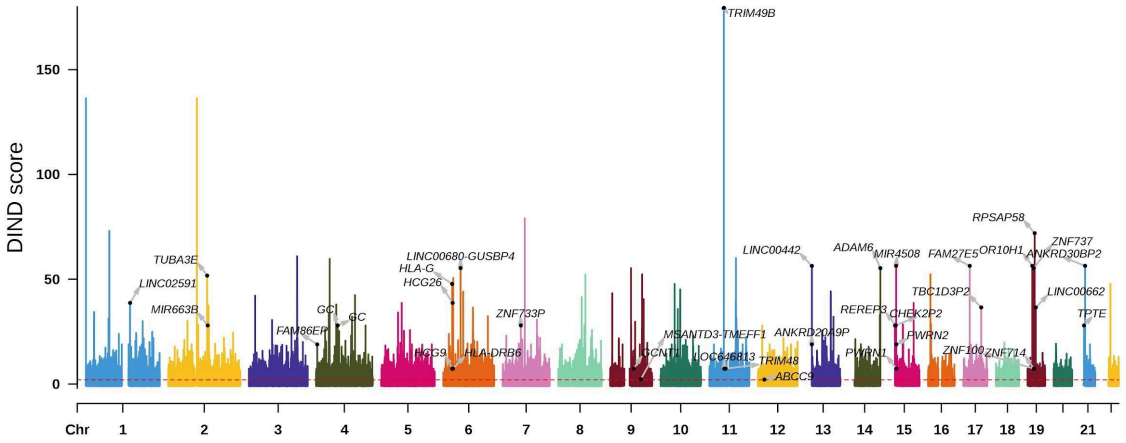



**Figure 5b.3.** Manhattan plot of the normalized DIND scores for SNPs (background colors) and TEs (black dots). Significant results (red-dotted line) have DIND scores ≥ 2. All genes mapped by significant TEs are annotated.

Results for the *signet* algorithm on the whole dataset of SNPs and TEs revealed one pathway confirmed by all three tests (nSL, DIND, H12): "Herpes simplex virus 1 infection"; on the other hand, one pathway was in common between DIND and H12: "Tight junction".

We cross-checked the significant results from Modenini et al. (2023), this work and the list of HAR-BRAIN genes as described in Materials and Methods (Table 5b.1). Four elements emerged as putatively under positive selection in the brain and with significantly different allele frequencies between schizophrenic and control individuals (Modenini et al., 2023). Two genes have also been identified as HAR-BRAIN genes, which means that they are more expressed in higher-order cognitive networks in humans compared to chimpanzees and macaques (Wei et al., 2019): TENM3 (Teneurin Transmembrane Protein 3) and EDIL3 (EGF Like Repeats And Discoidin Domains 3). The AluYb in TENM3 is present in all schizophrenic individuals and in half of the controls, only in a heterozygous condition; the other three Alus, on the contrary, are more frequent in controls. Finally, the AluYc1 in COL12A1 (Collagen Type 12 Alpha 1 Chain), that is under positive selection according to the nSL test and is a significant result for Modenini and colleagues (2023), acts as eQTL in brain cerebellum.

| Chr:position | TE type | Gene ID | HAR_gene | nSL | DIND | H1,2 | eQTL |
|---|---|---|---|---|---|---|---|
| 4:183647531 | AluYb | TENM3 | ✓ | | | ✓ | |
| 5:84516075 | AluYb8 | EDIL3 | ✓ | ✓ | | | |
| 6:75338236 | AluYc1 | COL12A1 | | ✓ | | | ✓ |
| 17:60376780 | AluYb | TBC1D3P2 | | | ✓ | | |

**Table 5b.1.** Transposable elements under positive selection in our brain samples and with significantly different allele frequencies between schizophrenic and control individuals, as found by Modenini et al. (2023).

Following the procedures described in the original VolcanoFinder paper (Setter et al., 2020), we looked at spurious peaks (Likelihood Ratio, LR > 20) and identified 5 peaks in the DLPFC dataset (Table 5b.2). The most significant result is for a peak corresponding to an AluYb6 located on chr20:32828034 (LR = 35.97) in the gene ASIP (Agouti Signaling Protein); the other peaks correspond to four single nucleotide variants.

| Chr | Position | LR score | Alpha | D | Gene ID |
|---|---|---|---|---|---|
| 3 | 100481209 | 20,715157 | 0,0006597 | 1,363773 | ABI3BP |
| 12 | 15548449 | 24,98861 | 0,0003995324 | 1,335753 | PTPRO |
| 12 | 15548498 | 24,133198 | 0,0003937782 | 1,335753 | PTPRO |
| 16 | 22487516 | 20,305244 | 0,0000476284 | 1,333521 | SMG1P1 |
| 20 | 32828034 | 35,974897 | 0,0001093879 | 1,160318 | ASIP |

**Table 5b.2**. Likelihood Ratio (LR) peaks from VolcanoFinder; “Alpha” and “D” values are also shown. The most significant result is an AluYb6 in ASIP’s promoter on chr20:32828034. “Chr” = chromosome; “Position” = position of the variant; “LR score” = Likelihood Ratio score; “Gene ID” = gene in which the variant is located.

Significant results for the three selection tests (nSL, DIND and H12) and from the VolcanoFinder analysis were compared with the lists of eQTL/sQTL identified by Cao and colleagues (2020) as described in Materials and Methods: we identified 74 eQTLs in the DLPFC group (52 for nSL, 7 for DIND and 15 for H12); as for sQTLs, we found 69 TEs (50 for nSL, 5 for DIND and 14 for H12).

The comparison between TEs identified by DIND, nSL, H12 and those identified by Rishishwar et al. (2018) highlighted that 12 TEs are common between the two studies (Table 5b.3).

| Chr:pos | Test | Rishishwar 2018 | Gene ID | Location | TE type |
|---|---|---|---|---|---|
| chr19:27836578 | DIND | chr19:27836578 | LINC00662 | TERMINATOR | AluYb6 |
| chr1:38447717 | nSL | chr1:38447717 | SF3A3 | INTRONIC | AluYa5 |
| chr1:51029967 | nSL | chr1:51029967 | FAF1 | INTRONIC | AluYc1 |
| chr3:163709222 | nSL | chr3:163709222 | LINC01192 | PROMOTER | AluYg6 |
| chr4:10632694 | nSL | chr4:10632694 | CLNK | INTRONIC | L1Ta1d |
| chr6:121077652 | nSL | chr6:121077652 | TBC1D32 | TERMINATOR | AluYa1 |
| chr7:12725160 | nSL | chr7:12725160 | ARL4A | PROMOTER | L1Ta |
| chr7:69969066 | nSL | chr7:69969066 | AUTS2 | INTRONIC | AluYb8 |
| chr9:76238787 | nSL | chr9:76238787 | ANXA1 | TERMINATOR | AluYb8 |
| chr10:43250102 | nSL | chr10:43250102 | BMS1 | PROMOTER | AluYa4 |
| chr16:5673584 | nSL | chr16:5673584 | RBFOX1 | PROMOTER | AluYe |
| chr11:10042452 | H12 | chr11:10042452 | SBF2 | INTRONIC | L1Ta1d |

**Table 5b.3**. Information about the 12 TEs common between Rishishwar et al., (2018) and this study. “Chr:pos” = chromosome and position of our TE; “Test” = test for positive selection; “Rishishwar 2018” = position of the TE identified by Rishishwar and colleagues, 2018; “Gene ID” = official gene symbol; “Location” = position of the TE; “TE type” = family/subfamily of the TE.

**Discussion**

Genes encoding brain-related proteins are among the most strongly conserved protein-coding genes in the human genome (Tuller et al., 2008). Conversely, several genes presenting signatures commonly associated with positive selection appear as causing brain diseases or conditions, such as dyslexia and autism (Dumas et al., 2021). However, the debate between researchers who point towards an influence of positive selection on genetic variants involved in the development of human cognitive functions/diseases, and researchers who claim the opposite, is still open. For example, two studies (Srinivasan et al., 2016; Polimanti & Gelernter, 2017) suggested that recent instances of positive selection acted on variants that confer risk of developing schizophrenia and autism. On the contrary, other studies by Yao (2020) and Gonzales-Peñas (2023) propose that the protective alleles are those actually under positive selection through a mechanism called “non-antagonistic pleiotropy”.

Some studies suggest that the primary substrate of evolution in the brain is regulatory changes in gene expression (Pollard et al., 2006; Changeux 2017) and splicing (Calarco et al., 2007): notably, TEs contribute in controlling the expression of genes at the transcriptional and post-transcriptional levels (for example, by acting as eQTLs or sQTLs), which is one of their significant functional effects on gene function and genome evolution (reviewed in Gebrie, 2023). In the past two decades, a great number of studies have revealed human-specific genomic changes, ranging from HARs to human-specific gene duplications, that constitute candidate species-specific modifiers of human brain development (Vanderhaeghen & Polleux, 2023). In particular, the prefrontal cortex is critical to many cognitive abilities that are considered specifically "human", and forms a large part of a neural system crucial for normal socio-emotional and executive functioning in humans and other primates (Teffer & Semendeferi, 2012).

In this work, we tested whether complex structural variants such as polymorphic retrotransposons could have recently experienced positive selective pressures and/or have been subject to more ancient events of adaptive introgression in individuals of European descent. To do so, we applied three different and recently developed haplotype-based tests for positive selection (nSL: Ferrer-Admetlla et al., 2014; DIND: Barreiro et al., 2009; and H12: Garud et al., 2015) and a method for the discovery of ancient adaptive introgression events (VolcanoFinder: Setter et al., 2020), another mutational force that shaped the human genome. We performed our tests on a previously generated dataset (Guffanti et al., 2018; Modenini et al., 2023) of 20 human genomes from the Dorsolateral Prefrontal Cortex (DLPFC) and compared them to other individuals of similar ancestry.

We successfully identified 502 polymorphic TEs putatively under positive selection in our dataset (410 Alus, 54 LINE-1s and 38 SVAs): of these, the AluY on chr2:133058082 in the promoter of MIR663B was found under positive selection according to the three tests (nSL, DIND and H12). For instance, a study on plasma biomarkers of Amyotrophic Lateral Sclerosis (ALS; Takahashi et al., 2015) evidenced a trend for an increase of MIR663B expression over time, even though when looking also at other miRNAs, their expression levels were not correlated with disease progression and changes in the patient conditions (Joilin et al., 2019). Another study (Gialluisi et al., 2016), which

investigated the possible correlation between copy number variants (CNVs) and reading/language performance, identified a large CNV (more than 500kb) covering the MIR663B region, possibly related to deficits in reading or speaking.

Two pathways emerged as significant from the *signet* algorithm: “Herpes simplex virus 1 infection” - common to all three tests (nSL, DIND and H12) - and “Tight junction”, in common between DIND and H12. Regarding the second pathway, six genes were found to be involved in “Tight junction”, even if none of them were in common between DIND and H12: PATJ (PATJ crumbs cell polarity complex component), AMOTL1 (angiomotin like 1), PPP2R2C (protein phosphatase 2 regulatory subunit Bgamma), PRKCZ (protein kinase C zeta), TUBA3E (tubulin alpha 3e) and TJP1 (tight junction protein 1).

Noteworthy, as highlighted in Figure 5b.1 (nSL score distribution of SNPs and TEs), the vast majority of the analyzed mobile elements have a significant negative score (305 TEs with a score $\leq$ -2), while only one TE is above 2. Technically, it means that these variants have experienced recent instances of positive selection on the derived allele, which in case of polymorphic TEs is, by definition, the presence of the mobile element (the ancestral state of these variants is known to be the absence of the insertion: Perna et al., 1992; Batzer et al., 1994). On the contrary, single nucleotide variants have a 0-centered distribution, with significant results having both positive and negative scores, as expected. TE sequences can be recruited by the host during adaptive processes and thus increase their frequency due to positive selection (recently reviewed in Bourgeois and Boissinot, 2019): indeed, the discovery of short interspersed elements - such as Alus - disproportionally present in gene-rich regions (Lander et al., 2001) is only explicable if these insertions have some benefits on the host genome and thus have been (and possibly still are) subject to positive selection (Oliver and Greene, 2009). However, it is not clear to what extent mobile elements can be the targets of such selective pressures (Villanueva-Cañas et al., 2017).

Four elements emerged as under positive selection in the brain and with significantly different allele frequencies between schizophrenic and control individuals (Modenini et al., 2023) (Table 5b.2). For example, the AluYb in TENM3 is under positive selection according to the H12 test and is present in all affected individuals and half of the controls, but only in a heterozygous condition. TENM3 stabilizes circadian rhythms by modulating the brain's response to light (Hunyara et al., 2023) and is "required presynaptically but not postsynaptically for the assembly of synaptic connections in the hippocampus" (Zhang X. et al., 2022). The TENM3 gene has also been identified as a HAR-BRAIN gene by Wei et al. (2019), who report that HAR-BRAIN genes are "key players in biological processes of nervous system development and neurogenesis"; moreover, selective pressures on cognitive networks promoting higher-order brain functions may have been associated with an increased risk of brain diseases (Crow, 1997; van den Heuvel et al., 2019). Findings from Wei and colleagues (2019) and from this study provide evidence for this hypothesis, with genes and related structural variants important for human brain evolution, found to play an important role in the onset of psychiatric disorders, such as schizophrenia.

To strengthen our findings, we compared the lists of TEs under positive selection for the nSL, DIND and H12 tests with the results from Rishishwar et al. (2018), who used fixation index (Fst) and population branch statistics (PBS) to detect signals of positive selection on polymorphic TEs: 12 elements are in common (Table 5b.3). For example, the AluYa4 in the promoter of BMS1 (Ribosome Biogenesis Factor 1), has been identified as under positive selection by H12, while previously detected as positively selected in Europeans by Rishishwar and colleagues (2018). This Alu acts as eQTL in the anterior cingulate cortex, suggesting a functional role for this element in regulating BMS1's expression in the brain. Indeed, according to the Human Protein Atlas (www.proteinatlas.org; Sjöstedt et al., 2020) the gene is characterized by medium expression in the adult neuronal cells of the Cerebral Cortex, while also being widely expressed in other tissues at the nuclear and nucleolar level. Interestingly, in a recent study (Yurdakul et al., 2023), ribosome biogenesis has been related to Autism Spectrum Disorder (ASD) and circadian clock: an upregulation of ribosomal protein-coding genes in ASD patients has been reported (Lombardo, 2021), suggesting an intensified ribosome biogenesis

related to neurodevelopmental disorders. Another relevant result is the AluYb8 in the gene AUTS2 (Autism Susceptibility Candidate 2), which is under positive selection according to nSL and is also detected by Rishishwar et al. (2018) in European samples. Variants of this gene have been widely studied for their putative causative role in ASD (first identified in Sultana et al., 2002), intellectual disability (Beunders et al., 2013), schizophrenia (Zhang B. et al., 2014; Ozsoy et al., 2020), attention deficit hyper-attention disorder (Talkowski et al., 2012) and addiction disorders (Schumann et al., 2011; Chen et al., 2013). According to the Human Protein Atlas, AUTS2 is expressed in several brain cell types, such as Inhibitory neurons, Oligodendrocytes, Excitatory neurons, Astrocytes and Microglial cells (www.proteinatlas.org; Sjöstedt et al., 2020). Furthermore, the first half of the AUTS2 sequence displayed the strongest statistical signal in a genomic screen differentiating modern humans from Neanderthals (Green et al., 2010); in addition, three evolutionary conserved noncoding intronic regions (HAR31, HACNS174 and HACNS369; Oksenberg et al., 2013) in AUTS2 have been found to be significantly accelerated when compared to primates (Pollard et al., 2006b; Prabhakar et al., 2006). Indeed, in line with our results on DLPFC samples, a more general analysis - such as that of Rishishwar et al. (2018) - has also highlighted putative positive selective pressures on variants with a role in brain development and diseases. The study concludes that if some polymorphic TEs have in fact been subject to positive selection, they play some functional role for their host genomes, as the QTL analysis also suggests.

Then, we searched for events of adaptive introgression using the VolcanoFinder method (Setter et al., 2020) and identified 5 LR peaks in the DLPFC dataset (Table 5b.3). The most significant result is a peak corresponding to an AluYb6 located on chr20:32828034 (LR = 35.97) in the promoter of the gene ASIP (Agouti Signaling Protein). However, to our knowledge, limited literature exists about the role of ASIP in brain development or diseases (Jain et al., 2023), since so far the gene has been widely studied for its roles in obesity, type-2 diabetes, cancer, hair and skin pigmentation, particularly using mouse models of human diseases.

In conclusion, our study promotes the idea that complex structural variants, such as polymorphic retrotransposons, have experienced instances of positive selection and that these evolutionary forces acted on genes specifically involved in brain development and cognitive functions. Moreover, our findings also point towards the hypothesis that variants under positive selection could also have an important role in neurological and psychiatric disorders, such as schizophrenia. However, further studies are needed to elucidate the actual functional role of the identified TEs.

## 5c. Differentially expressed retrotransposons as biomarkers of Alzheimer's disease

Fabio Macciardi[1,#], Maria Giulia Bacalini[2], Ricardo Miramontes[3], Alessio Boattini[4], Cristian Taccioli[5], Giorgia Modenini[4], Rond Malhas[3], Laura Anderlucci[6], Yuriy Gusev[7], Thomas J. Gross[3], Robert M. Padilla[3], Massimo S. Fiandaca[3], Elizabeth Head[8], Guia Guffanti[9], Howard J. Federoff[8], Mark Mapstone[3]. *A retrotransposon storm marks clinical phenoconversion to late-onset Alzheimer's disease*. Geroscience. 2022 Jun;44(3):1525-1550. doi: 10.1007/s11357-022-00580-w

Affiliations:
1: Department of Psychiatry and Human Behavior, UCI, Irvine, USA
2: IRCCS Istituto Delle Scienze Neurologiche Di Bologna, Bologna, Italy
3: Department of Neurology, UCI, Irvine, USA
4: BiGeA Department, University of Bologna, Bologna, Italy
5: MAPS Department, University of Padova, Padua, Italy
6: Department of Statistical Sciences "Paolo Fortunati", University of Bologna, Bologna, Italy
7: Georgetown University Medical Center, Washington, DC, USA
8: Department of Pathology, UCI, Irvine, USA
9: Department of Psychiatry, McLean Hospital, Harvard Medical School, Boston, USA
#: correspondence to fmacciar@hs.uci.edu

### Background

In the previous chapters, we reported how non-reference TEs could have had (and still have) a strong impact on brain disorders (such as addictions and schizophrenia), and also on the recent evolution of cognitive traits. To complete our work, we also performed a study on the role of reference TEs in neurodegeneration, in particular by focusing on Alzheimer's disease (AD) onset.

Recent works suggested that the reactivation of (otherwise) transcriptionally silent TEs can impact the neural homeostasis during pathological aging and might induce brain degeneration, either by dysregulating the expression of genes and pathways implicated in cognitive decline and dementia or through the induction of immune-mediated neuroinflammation, resulting in the elimination of neural and glial cells (Frost et al., 2014; Colombo et al., 2018; Liu et al., 2019; Jonsson et al., 2020; Ochoa Thomas et al., 2020). Studies in Alzheimer's disease, and other tauopathies such as progressive supranuclear palsy (PSP), have shown alterations in TE expression profiles that suggest a potential involvement in Tau-dependent pathological mechanisms leading to neurodegeneration (Frost et al., 2014; Sun et al., 2018).

Widespread Tau-dependent chromatin decondensation leads to the re-expression of otherwise silenced TEs, without reactivating TE retrotransposition (Protasova et al., 2017), but altering their expression.

Such a Tau-induced expression of TEs has been associated with cognitive decline in manifest AD, in association with increased neurofibrillary tangles (NFTs) found in post-mortem AD brains, and in support of a proposed pathogenic role of TEs in neurodegeneration (Guo et al., 2018).

In this work we tested the hypothesis that differentially expressed TEs (DE-TEs) in blood could be used as biomarkers of cognitive decline and development of AD by evaluating the differential expression of TEs within a unique sample of subjects from a late-onset Alzheimer's disease (LOAD) cohort for which we have RNA-sequencing data obtained before and after their phenotypic conversion (*pheno*-conversion) from the pre-symptomatic to the symptomatic forms of the disease (Mapstone et al., 2014).

**Subjects and Methods**

To test our hypothesis, we used a previously validated RNA-based analytical pipeline (Guffanti et al., 2018) on 25 aging subjects (age ≥ 75) that developed late-onset Alzheimer's disease over a relatively short period of time (12-48 months), for which blood was available before and after their *pheno*-conversion, and a group of cognitive stable subjects as controls, represented by 64 age- and sex-matched individuals that have retained normal cognition along the whole 5 years of observation. Individuals were independent, community-dwelling older adults, without known diagnosis of AD or mild cognitive impairment (MCI), nor other major neurological or medical illnesses (Mapstone et al., 2014; Fiandaca et al., 2015). 525 participants were enrolled over the course of this 5-years study: each individual underwent a fasting blood draw and thorough neuropsychological testing at the time of entry and yearly thereafter, for a maximum of 6 visits. After year 3 of the study, a biomarker discovery cohort of participants that met strict neuropsychologically defined criteria for either normal cognition (NC), newly diagnosed amnestic MCI (aMCI), or AD were defined. In addition, 25 participants were identified as entering the study with normal cognition (NC), but over the course of the study developed criteria for aMCI and/or AD. The latter individuals were designated as ConverterPre (pre-conversion), while meeting NC criteria, and ConverterPost (post-conversion), once meeting the neuropsychologically defined criteria for either aMCI or AD.

RNA sampling, processing and storage, as well as the RNA expression analysis methods can be found in Macciardi et al., 2022. Briefly, prior to blood collection, participants' body and physiological measurements (height, weight, blood pressure, pulse, temperature), as well as list of current medications, and whether food or drink other than water before midnight had been consumed were recorded. Total RNA was extracted using the PAXgene Blood RNA Kit (# 762,164, Qiagen, Inc., Germantown, MD, USA), according to the manufacturer's instructions. The isolated blood-derived RNA was quantified using a NanoDrop ND-1000 spectrophotometer (Thermo Fisher Scientific, Inc., Waltham, MA, USA), cataloged, and stored at −80 °C until ready for further analysis. Total RNA specimens from selected subjects were shipped to Expression Analysis Inc. (EA, a Quintiles Company, Durham, NC, USA) for RNA sequencing (RNAseq) analysis. Extracted *fastq* files from the provided Illumina *bam* files were scanned to map and quantify the level of expression for each TE at their unique genetic locations with a previously developed and experimentally validated protocol (Guffanti et al., 2018). Our RNA mapping strategy for TEs is based on modifications of the Trinity Genome Guided (GG) assembly protocol (Grabherr et al., 2011; Haas et al., 2013). Firstly, using HISAT2 (Pertea et al., 2016), we aligned raw RNA reads to the TE reference genome, which was extracted from the Repbase/RepeatMasker database (v4.1.0) of the human genome GRCh38. In this first step, the goal was to sort out the reads that potentially mapped to the TE reference genome and discard the reads that did not. The selected reads were then separately submitted to the Trinity-GG algorithm, which assembled these reads into transcripts that represent the *de novo* assembled transcriptome for TEs. Using Megablast, each de novo assembled TE transcript was aligned to the RepeatMasker reference, and we filtered out all transcripts that show less than 95% identical matches and that align for less than 90% of their length with the reference TE.

The expression of each discrete TE transcript was quantified using Kallisto (v.0.43.0) (Bray et al., 2016), generating matrices with TPM (transcript per million) values, where TPM is the transcript count of each TE divided by the sum of the transcript counts of each sample, multiplied by one million. TPMs were cross-sample normalized for subsequent analyses with the TMM (trimmed mean of M values) normalization approach using the edgeR Bioconductor package (Robinson & Oshlack,

2010; McCarthy et al., 2012). The same package was used to test for differential expression (DE) of TEs in pre/post conversion samples (ConverterPre/ConverterPost): edgeR keeps only those transcripts that have at least 1 read per million in at least 2 samples, and uses the differential analysis of sequence read count data for paired samples for the comparison of data before and after the phenotypic onset of AD. First, a design matrix was generated without an interaction term. This was then applied to a generalized linear model to normalize expression data. Finally, likelihood ratio tests were performed for ConverterPre vs ConverterPost individuals.

To initially evaluate a possible functional role of the DE-TEs that were significant in the ConverterPre vs ConverterPost and ConverterPre vs NC comparisons, respectively, we looked at the annotations of their neighboring genes within the human genome hg38 with the software GREAT (McLean et al., 2010). To constrain our analyses to the hypothesis of a *cis*-regulatory function of TEs, we considered only those protein-coding genes that lie within a distance of 5,000 bp, either upstream or downstream of the TE, using all the transposable elements present in the RNA-sequencing analysis as background. Finally, using multiple annotation sources, an estimate of enrichment was determined for biological and molecular functions for those gene families with identified annotated genes using GREAT (http://great.stanford.edu/public/html/).

Then, to investigate the chromatin states of the DE-TEs in our sample, the Core 15-state model from the Epigenomics Roadmap website was used (https://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html#core_15state). Hg38 coordinates were converted to hg19 coordinates using the *liftOver* Bioconductor package (https://master.bioconductor.org/packages/release/workflows/html/liftOver.html), for compatibility with the reference genome of the Epigenomics Roadmap. TEs that mapped in genomic regions that did not successfully convert from hg38 to hg19 were removed. The Bedops software (Neph et al., 2012) was used to assess the overlap between TE coordinates and chromatin states from 8 different tissues including adult blood, adult brain regions (E062: peripheral blood mononuclear primary cells; E073: brain dorsolateral prefrontal cortex; E072: brain inferior temporal lobe; E067: brain angular

gyrus; E071: brain hippocampus middle; E074: brain substantia nigra; E068: brain anterior caudate; E069: brain cingulate gyrus), 3 fetal brain regions (E081: fetal brain male; E082: fetal brain female; E070: brain germinal matrix), and neuronal cultures (E007 and E009: H1 derived neuronal progenitor cultured cells; E010: H9 derived neuron cultured cells). The *mix.heatmap* function from the CluMix R package (Hummel et al., 2017) was used to generate heatmaps of the Core 15-state model analysis data. Similarities between subjects were measured by Gower's general similarity coefficient. Similarities between variables were based on distance correlation. Standard hierarchical clustering, with default Ward's minimum variance method, was applied to obtain dendrograms of the considered subjects. Variations among the considered 14 tissues were represented by applying Kruskal's non-metric MDS to distance correlation, as implemented in the isoMDS function from the MASS library of R software (Venables & Ripley, 2002). Colors for the 15 chromatin states were set using color codes provided by the Roadmap Epigenomics Project (https://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html). Tissues with multiple chromatin states were colored in blue and labeled as "Mx" (i.e., mixed chromatin states).

We also used the R-Bioconductor package *Monocle* (Qiu et al., 2017) to analyze time-dependent RNA trajectories and identify the *pre* to *post* transition path in the group of individuals that developed LOAD. We performed the analysis of TE RNA transcript expression values to sort individuals in a pseudotime order. After converting TPM values into RNA counts via the *relative2abs* function, implementing the algorithm called Census (Qiu et al., 2017), we normalized the RNA counts across transcripts via the *estimateSizeFactors* and *estimateDispersion* functions, and filtered out transcripts below the expression threshold of 0.1, while retaining transcripts expressed in at least 4 individual RNAs of the dataset. The time-dependent trajectory analysis was performed on a set of transcripts selected to be DE at the threshold q-value $< 0.01$ between time points comparing transcripts at the ConverterPre and ConverterPost conditions. After applying the data dimensionality reduction, using the Discriminative Dimensionality Reduction with Trees (DDRTree) method, the RNAs of all individuals were ordered along the trajectory using the *orderCells* function. The information on the collection time was leveraged to identify the start point of the pseudotime. Then, the identified start

state was used as root to reorder RNAs. To find transcripts that change as the RNAs make progress along the pseudo-temporal trajectory, we tested for DE transcripts as a function of the pseudotime, recording the progress of each RNA through the developmental path. To identify patterns of covariation of transcripts along the pseudotime, we used the *plot_pseudotime_heatmap* function that generates smooth expression curves for each transcripts and clusters them based on profile similarity.

In addition to the previous analyses, we were interested in identifying possible TE biomarkers for the comparisons "ConverterPre vs ConverterPost" and "ConverterPre vs NC" through a machine learning (ML) technique. Our goal was to identify those TEs that accurately discriminate between the groups of patients and predict their health state with significant accuracy. First, the Shannon entropy value of the expression levels for each TE was calculated, using a function created specifically for this study (see Supplementary Methods from Macciardi et al., 2022). Shannon's entropy is a measure that estimates the amount of information present within a message. This step allowed us to remove all the TEs from the dataset that did not have a sufficient amount of information. 10,000 TEs, ranked by the entropy of their level of expression, were retained for further analyses. The data were then organized in a matrix whose rows matched the individual samples and selected TEs were represented in the columns. This matrix was then used for selecting features based on the Boruta algorithm, with the Boruta R package applied to Random Forests (Kursa & Rudnicki, 2010). The parameters used to run Boruta were: $\text{p-value} \leq 0.05$, $\text{ntree} = 10{,}000$, $\text{maxRuns} = 100$. The remaining functionalities obtained by Boruta were used to discriminate those TEs that were representative for a particular class of patients with respect to any other (ConverterPre vs ConverterPost or ConverterPre vs NC). Next, our dataset was divided into a training and test set, selecting 70% and 30% of the samples, respectively, using the Caret and Ranger R packages (Kuhn, 2008; Wright & Ziegler, 2017) with the "down" parameter set to "TRUE" in order to take into account any variability imbalances within classes. The generalizability of our models was validated using a five times cross-validation and by the "train" function of Caret. Model performance was assessed using standard functions implemented in Caret. The estimation of the AUC values for the ConverterPre vs ConverterPost and ConverterPre vs NC (ROC curves) was generated using the pROC R package (Robin et al., 2011).

**Results**

Based on quality measures, 799,853 and 624,793 RNA transcripts were retained from the ConverterPre vs ConverterPost and ConverterPre vs NC comparisons, respectively. All transcripts are putatively mapping to the reference sequences of discrete TEs reported in RepeatMasker/Repbase (v4.1.0). After QC to remove transcripts that are mapping to multiple locations within the genome, the number of transcripts reduced to 424,511 (ConverterPre vs ConverterPost) and 489,694 (ConverterPre vs NC) elements, aligning to 338,447 and 373,159 unique reference TE loci, respectively. Approximately 10-14% of uniquely mapped reference TE loci belong to evolutionary recent TE families.

The DE analysis from the ConverterPre vs ConverterPost comparison identified 1,790 TEs with significant expression differences between these two timepoints (logFC ± 1.5, logCMP > 5.3, nominal p-value < 0.01). These DE transcripts mapped both over- and under-expressed TE elements: interestingly, LINEs, LTR elements and SVAs are significantly over-expressed, while SINEs are under-expressed. A PCA on the significant DE-TEs from the ConverterPre vs ConverterPost (n=1,790) and from the ConverterPre vs NC (n=503) revealed that there is no clear distinction between ConverterPre, ConverterPost and NC conditions, indicating that no systematic differences in TEs expression exist among the 3 groups (Figure 5c.1).

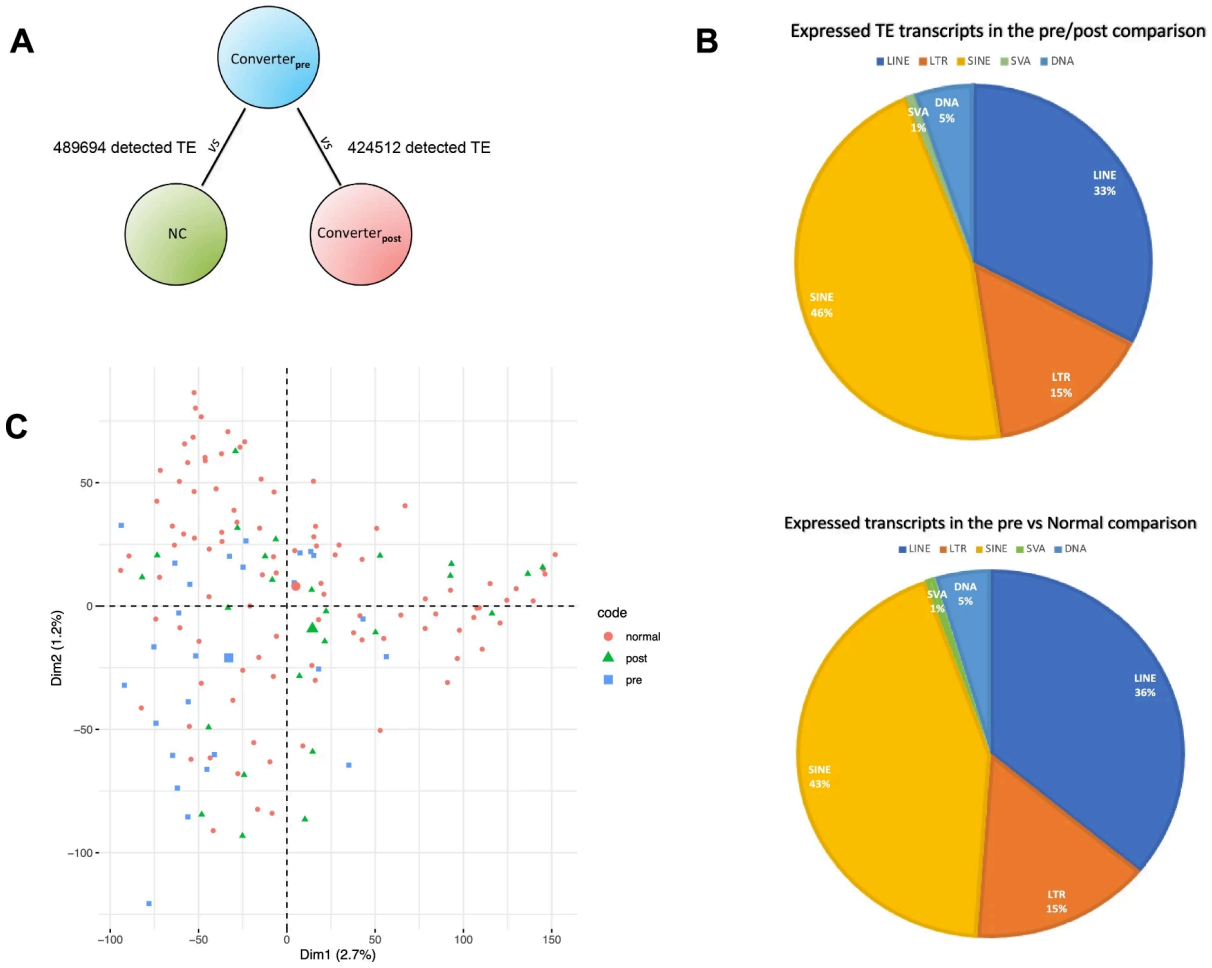



**Figure 5c.1. A)** A graphical representation of the comparisons with the QC numbers of observed TE-mapping transcripts in the ConverterPre vs ConverterPost and ConverterPre vs NC samples. **B)** The relative proportion of expressed TEs by classes in the 2 comparisons. **C)** The first 2 dimensions of PCA for normal, pre, and post subjects that do not present any preferential sub-clustering.

As introduced earlier, we identified 1,790 transcripts mapping to reference TEs that were DE between ConverterPre and ConverterPost samples: 1,543 with higher expression values and 247 with a lower expression values in ConverterPre than in ConverterPost states (Figure 5c.2: A and B). Up-regulated DE-TEs are significantly enriched in LINE and long terminal repeat (LTR) elements, while down-regulated DE-TEs are enriched in LINEs and composite repetitive elements (SVAs) (Figure 5c.2, C). Both up- and down-regulated DE-TEs were depleted in SINE (Alu) elements (Figure 5c.2, C). Within the over-expressed TEs, 70 are evolutionarily recent LINE-1s (L1HS, L1P1/2/3 or L1PA2-4) and at least one L1HS element on chr6:24811658-24817706 is insertionally polymorphic,

with highest frequencies in African populations from the 1KGP (Yoruba = 85.42%; Luhya = 84%; Europeans (CEU) = 46%; Chinese = 56%; Indian Telugus = 56%), and putatively acting as a weak enhancer of the RIPO2 gene. We also observed 24 evolutionary recent HERVs (HERVK-int, LTR5_Hs, LTR7) and 18 SVAs. Within the under-expressed TEs, LINE elements are a mix of evolutionarily recent and old elements, and 10% of LTRs are represented by HERV-K family elements.

Importantly, we cannot exclude *a priori* that these 1,790 TEs have been identified as differentially expressed in the ConverterPre vs ConverterPost comparison because of the longitudinal design of the study (i.e., these TEs have an age-dependent expression), independently of the conversion to aMCI/LOAD. To rule out this possibility, we evaluated whether the 1,790 DE-TEs showed an age-dependent expression in the NC group. NC subjects did not cluster according to their age for these TEs, and accordingly the first two components of the PCA calculated on the expression values of the 1,790 TEs did not show association with age in the NC group (Supplementary Figure S2). Collectively, these observations suggest that the differential expression of the 1,790 TEs that we identified in the ConverterPre vs ConverterPost comparison cannot be simply ascribed to the fact that subjects were evaluated at two timepoints.

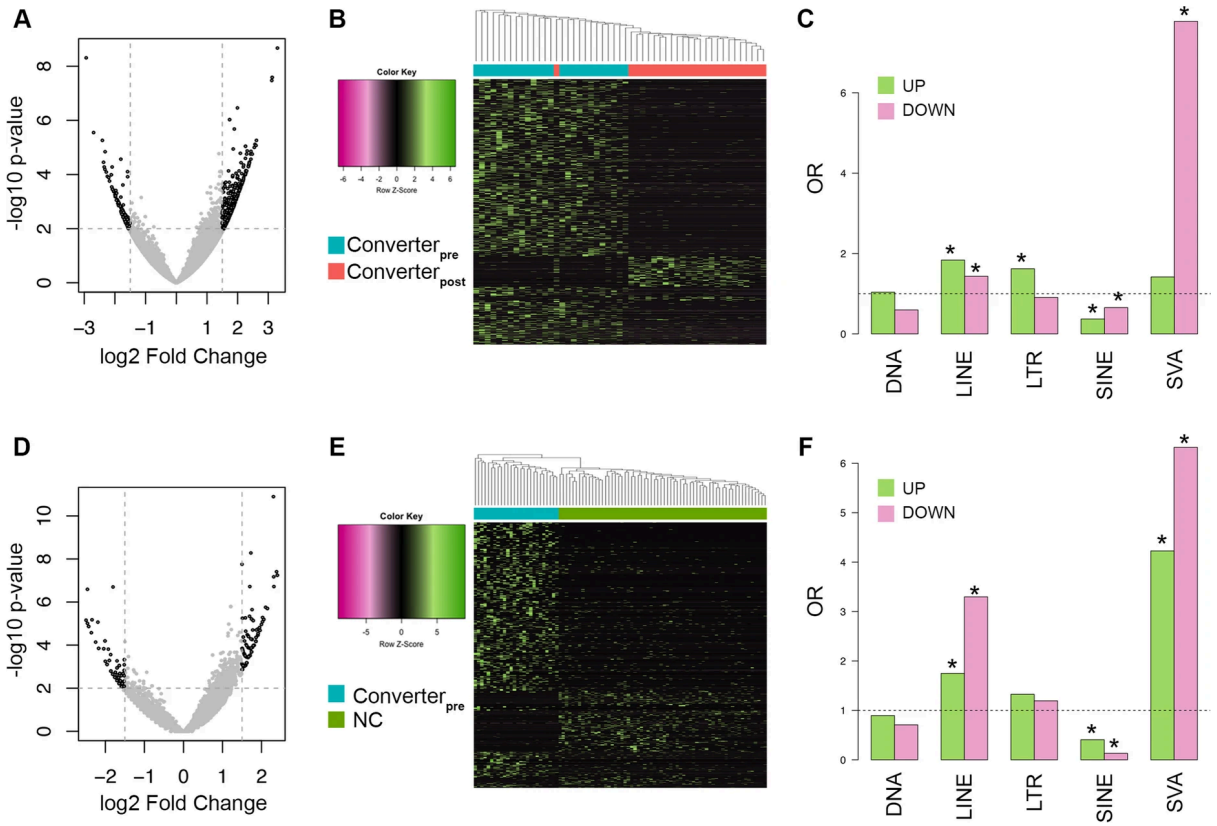



**Figure 5c.2**. Differential analysis of TE expression. **A)** Volcano plot of the results from the differential expression analysis of TE transcripts in the ConverterPre vs ConverterPost comparison. Significant RNA transcripts at $\log FC \pm 1.5$ and p-value $\leq 0.01$ are highlighted in black. **B)** Scaled heatmap and unsupervised hierarchical clustering of the log2 TMM values of the 1,790 TEs identified as differentially expressed in the ConverterPre vs ConverterPost comparison. Samples are annotated with different colors according to the group (ConverterPre or ConverterPost). **C)** Enrichment analysis for up- and down-regulated differentially expressed TE transcripts according to their class. Stars mark significantly enriched TE classes (Fisher's exact test p value $\leq 0.01$). **D, E, F)** The panels report the same plots described above, but for the ConverterPre vs NC comparison.

The analysis with GREAT on the ConverterPre vs ConverterPost comparison shows that the most enriched biological families (adjusted p-value $< 0.01$) were related to molecular pathways already known to be involved in AD, such as "negative regulation of autophagosome", "negative regulation of autophagy", and "positive regulation of dopamine receptor signaling". Instead, the most enriched gene families in the ConverterPre vs NC comparison show an involvement in the "cellular protein modification process", "protein modification process" and "macromolecule modification" (with

enrichment in molecular functions related to “regulation of skeletal muscle fiber development”, “regulation of myotube cell development” and “negative regulation of proteasomal activity”).

To characterize the potential peripheral blood biomarkers of early neurodegeneration, we considered the chromatin states of the genomic regions harboring the DE-TEs, both in blood and brain tissues, according to Epigenome Roadmap data. For the ConverterPre vs ConverterPost comparison we found that 67% of the over-expressed transcripts and 56% of the under-expressed transcripts overlap or intersect with signatures of functionally active chromatin states in blood cells (Figure 5c.3, A). Interestingly, we found that over-expressed TEs in the ConverterPre vs ConverterPost comparison are significantly enriched in TxWk (actively transcribed) and Enh (enhancer) chromatin states. For the ConverterPre vs NC comparison, 62% of the over-expressed DE-TEs mapped in genomic regions with active chromatin marks in blood cells, while down-regulated TEs mapped mainly to inactive regions and only 40% of down-regulated DE-TEs mapped in active chromatin regions (Figure 5c.3, B). We then considered the chromatin state of the DE-TEs using the Epigenome Roadmap Core 15-state model from brain regions. We found that adult brain tissues and fetal brain/germinal tissues, despite organizing in two distinct clusters, show a rather similar profile of active and quiescent chromatin regions (Pehrsson et al., 2019) to those characterizing the peripheral blood cells (Figure 5c.3, C and D). This observation suggests that at least a fraction of the DE-TEs that we identified in whole blood have a similar epigenetic regulation in brain tissues. Indeed, we found that 61% of the DE-TEs in the ConverterPre vs ConverterPost comparison and 67% of the DE-TEs in the ConverterPre vs NC comparison were also expressed in the human dorsolateral prefrontal cortex (DLPFC) using previous data generated by our lab (Guffanti et al., 2018).

Most of the chromatin regions that overlap with the significant DE-TEs and presenting with an active Core 15-state model (suggesting a possible functional role as either enhancers or promoters) are also functionally active within adult brain tissues as well as fetal brain tissues. In the ConverterPre vs ConverterPost comparison, 11 DE-TEs are marked as enhancers (7_Ehn, yellow) in all adult brain tissues, but not in peripheral blood. Some of these DE-TE insertions map onto genetic regions linked to Alzheimer’s disease. For example, a LINE-2 on chr1:10075287-10075497 maps within the second

intron of the UBE4B gene (Gireud-Goss et al., 2020), and a LINE-2 on chr7:105246376-105246653, found in the ConverterPre vs NC comparison, lies within the SRPK2 gene (Wang ZH et al., 2017). Moreover, a LINE-1 on chr6:36594353-36605600 in ConverterPre vs ConverterPost comparison was found to be transcriptionally active (1_TssA, red) in all adult brain tissues, and maps within the SRSF3 gene, known to regulate the innate immune response in resident microglia (Boutej et al., 2017).

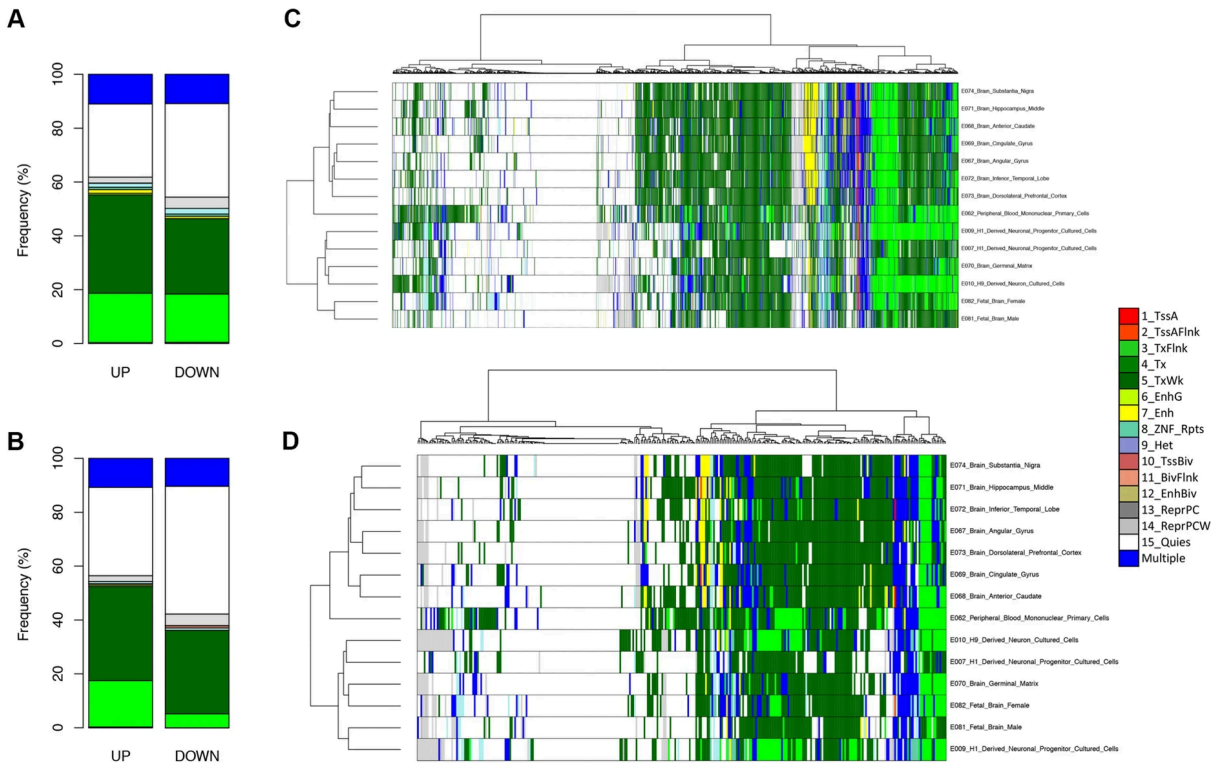



**Figure 5c.3**. Chromatin states of DE-TEs. **A-B)** show the distribution of up- and down-regulated DE-TEs across the chromatin states included in the Core 15-state model in blood cells, considering ConverterPre vs ConverterPost (A) and ConverterPre vs NC (B) comparisons. **C-D)** are the heatmaps with unsupervised clustering of the chromatin states in blood cells, and adult and fetal brain tissue considering the genomic regions overlapping with DE-TEs from ConverterPre vs ConverterPost (C) and ConverterPre vs NC (D) comparisons. In all the plots, colors of the chromatin states are shown in the legend and correspond to those used in the Epigenomic Roadmap website; DE-TEs whose genomic location encompasses multiple chromatin states are colored in blue.

TE transcriptional profiles recovered with Monocle appeared to cluster according to the individual RNAs' collection time and scattered along the temporal trajectory of *pheno*-conversion to LOAD, in a pseudotime-dependent manner (Figure 5c.4, A). Overall, the transition from being in a ConverterPre into a ConverterPost state requires about 45 "pseudotime" discrete units, representing a striking approximation of the observed 12 to 48 months that these subjects required for their clinical *pheno*-conversion. Such TE transcriptional changes lead to clustering of ConverterPre RNAs at the very beginning of the pseudo-temporal trajectory while the ConverterPost RNAs are distributed at the opposite extreme of the pseudo-temporal path to *pheno*-conversion (Component 1 in Figure 5c.4, A and B). This temporal distribution reflects the patterns of transcriptional activation of TE classes at the ConverterPre and ConverterPost stages of LOAD development. The pseudotemporal reconstruction shows that TEs change consistently with the development of LOAD across individuals, mimicking almost entirely the expected timing of the transition from ConverterPre to ConverterPost stages within these particular individuals. Notably, there is a remarkable differentiation of the pseudotemporal starting points across the ConverterPre stage(s), with individuals clustering at different positions along Component 2 of Figure 5c.4 A and B, suggesting a degree of heterogeneity of TE-identified ConverterPre conditions across individuals (Figure 5c.4, B).

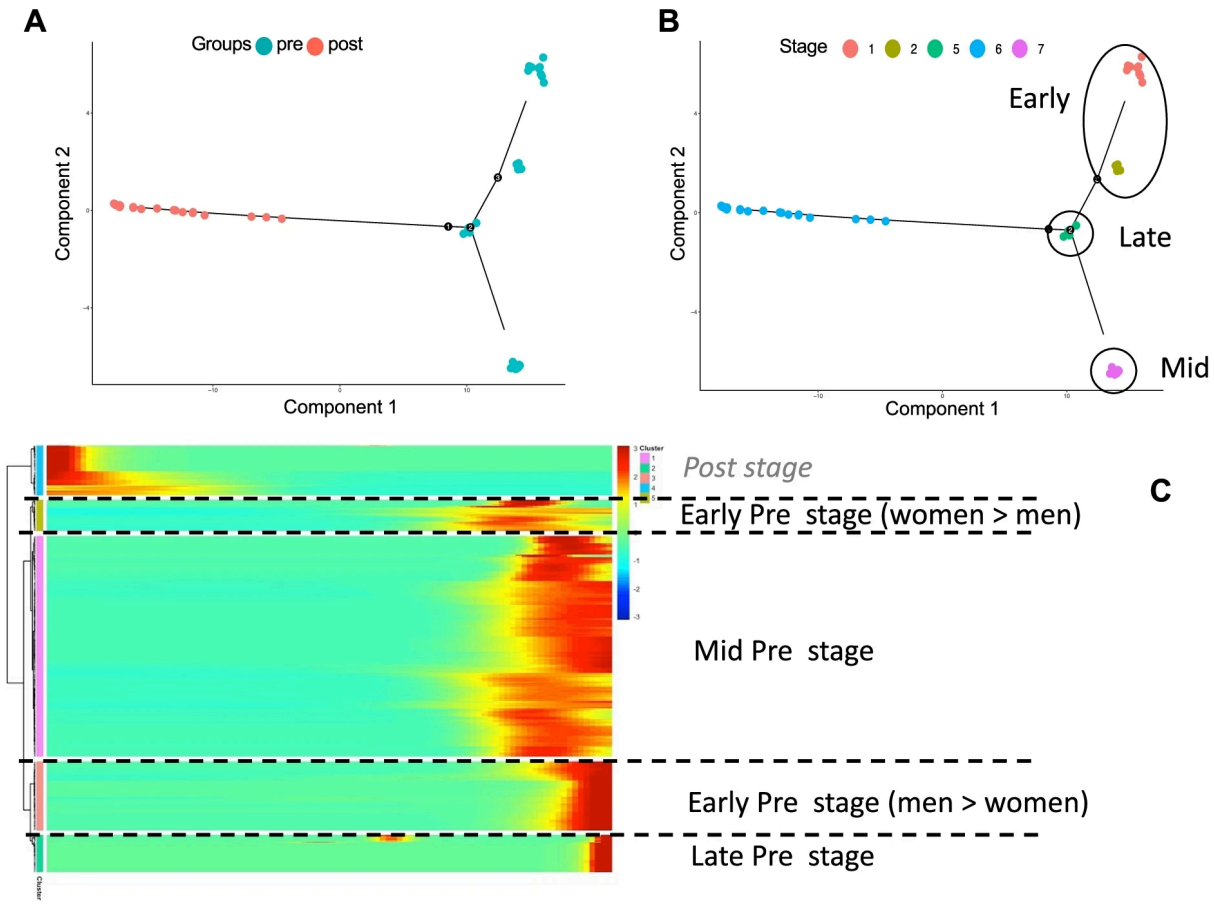



**Figure 5c.4. A and B** show the pseudotime continuum from a ConverterPre (dots on the right side) to a ConverterPost (dots on the left side) for the subjects that developed AD during the period of observation. Dots represent subjects: in A, blue dots are subjects at their ConverterPre condition and red dots are those at their ConverterPost condition. In B, blue dots show the ConverterPost condition for subjects; the other colors show different ConverterPre stages. **C** shows a heatmap expression matrix for significant DE-TEs at the 3 (early, mid, and late) ConverterPre stages in addition to the ConverterPost phase.

Once we ordered the ConverterPre/ConverterPost individuals in stages of disease development, we sought to identify which TEs dynamically change as a function of disease stage when the individual RNAs progress through disease development. For each TE, we modeled its expression by fitting two models (full and reduced) that differ based on whether the individual RNA classifications are explicit or not (Qiu et al., 2017) (Figure 5c.4, C). A total of 2,408 TEs appeared regulated during the transition to a clinically evident (manifest) phase of LOAD. We further explored whether specific families of TEs are more highly expressed in an individual's ConverterPre RNA type compared to another, and

whether specific classes of TEs tend to be co-expressed along the pseudo-temporal trajectory. At the threshold of false discovery rate (FDR) < 0.01, the Monocle cluster analysis identifies 5 groups of TEs that display patterns of similar expression within each cluster, but no TE classes or families appeared enriched across these 5 clusters. The partition of ConverterPre individual RNAs into separate subgroups seems rather consistent with differing preliminary stages of progression of TE activation to manifest stages of LOAD (ConverterPost). The analysis of the pseudotemporal trajectory indicates that the TE transcriptional activity delineates three different branches within global ConverterPre TE transcripts. This finding appears dependent on the particular time an individual is analyzed, prior to the onset of disease at his/her own ConverterPre stage, and shows consistency in partitioning the ConverterPre stage into early, mid, and late phases, with each phase signaling the time to AD onset (early: 36–48 months; mid: 18–36 months; late: < 18 months) (Figure 5c.4, B). A total of 1,006 TEs characterize these 3 phases of the ConverterPre state along a "dynamic" pseudotime trajectory, with 106 TEs overlapping with those found significant as DE-TEs in the ConverterPre to ConverterPost comparison (Figure 5c.4, C). The early phase is also characterized by a sex effect, further generating two subgroups with different women:men ratios. Such finding suggests the presence of heterogeneity at the "pre" stage, due to both sex and time-to-disease-onset effects. Each individual at his/her ConverterPre state is thus characterized by a specific TE signature that marks his/her progression toward the ConverterPost state, which does not show signs of heterogeneity related to TEs' expression (Figure 5c.4, B).

Using the described ML approach, we obtained eight predictive biomarkers by comparing the ConverterPre vs ConverterPost states in samples of subjects that *pheno*-converted to manifest LOAD, producing a classification accuracy of 78% (Table 5c.1, A). In particular, a L1M5 element is located within the intron of the USP25 gene on chromosome 21, whose trisomy is associated with Down syndrome (DS; trisomy-21), a condition associated with a high AD risk. USP25 is implicated in activating microglia, and its over-expression allows the de-ubiquitination of a series of molecular substrates that have been associated with synaptic abnormalities and related cognitive deficits. Removal of USP25 reduces neuroinflammation and rescues synaptic and cognitive functions in a

knockout mouse model (Sanchez-Valle et al., 2017; Zheng et al., 2021; Soleimani Zakeri et al., 2020; de Yebra et al., 2004; Castillo et al., 2017). When analyzing the comparison of ConverterPre vs NC individuals, we also found few significant TEs, most of which are not localized in protein-coding genes and do not have an already known specific relationship with AD. Furthermore, these TEs as biomarkers have an accuracy that is lower (69%) compared with that of the ConverterPre vs ConverterPost condition (Table 5c.1, B).

Finally, PCAs and related ROC curves confirmed that both RNA-sequencing DE analyses and the ML approach identified TEs that correctly discriminate between the patient groups, even when a small number of predictive biomarkers, those selected with the machine learning (ML) algorithm, are used in the models as reported in Supplementary Figure S3.

| Chr | Start | End | TE | Gene |
|---|---|---|---|---|
| 1 | 108,926,372 | 108,927,695 | L1M3 | GPSM2 |
| 1 | 174,901,607 | 174,902,942 | L1PA10 | RABGAP1L/KIAA0471 |
| 7 | 149,486,060 | 149,486,843 | L1ME3D | ZNF746 |
| 9 | 124,885,506 | 124,887,215 | L2a | GOLGA1 |
| 11 | 88,331,047 | 88,331,958 | MER21A | CTSC |
| 17 | 45,631,040 | 45,631,664 | MER77B | LINC02210 |
| 21 | 15,738,037 | 15,738,674 | L1M5 | USP25 |
| X | 17,096,688 | 17,097,359 | L1MEd | REPS2 |

| Chr | Start | End | TE | Gene |
|---|---|---|---|---|
| 22 | 23,900,208 | 23,900,715 | MER9a2 | NA |
| 16 | 67,141,398 | 67,142,927 | MER52A | C16orf70 |
| 2 | 26,305,911 | 26,306,395 | LTR15 | AC10896.1 |
| 19 | 11,853,714 | 11,854,477 | HERVK3-int | ZNF439 |
| 17 | 67,398,160 | 67,399,008 | HSMAR1 | PITPNC1 |
| 2 | 97,505,313 | 97,505,837 | MER1A | ANKRD36B |
| 20 | 44,217,986 | 44,218,725 | L1ME4b | OSER1-DT |
| 19 | 54,668,341 | 54,669,123 | L1M5 | LILRB4 |

**Table 5c.1, A (left) and B (right).** TEs selected by machine learning analysis. The 8 TEs are able to discriminate between ConverterPre/ConverterPost and ConverterPre/NC conditions with an AUC accuracy of 78% and 69%, respectively. Chr = chromosome; Start = TE start position; End = TE end position; TE class = TE class type; Gene = gene in which a TE is located.

## Discussion

A few published reports have suggested that TEs show a differential expression in patients with Alzheimer's disease (AD) compared to healthy aged controls, using case–control, retrospective approaches. Here, we have shown that the expression of TEs is massively dysregulated before the clinical manifestations of LOAD using RNA-sequencing data at both the preclinical (ConverterPre) and clinically manifest (ConverterPost) stages of disease using data from the same subjects. To our knowledge, our analysis is the first of its kind, using data collected from a prospective longitudinal

cohort of subjects known to have started in NC state and *pheno*-converted to LOAD over a 12–48-month time-frame. Our prospective design supports the hypothesis that the functional expression of our genome is altered through DE-TEs in subjects that are in a preclinical stage of LOAD, when they are otherwise clinically and cognitively undistinguishable by other NC subjects that will not go on to develop the disease. Our findings suggest that DE-TEs may be used as peripheral biomarkers heralding the future development of LOAD within a specific time-frame, although the exact span of such time-frame needs to be more carefully investigated. Our current and experimentally tested time-frame ranges between 12 and 48 months before the clinical onset of the disease, but, at least in principle, subjects that will develop AD at some point in time during their life could present with a TE's genomic dysregulation even 10 or 20 years (or more) before the clinical onset of AD. Moreover, many of the DE-TEs that we detected in blood leukocytes appear to be functionally expressed enhancers or alternative promoters also in specific brain regions related to AD, using an in silico computational analysis of the Epigenome Roadmap database. These DE-TEs that appear also putatively expressed in brain regions are implicated in either memory and/or other cognitive functions (notably, within the hippocampus, the anterior caudate, and the inferior temporal lobe, among others). Therefore, our findings might also direct future analyses investigating novel genomic elements that may regulate regional brain genomic mechanisms involved in developing AD. Few previous studies have investigated the possibility to use blood as a surrogate of brain in transcriptomic investigations (Cai et al., 2010), while more papers evaluated the blood–brain correlation for methylation analyses (Edgar et al., 2017), but to the best of our knowledge, at present there are no studies systematically comparing TE regulation and expression between human brain and blood. It is worth noting, however, that 61% of the DE-TEs in the ConverterPre vs ConverterPost comparison and 67% of the DE-TEs in the ConverterPre vs NonConverter comparison were also expressed in the human dorsolateral prefrontal cortex (DLPFC) according to previous data generated by our lab (Guffanti et al., 2018).

Expressed TEs are a large group of genomic elements, collectively classified as ncRNAs. While progressively better identified and known by their genomic locations (Hoyt et al., 2022), our current

knowledge regarding their functional role(s) remains incomplete. TEs have been considered enhancers or alternative promoters often associated with time- and tissue-dependent regulation of gene expression, as regulators of splicing sites, or contributing to domain rearrangement with pre-existing functional elements, producing novel composite architectures via exon shuffling, thereby leading to the genesis of genes with novel functionalities (Cosby et al., 2021). Additionally, especially in pathological conditions, commonly silenced TEs can be re-expressed due to loss or malfunction of TE-silencing mechanisms. When inappropriately (re-)expressed, TEs can lead to cellular death via multiple mechanisms, but usually involving the direct or indirect activation of the immune system. At present, we do not know whether the DE-TEs that we have observed in the development of AD are the primary mechanism driving neurodegeneration (etiological agents) or are acting as a secondary mechanism (pathogenic elements) unleashed by loss of TE-silencing mechanisms. We have identified, quantified, and evaluated a large number of DE-TEs that are nonetheless altering the functional architecture of the genome, under the assumption that expressed TEs act as non-coding RNAs regulating gene expression. Remarkably, other than their better-known role in cancer evolution, TEs have been proposed as pathogenetic elements in various neurological and psychiatric disorders (Guffanti et al., 2014; Reilly et al., 2013), despite our still limited understanding of their specific pathobiological mechanisms within the brain. We detected a significant overexpression of LINE1 elements prior to the onset of clinical manifestations of aMCI or LOAD, a sort of LINE1 storm, adding further support to the potential role of TEs in the genesis of certain neurodegenerative disorders. LINE1 re-expression has already been documented in senescent cells (De Cecco et al., 2013; Sedivy et al., 2013) and in the inflammatory and oxidative stress associated with cellular aging, dubbed senescent-associated secretory phenotype (SASP). SASP features an active expression of LINE1 elements and promotes neurodegeneration through the clearance of aging neural and glial cells by the immune system, activated by an unspecified chemical neuroinflammation (De Cecco et al., 2019; LaRocca et al., 2020). As our group and others have noted in SASP, most over-expressed LINE1s are evolutionarily recent, with many elements appearing to be human-specific, a finding that has yet to be confirmed by others (LaRocca et al., 2020). Most of our overexpressed LINE1 transcripts overlap with signatures of transcription regulation, as reported in the Epigenome Roadmap

data: genic enhancers or Transcription Start Sites (TSS). These signatures of transcription are present in normal blood cells and in both adult and embryonic brain tissues of varying developmental stages. Noteworthy, 85% of the genes containing LINE1 elements in their ORFs are brain-expressed, according to the Brain Atlas database (Sjostedt et al., 2020). We can question whether these overexpressed TE elements could also potentially dysregulate brain genes: 20 of these genes are actually already known to be associated with a "dementia" phenotype and 7 specifically with AD. Whether the TEs that we identified as DE in peripheral blood are also DE and have an effect in the human brain remains an open question. However, under the only functional assumption that we have considered here (that LINE1s can act as *cis* regulators of gene expression), we acknowledge that the genes putatively regulated by these DE LINE1s are also associated with "circadian gene expression" or "interferon-mediated immune response to pathogen-associated DNAs" pathways. Not surprisingly, therefore, one of these genes produces the amyloid precursor protein (APP), and is potentially regulated by a full-length (6,025 nucleotides) human-specific L1PA2 element, presenting as an enhancer with a weak-transcription signature.

Another LINE1 selected by the machine learning predictive algorithm among the 8 TEs that classify ConverterPre individuals with 78% accuracy is a L1M5. This L1M5 presents with a signature of a weak enhancer and is located within the first intron of the USP25 gene. The same gene is also tagged by a second LINE1 (a L1M1) in the 4th intron, again showing a signature of a weak enhancer. The USP25 gene has already been shown to be greatly expressed in the brains of DS patients than in controls (Lockstone et al., 2007) and overexpression of USP25 in a murine model of DS-AD, particularly in hippocampal CA1 cells, results in microglial activation inducing both synaptic and cognitive deficits (Zheng et al., 2021).

In addition to LINE1s, we found other DE-TEs in our sample set. HERVs are known to be highly expressed in human embryonic stem cells (hESCs), with HERV-H and -K considered markers for pluripotency (Santoni et al., 2012; Grow et al., 2015; Wang et al., 2014; Lu et al., 2014). Progressively silenced during cell differentiation, HERVs still represent one of the largest sources of regulatory

elements (mostly enhancers) under both physiological and pathological conditions, and show tissue-dependent specificity (Deniz et al., 2020; Suntsova et al., 2013 and 2015; Reilly et al., 2015; Stearrett et al., 2021). In addition to other neuropsychiatric diseases, HERVs have also been proposed to play a role in neurodegeneration, possibly altering the functional architecture of the genome and contributing to cell death. Within the 1,790 DE TEs identified in the ConverterPre vs ConverterPost comparison, HERVs/LTRs represent about 25% of the TEs, with not less than 10% being human-specific, and mostly represented by HERV-K elements. Whether HERV-K elements contribute to characterizing certain pathways noted to be enriched in the ConverterPre vs ConverterPost comparison, in addition to the genes putatively controlled by them as regulators, remains uncertain, due to the current imprecise knowledge base for biological effects associated with HERV sequences. About 50% of the DE SVAs that we detected in the ConverterPre vs ConverterPost comparison and 35% of those in the ConverterPre vs NC comparisons belong to the E and F sub-clades, indicative of the more evolutionary recent SVA elements in our genome (Hoyth et al., 2022). Moreover, they continue to appear to be transpositionally active, or at least can co-mobilize 3′ or 5′ DNA flanking regions to new genomic loci using TE-mediated transduction (Hoyth et al., 2022). They represent, therefore, one of the most active mechanisms to generate structural variation. Alus, which are significantly depleted in our ConverterPre vs ConverterPost comparisons, seem to act by the same mechanisms observed in SVAs. Thus, these data support the idea that TEs expression, including those associated with SVAs and Alus, are important for risk profiling in preclinical LOAD.

We have shown that TEs can be profiled in a pseudotime model of LOAD development, further suggesting their involvement in a disease fate decision along a pathological continuum. To obtain further insights as to which family of TEs is more highly expressed in the ConverterPre vs ConverterPost groups, first we examined whether specific classes of TEs are typically co-expressed along the development of the disease. Using a cluster analysis, we found TE expression profiles along pseudotime trajectory clusters according to different stages of the LOAD developmental process. At the threshold of FDR < $1 * E^{-3}$, the cluster analysis identifies 5 groups of TEs that display patterns of similar expression within each cluster. The unique expression within the ConverterPost group is

clearly different from the 4 associated sub-stages within ConverterPre. It remains somewhat puzzling as to the significance of the 4 different clusters of subjects defined within the ConverterPre stage of disease, although they likely represent clinical heterogeneity. While failing to meet significance due to the limited sample size of our dataset, we also noted that these ConverterPre clusters are characterized by a different time-to-disease and a different sex ratio. Such discrimination allows us to define these clusters into two early, a mid, and a late ConverterPre transition stage to clinical LOAD. The two early ConverterPre clusters are best defined via the women:men ratio, and define subjects at farthest timepoints away from phenoconversion to LOAD. Importantly, the four ConverterPre clusters display significantly dysregulated TEs (retrotransposon storm) compared to clusters noted in the NC and ConverterPost groups.

In summary, TEs appear to be involved in a profound reorganization of the functional architecture of the genome in LOAD (and probably other age-dependent diseases). Based on our analyses, DE-TEs at specific ConverterPre timepoints appear to accurately identify those individuals that are at risk of *pheno*-converting to LOAD. Two different analytical methodologies were used to define such biomarkers, using either: 1) all the DE-TEs identified between ConverterPre and NC or 2) a machine learning algorithm that makes use of a much reduced number of DE-TEs, after a thorough control of entropy reduction. While it is not surprising that the whole set of DE-TEs (1,790 elements) can fully discriminate between the ConverterPre and ConverterPost stages of LOAD development, it is interesting to note that only 8 TEs are required to discriminate subjects between ConverterPre and ConverterPost, with about 80% accuracy. Whether the latter result might be suggestive of a more biologically relevant set of TEs within the preclinical stage of LOAD, or is a consequence of the entropy reduction algorithm, remains unresolved, but it will require further elucidation.

# 6. Final remarks

## 6a. Weaknesses of the studies

The first limitation to the above studies is the use of short-read sequences to detect and identify complex structural variants such as polymorphic retrotransposons. From one side, the use of short-read Next Generation Sequencing (NGS) technology allows a fast detection and genotyping of most TEs, as demonstrated by the power and accuracy of the MELT pipeline (Gardner et al., 2017; Watson et al., 2021). On the other hand, short reads do not allow the identification of TEs located in repetitive/complex regions of the genome, as well as to detect the variation within the TE's sequence itself. In the last years, long-read sequencing provided scientists with "highly accurate and more complete human genome sequences that span more of the repetitive regions compared to short-read Illumina sequencing" (Devine, 2023), such as that of the Telomere-to-Telomere project ("T2T": Nurk et al., 2022). In future studies, the use of long-read sequencing will enable researchers to study not only Transposable Elements (TEs) presence or absence (i.e., allelic frequencies and genotypes), but also variations in their sequence that could lead to a different function or expression of the mobile element itself. Furthermore, the availability of more complete human genome sequences (i.e., the "T2T" genome) will increase the possibilities to find and annotate TEs in particularly complex and repetitive regions of the human genome, such as centromeres and telomeres. Similarly, the advent of a human pangenome reference sequence (Liao et al., 2023) will result in a more precise identification and annotation of structural variants (Massarat et al., 2023; Devine, 2023). This will provide scientists with a more complete comprehension of the mobile element's evolution (in terms of sequence variation) and time of insertion (for instance, before/after the split between Neanderthals/Denisovans and Anatomically Modern Humans).

After looking for the presence/absence of mobile elements, we performed specific tests to highlight a possible association between polymorphic TEs and a defined phenotype/disease (such as living at high altitude, addiction/substance abuse and neurological disorders), but these tests have some

limitations. In particular, the analyses performed with GEMMA (Zhou and Stephens, 2012) and Beagle (Browning and Browning, 2007; Browning, Zhou and Browning, 2018) were carried out on a limited set of variants and individuals. However, we must take into consideration that GEMMA corrects for the potential presence of population stratification, a typical feature of small/isolated populations. Where possible, we also performed haplotype-based analyses with Beagle to strengthen our findings by also taking into account the genomic surrounding of the previously identified TEs (see chapters 4b and 5a). In these two particular studies, we tried to corroborate the significant results of frequency-based analyses and association tests by assessing the potential influence of the presence/absence of TEs in the context of their genomic landscape, allowing us to obtain a more accurate identification of candidate structural variants for future analyses. For instance, in chapter 4b, we reconstructed haplotypes including the polymorphic TEs identified with GEMMA, and confirmed two significant results: the intergenic Alus on chromosome 6 that are associated with the "alcohol" phenotype. Similarly, in chapter 5a, the haplotype-based analysis was performed on the 38 candidate TEs (for an increased risk of developing schizophrenia) which emerged from the allelic frequencies-based analysis, and again two haplotypes including mobile elements were identified as possibly associated with the schizophrenic condition.

Analogously, sample size problems could also affect the outcome of selection analyses. In chapter 5b, we describe our work on positive selection acting on polymorphic TEs with a particular focus on brain development and diseases. In this case, we used three different and complementary tests for positive selection that work well also with a low number of samples: nSL (Ferrer-Admetlla et al., 2014), DIND (Fagny et al., 2014) and H12 (Gardu et al., 2015). Moreover, to reinforce our study, we cross-checked our results with those of Rishishwar and colleagues (2018), who used a completely different approach to find traces of positive selection acting on polymorphic TEs: interestingly, 12 variants were identified by both studies, therefore confirming the validity of our approach.

Lastly, this thesis, being founded in a bioinformatic approach, presented only *in silico* studies, which are particularly useful for detecting and identifying potential candidates for *in vitro* and/or *in vivo*

analyses. Indeed, future *in vitro/in vivo* analyses on the identified TEs could elucidate their functional role, for example on nearby genes, and also their effects on the host's genome.

## 6b. General discussion and conclusions

Transposable elements mobilization in germline cells can enhance diversity through the creation of potentially inheritable insertions, thereby generating human-specific polymorphisms. Therefore, we decided to use TE polymorphisms to study both ancient and modern human variability. As shown by the studies carried out for this thesis, polymorphic TEs are exceptionally useful as variability and ancestry markers for human populations, as suggested also by other Authors (Rishishwar et al., 2015; Gardner et al., 2017; Watkins et al., 2020). For instance, Principal Component and Admixture analyses performed on different datasets (Figures 4a.2, 4b.2, 5a.1 and 5c.1C) show how polymorphic TEs are valuable genetic markers to reconstruct the genetic history of modern human populations, and also to highlight possible population substructures: see in particular Figures 5a.1 and 5c.1C, in which we showed how affected individuals (i.e., patients with schizophrenia or Alzheimer's Disease) and controls do not differ from each other and do not form any particular sub-clustering.

Genome-wide association studies (GWAS) usually rely on single nucleotide polymorphisms to unveil a possible association between a phenotypic trait and a genetic variant. Since we focused on the analysis of structural variants, we performed association studies using polymorphic TEs, applying two different methods: an association test after inferring haplotypes (see chapters 4b and 5a) and an association test using only information about the TE's genotype (chapters 4a and 4b). Both approaches returned several significant results, confirming that polymorphic TEs can be used not only to study human populations variability and adaptation to particular environments, but also that they can inform us about the susceptibility to specific disorders, such as schizophrenia and addiction.

Functional roles of retrotransposons are also important: as highlighted in the introduction section (chapters 1b-c), TEs can be co-opted by the host's genome and intervene in fine-tuning expression and alternative splicing of nearby genes (see for example Cao et al., 2020). Therefore, we retrieved information about functional roles of the most significant TEs emerged by our analyses: notably, several mobile elements act as expression and/or alternative splicing quantitative trait loci, and when

looking at TEs specifically studied in the context of an increased risk of developing schizophrenia or impacting cognitive abilities, a non-negligible number of those retrotransposons act as eQTL/sQTL specifically in brain tissues.

Recent works suggests that the reactivation of transcriptionally silent TEs can impact the neural homeostasis during pathological aging and might induce brain degeneration (Frost et al., 2014; Colombo et al., 2018; Liu et al., 2019; Jonsson et al., 2020; Ochoa Thomas et al., 2020). Even when they do not actively transpose, mobile elements can induce variation and diseases, for example when they escape silencing and their expression is therefore altered, as demonstrated for several cancers (reviewed in: Burns, 2017; Anwar et al., 2017; Grundy et al., 2022) and proposed for brain disorders (reviewed in: Misiak et al., 2019; Ahmadi et al., 2020. See also Guo et al., 2018 and Macciardi et al., 2022).

Our work on TEs expression in the early stages of late onset Alzheimer's disease (LOAD) not only highlighted how a very high number of TEs are differentially expressed between subjects who will not develop LOAD and subjects who will, but also that some of these differentially expressed TEs could be potentially used as a peripheral biomarker for those individuals that are at risk of converting to LOAD years before the manifest stages of the disease (Macciardi et al., 2022).

At this point of the discussion, we would like to introduce an important question: during human evolution, have transposable elements experienced negative/purifying selection or were they subject to positive selection? Are transposable elements negative or positive players in human genome evolution, and in particular in the evolution of cognitive abilities typical of *Homo sapiens*?

Far from being "junk" and devoid of functions, transposable elements have been proved to have both positive and detrimental effects on the host's genome (reviewed in Reilly et al., 2013; Bourque et al., 2018; Gebrie, 2023), especially when looking at cognitive functions and central nervous system development (Guffanti et al., 2016; Suarez et al., 2018; Ferrari et al., 2021). For instance, the analysis of the 1KGP dataset identified ~16,000 polymorphic TE loci, 93% of which show a worldwide allele frequency < 5%, indicating that overall polymorphic TEs are indeed deleterious and have faced

purifying selection (Rishishwar et al., 2015). On the other hand, the same Author suggests how several TEs are under positive selection and possibly contributed to modern human populations differentiation (Rishishwar et al., 2018). For example, the adaptation of Tibetan and Sherpa highlanders to the extreme environmental conditions of the Himalayan Arc could have been influenced by polymorphic TEs (as we reported in chapter 4a and in Modenini et al., 2024), which not only participated in the physiological adaptation to high altitude of these populations, but also have probably experienced recent instances of positive selection - as fixation index and population branch statistics suggest - that possibly contributed to the differentiation of these human groups.

Furthermore, our study on the signatures of positive selection in the human brain (chapter 5b) suggests that hundreds of polymorphic TEs possibly have experienced recent instances of positive selection, according to different tests, and have been subject to adaptive introgression, a process in which "beneficial variants acquired from archaic humans may have accelerated adaptation and improved survival in new environments" (Racimo et al., 2015). Just a few papers studied the evolutionary forces acting on TEs in the context of cognitive abilities and neurological diseases, such as schizophrenia and autism spectrum disorder (ASD). Among them, some Authors point towards a positive evolutionary force acting on schizophrenia and ASD risk variants (Srinivasan et al., 2016; Polimanti & Gelernter, 2017), while others suggest that the protective variants are those actually under positive selection through a mechanism called "non-antagonistic pleiotropy" (Yao et al., 2020; Gonzales-Peñas et al., 2023). Our current results do not clearly point toward one of the two hypotheses, even if some TEs, that are more present in schizophrenic subjects and are putatively under positive selection, seem to confirm the hypothesis proposed by Srinivasan (2016) and Polimanti & Gelernter (2017).

To sum up, our research firstly highlights how polymorphic transposable elements are extremely valuable genetic markers for studies on human ancestry and evolution, as shown by results from classical population genomics analyses (i.e., PCA and Admixture). Secondly, we verified if polymorphic TEs can be used to perform association studies, in order to look for structural variants associated with particular phenotypes or diseases. Third, by cross-checking our results with publicly

available lists of TEs, we showed how some of the detected polymorphic elements can act as expression and/or alternative splicing modulators of nearby genes. Finally, we tested whether polymorphic TEs may have been subject to ancient adaptive introgression events or may have experienced recent instances of positive selection, as suggested by other Authors.

In conclusion, our study underlines the importance of using polymorphic transposable elements as genetic markers for the study of human evolution, especially when looking for particular effects on human phenotypes, such as body height/body mass index, and brain diseases, including addiction and neurological disorders.

# 7. Appendix: the rapid expansion of the APOBEC3 family of proteins is possibly correlated with the evolution of retrotransposons in primates

Giorgia Modenini[1], Paolo Abondio[1,2], Alessio Boattini[1,#]. *The coevolution between APOBEC3 and retrotransposons in primates*. Mob DNA. 2022 Nov 29;13(1):27. doi: 10.1186/s13100-022-00283-1

Affiliations:
1: Bigea Department, University of Bologna, Bologna, Italy
2: Department of Cultural Heritage, University of Bologna, Ravenna, Italy
#: correspondence to alessio.boattini2@unibo.it

**Background**

In the introduction, we explained how retrotransposons activity in humans is counteracted by numerous defense mechanisms. Here, we will introduce our hypothesis about the co-evolution between apolipoprotein B mRNA-editing catalytic polypeptide-like 3 (APOBEC3 or A3) and retrotransposons in primates (Modenini et al., 2022).

Located on human chromosome 22, the APOBEC3 genes encode for deaminase proteins that can catalyze the deamination of cytosine-to-uracil (C to U) on single-stranded DNA and/or RNA. APOBEC3 genes are part of the AID/APOBEC superfamily of proteins involved in immunity, metabolism and infectious diseases (reviewed in Conticello, 2008) and are present only in placental mammals (Conticello et al., 2008; Rogozin et al., 2007). In most primates and *Homo*, the APOBEC3 family of mutator proteins is represented by seven members: APOBEC3A/B/C/D/F/G/H, first annotated by Jarmuz et al. (2002).

Initially studied for their capacity of inhibiting a wide range of exogenous viruses, such as Human/Simian immunodeficiency virus (HIV/SIV) (Sheehy et al., 2002; Wang et al., 2018) and hepatitis B virus (HBV) (Suspène et al., 2011), more recently APOBEC3 genes have been recognized for their important role in counteracting the mobilization of endogenous retroviruses and other retrotransposons, such as Alus and LINE-1s. Interestingly, in primates and *Homo* A3 proteins have been faced with strong positive selection, duplications and fusions that gave rise to the currently known seven members of the APOBEC3 gene cluster. Such expansion is a consequence of the co-evolution between A3 proteins and their counterparts, i.e. viruses and retrotransposons (Konkel et al., 2010; Ito et al., 2020).

**Evolution of retrotransposons in primates**

In primates, half of the genome consists of TEs, and about 60% of the repetitive elements are represented by LINEs and SINEs in all investigated species of primates, suggesting their evolutionary importance across simians and prosimians (Lee et al., 2015).

The emergence of the primate-specific Alu family of retrotransposons dates back to the radiation between simians and prosimians, ~100 Mya (million years ago), with a major expansion between 50 and 25 Mya (Shen et al., 1991). The oldest subfamily AluJ predates the division between Strepsirrhini and Haplorrhini (~ 86 Mya) (Figure 7.1); the subfamily AluS derived from AluJ just before the divergence between Platyrrhini and Catarrhini, and took over amplification about 55 Mya. Finally, the youngest subfamily of Alu, AluY, evolved from AluS and expanded in the Catarrhini lineage, with AluYa5 and AluYb8 being the most represented elements in humans.

The evolutionary history of LINE-1s is far less characterized. Early in primates' evolution as many as three lineages were present and active in parallel for 30 My (Khan et al., 2006): L1MA, L1PB and L1PA. L1PA succeeded and remained active within the anthropoid lineage leading to the human specific L1PA1 (Smit et al., 1995). Nowadays, the most active L1 subfamily in the human genome is L1-Ta1 (Boissinot et al., 2004), however some pre-Ta elements are still capable of retrotransposition (Kazazian et al., 1988; Beck et al., 2010).

SVA family, which is represented by seven members (named SVA_A-F), is more recent than Alu and L1: indeed, the lack of SVAs in old world monkeys suggests that SVAs are hominid specific retroelements (Wang et al., 2005). Subfamily age estimates based upon nucleotide divergence indicate that the expansion of four SVA subfamilies (SVA_A-D) began before the divergence of human, chimpanzee, and gorilla, while subfamilies SVA_E and SVA_F are restricted to the human lineage (Wang et al., 2005). SVAs took over expansion in Great Apes, with 1,800-2,500 elements identified in orangutan (Locke et al., 2011) and chimpanzee (Chimpanzee sequence and analysis consortium, 2005) genomes, respectively, and approximately 2,700 elements found in humans. SVAs are composite elements, as we described in the introduction, but they are not the only composite TEs in primates: for instance, LAVA (L1-Alu-VNTR-Alu), PVA (PTGR2-VNTR-Alu) and FVA (FRAM-VNTR-Alu) elements have been identified in gibbons (Carbone et al., 2012; Ianc et al., 2014). They combine

simple repeats, Alu fragments, a VNTR and variable 3' domains, which are, except for PVA, derived from other retrotransposons (Damert 2015). Notably, the central domain of VNTR composites evolved in a lineage-specific manner which gave rise to distinct structures in gibbon LAVA, orangutan SVA, and human/chimpanzee SVA (Lupan et al., 2015), suggesting an inextricable link between TEs and primate genomes that lead to speciation, radiation and evolution of primates.

Finally, the most ancient groups of human endogenous retroviruses (HERVs) appeared before the divergence of Platyrrhini and Catarrhini, around 40 Mya (HERV-L and H). Younger HERV groups, such as HERV-E and HERV-K(HML-2), have been acquired after the separation of Platyrrhini and Catarrhini (Grandi et al., 2020).

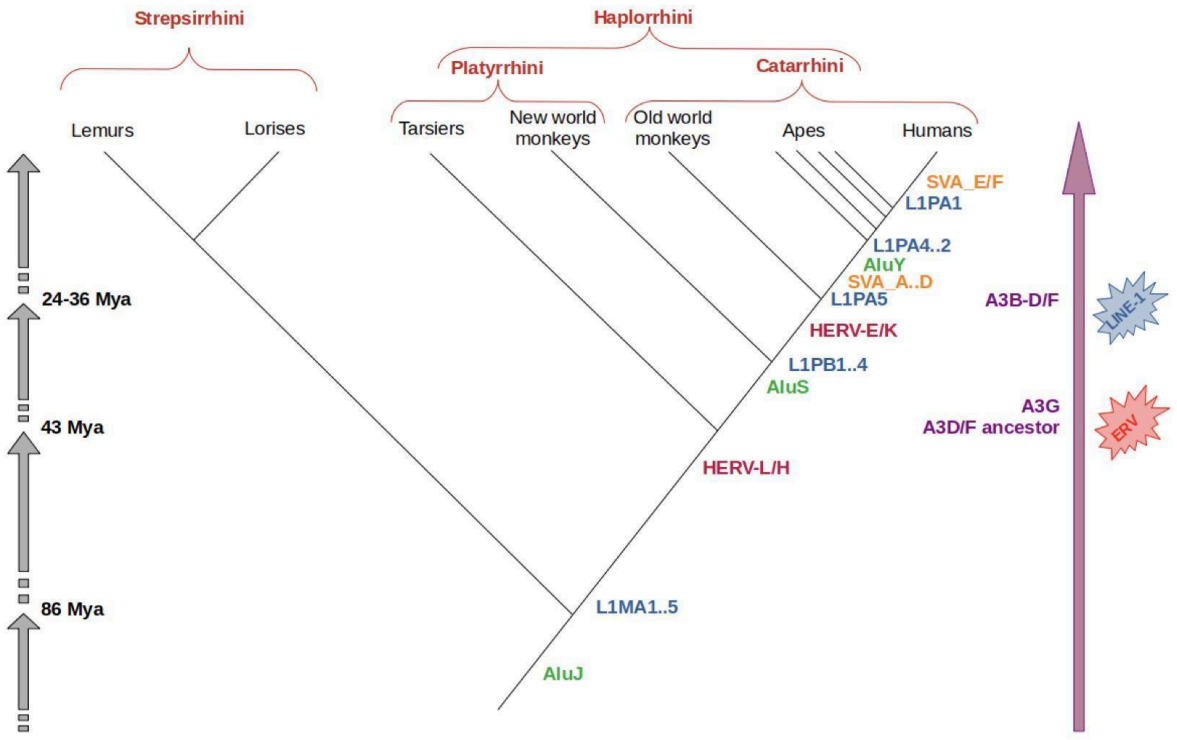



**Figure 7.1**. Evolutionary tree of primates and retrotransposons. Alu, L1 and HERV are more ancient than SVAs, which are Hominoidea-specific (Apes+Humans). The origin of different APOBEC3 genes is concurrent with the explosion of specific retrotransposon families, i.e. HERVs and L1s: A3G appeared just after the split of Simiiformes 43 Mya and during the invasion of ERVs, while A3B and A3D/F originated during the invasion of LINE-1 and the split between old world monkeys and Hominoidea.

**Evolution of the APOBEC3 family of proteins in primates**

APOBEC3A/C/H have a single cytosine deaminase (CD) domain. By contrast, APOBEC3B/D/F/G have two CD domains, of which only the C-terminal CD2 is catalytically active (61). All A3 proteins share at least one zinc (Z)-coordinating catalytic motif, and A3 genes possess either one or two conserved zinc-coordinating motifs, in which the zinc is coordinated by a histidine and two cysteines. Z motifs can be classified into three groups (Z1, Z2, Z3), all sharing the consensus amino acid signature His-X-Glu-X23–28-Pro-Cys-X2–4-Cys (where X can be nearly any residue) (Conticello, 2008; Jarmuz et al., 2002; Münk et al., 2008; Münk et al., 2012).

The existence of three paralog zinc-coordinating motifs in the sequence of the seven APOBEC3 members in the primate lineage suggests a complex sequence of duplications and fusions that gave origin to the current ensemble of mutator proteins (Ito et al., 2020; Münk et al., 2012; Uriu et al., 2021). Specifically, primates carry three Z1 paralogs, seven Z2 paralogs, and one Z3 paralog distributed across the APOBEC3 gene locus on chromosome 22 (Ratcliff & Simmonds, 2021). In modern humans, these eleven A3 open reading frames contribute to the seven genes by encoding either a single Z domain or a fusion of two (A3Z2-A3Z2 or A3Z2-A3Z1) in a complex organization (Münk et al., 2012). These three motifs certainly existed at least as far back as the separation between placental mammals and marsupials, 148 Mya, and may have originated from a single gene copy, possibly predating egg-laying mammals (247 Mya) (Münk et al., 2012). Moreover, Münk and colleagues show that most duplications and rearrangements in the Z1 and Z2 groups, especially for the primate lineage, have happened over the last 100 My. When compared with their sister group, the AICDA genes, the Z groups all show a higher evolutionary rate (AICDA: 7.41x10-4 substitutions per site per My; A3s: 2x10-3 substitutions per site per My), but there is a significant decrease in the evolutionary rate of the Z groups over the last 100 My ($p$-value $< 0.0007$). Therefore, the A3 genes have a higher rate of substitutions than their sister groups, but the same rate has steadily reduced over time. The Z1 group has split twice: once around the basal divergence of primates (around 75 Mya), and again around the origin of the Hominoidea lineage (26 to 34 Mya) (Münk et al., 2012). The phylogenetic relationships of the Z2 group are more complex to reveal, especially with regards to the

primate lineage, but Münk and colleagues argue that a first duplication event (or even two) may have happened around the separation between Haplorrhini and Strepsirrhini (86 Mya) and certainly before the diversification of the Simiiformes (43 Mya); based on sequence similarity, the several copies of Z2 that can be found in humans have definitely appeared by duplication, but their phylogeny is intricate and separation estimates could not be clearly supported (Münk et al., 2012).

Recently, Uriu and colleagues have performed a complete reannotation of the APOBEC gene family in primates, specifically highlighting the phylogenetic subclassification of the A3 zinc domains (Uriu et al., 2021). Their work confirmed the amplification of the Z1 and Z2 domains in this lineage, together with an accelerated increase in diversification and complexity over time, especially with respect to Z3. By comparing sequences of Prosimians, New World Monkeys (NWM), Old World Monkeys (OWM) and Hominoidea, they suggest that the Z3 domain was preserved in the Simiiformes but lost in the Prosimians, while the generation of genes with multiple catalytic domains that have been conserved up to the present has first occurred in the common ancestor of Simiiformes (Uriu et al., 2021). Repeated instances of amplification, duplication and gene conversion have, then, produced the variety of A3 genes that can be observed across Simiiformes today. Interestingly, these events have been accompanied by the peak invasions of mobile elements in human DNA: specifically, ERVs peaked around the origin of A3G in the common ancestor of the Simiiformes, while LINE1 peaked around the origin of A3B, D and F in the Catarrhine clade (OWMs and Hominoidea) (Uriu et al., 2021). Ito and colleagues (2020) explored the relationship between intact A3Z domains and ERV insertions in the mammalian genome and highlighted an acceleration in the accumulation of Z domains over an increase of ERV insertions in primates (Ito et al., 2020). At the same time, they suggested a parallel increase in the quantity of G-to-A mutations in primate ERV sequences and a higher estimated proportion of ERV insertions in the ancestor of Simiiformes, which was not subsequently carried on in the NWMs (Ito et al., 2020). Moreover, sequence analysis allows detection of residue conservation in the catalytic domains across all Z groups, as well as specific amino acid residues that are characteristic of each group (Uriu et al., 2021). These observations suggest a notable relationship between primate evolutionary radiation, proportion of transposable element insertions

over time and amplification of the defensive repertoire that brought to the variety of A3 genes observable in our species.

**Overview of APOBEC3 functions**

A3 genes are involved in various functions, from viral and retrotransposon restriction to cancer progression (Knisbacher et al., 2016). Indeed, several recent studies have described the role and mechanisms of action for this protein family in the context of cancer-related DNA mutagenesis, as it is becoming more and more clear that prevalent signatures of instability in cancer cell genomes are due to APOBEC3 activity on transiently exposed single-strand DNA (for example, during DNA mismatch repair and lagging strand replication) (Green et al., 2016; Lei et al., 2018; Mas-Ponte & Supek, 2020; Bergstrom et al., 2022; Petljak et al., 2022; Jakobsdottir et al., 2022). This activity leaves signatures along the double helix that are clearly traceable to A3 family members and are found predominantly in cancer cells (DeWeerd et al., 2022). As the structural details of A3s interaction with nucleic acids are being unveiled (Maiti et al., 2021), the ambivalent effect of these protective enzymes is also being highlighted, as an elevated expression of APOBEC3s may provide a reason for aberrant cancer-inducing somatic mutations in human papilloma virus (HPV) (Smith & Fenton, 2019; Revathidevi et al., 2021; Warren et al., 2022) and HBV (Zhang Y. et al., 2021) infections, as well as an extensive range of other tumor types (DeWeerd et al., 2022; Swanton et al., 2015; Guo et al., 2022), even in the context of inflammation (Liu et al., 2021).

In fact, A3s strongly inhibit various types of exogenous viruses, including herpesvirus, parvovirus, papillomavirus and hepadnavirus (Suspène et al., 2011; Baumert et al., 2007; Vartanian et al., 2008; Narvaiza et al., 2009). Sheehy et al. (2002) isolated a gene that restricts HIV-1 replication, identified as APOBEC3G (Sheehy et al., 2002). In HIV-1 and other viruses, the virion infectivity factor (Vif) is a potent regulator of virus infection and replication and is consequently essential for pathogenic infections in vivo (Fisher et al., 1987; Strebel et al., 1987; Gabuzda et al., 1992; von Schwedler et al., 1993; Desrosiers et al., 1998). Vif interacts with A3G, triggers the ubiquitination and degradation of A3G via the proteasomal pathway, by binding A3G and a Cullin5-ElonginBC E3 ubiquitin ligase complex which results in the proteasomal degradation of A3G. Therefore, Vif is required during viral

replication to inactivate the host cell antiviral factor A3G (Donahue et al., 2008). Indeed, the presence of a mutant Vif results in a failure to bind A3G, which in turn results in A3G incorporation into assembling virions with loss of viral infectivity (Donahue et al., 2008).

A3 proteins also inhibit the mobilization of endogenous retroviruses and other retroelements, such as Alu and L1. For instance, Esnault and colleagues (2005) demonstrated that A3G can interfere with the mobilization of murine ERV elements, such as IAP and MusD, by inducing G-to-A hypermutations in the proviral DNA plus strand (Esnault et al., 2005). In recent years, most A3 family members have been shown to be able to counteract the activity of Alus and L1s in humans and primates, both in the nucleus and in the cytoplasm. For instance, A3G is able to repress Alu retrotransposition without interacting directly with L1 (Chiu et al., 2006; Hulme et al., 2007), in fact A3G can inhibit L1-dependent retrotransposition by sequestering Alu RNAs in the cytoplasm, therefore being away from the nuclear L1's enzymatic machinery. Different A3 proteins have diverse cellular localization patterns: A3A/C/H act both in the cytoplasm and in the nucleus; A3B only in the nucleus; A3D/F/G are active in the cytoplasm (Aria et al., 2012). Given these critical functions, it is no surprise that the A3 family is being studied in the context of cancer, antiviral and immune-related drug discovery (Bennett et al., 2018; Green & Weitzman, 2019; Duan et al., 2020; Grillo et al., 2022).

**The evolutionary arms race between APOBEC3 and retrotransposons**

The evolutionary arms race (Dawkins & Krebs, 1979) is an ongoing struggle between competing sets of co-evolving genes, phenotypic/behavioral traits or species, that develop escalating adaptations and counter-adaptations against each other.

Retrotransposons in humans are counteracted by different mechanisms, for example the Piwi-interacting RNA (piRNA) pathway and the Krüppel-associated box zinc finger (KRAB-ZNF) proteins (reviewed in Goodier, 2016), which are able to repress TEs mobilization and expression. In a similar way, some components of the APOBEC3 gene cluster are involved in the control of retrotransposons. Indeed, the rapid co-evolution between the A3 locus and different retroviruses, and

the positive selection acting on A3 genes are signals of the continuous arms race that characterized A3s, viruses and retroelements (Sadeghpour et al., 2021; Zhang & Webb, 2004; Sawyer et al., 2004).

First discovered by Sheehy and colleagues as a defense against HIV-1 virus (Sheehy et al., 2002), A3G is able to repress ERVs mobilization in both mouse and human cells, by inducing G-to-A hypermutations in the nascent DNA of ERV elements, such as IAP and MusD in mice and HERV in humans (Esnault et al., 2005). Therefore, by editing viral genetic material, it provides an ancestral wide cellular defense against endogenous and exogenous invaders.

Other proteins of the A3 family can counteract LTR retrotransposons' activity: A3A and A3B. A3B acts similarly to A3G, by specifically interacting with the ERV Gag protein in co-expressing cells and inducing extensive editing of ERV reverse transcripts (Bogerd et al., 2006). On the contrary, A3A, which can restrict ERVs in human cells by 100-fold (compared to a 4-fold inhibition of these elements by A3G), fails to package detectably into ERV virus-like particles and does not edit ERV reverse transcripts (Bogerd et al., 2006).

Inhibition of L1 by A3 occurs at the post-transcriptional level by a deamination-dependent or independent mechanism. The most active enzyme (with respect to L1) A3A has deaminase activity and converts C-to-U in the first strand of the L1 cDNA transcript. As a result of such modification, the deamination of transiently exposed DNA leads to the truncation/abortion of retrotransposition (Richardson et al., 2014). A different mechanism has been identified for A3C and A3D: acting by a deamination-independent mechanism, the enzyme blocks the L1 reverse transcription reaction by interacting with the L1 complex of ribonucleoprotein (RNP) and ORF1 in the cell cytoplasm (Horn et al., 2014; Liang et al., 2016).

Recently, Uriu and colleagues (2021) investigated the evolutionary forces that drove the generation of the youngest A3 members, i.e. A3B and A3D/F. Notably, the invasion of LINE-1 and Alu peaked around the age of the common ancestor of Catarrhini (29 to 43 Mya), concurrently with the generation of A3B and the duplication of A3D/F, suggesting that the origin of these A3 genes in the common ancestor of Catarrhini could be driven by the invasion of LINE-1 and Alu (Uriu et al., 2021). The

same Authors suggest that the origin of A3G dates back to the age of the common ancestor of Simiiformes (67-43 Mya), when there was an invasion of ERV elements. Indeed, A3B potently suppresses the growth of LINE-1 (Stenglein & Harris, 2006; Wissing et al., 2011; Marchetto et al., 2013), whereas A3F inhibits the replication of vif-deleted HIV-1 (Liddament et al., 2004), HERV-K (Lee & Bieniasz, 2007) and LINE-1 (Stenglein & Harris, 2006). Altogether, these findings suggest that retrotransposons invasion in the common ancestor of Catarrhini and Simiiformes was a driving force of the powerful co-evolution between TEs and A3 proteins (Uriu et al., 2021).

Interestingly, DNA editing of retrotransposons has been proposed to be a source of genome evolution, in fact DNA editing by APOBEC3 can induce many mutations in a single event. That way, a given element could change to such an extent that its evolutionary trajectory could be altered (Knisbacher et al., 2016). With the help of new mutations, retrotransposons' sequences can vary significantly, and these elements can acquire new and diverse functions in the host genomes. For instance, they can still play a functional role as exapted enhancers or transcriptional start sites (Rangwala et al., 2009; Deininger, 2011; Su et al., 2014; Babaian & Mager, 2016), by inserting Transcription Factor Binding Sites (TFBS) (Emera & Wagner, 2012; Lynch et al., 2015) or by acting as novel RNA genes such as long non-coding RNAs (lncRNAs) (Hezroni et al., 2015). TEs can also affect translation regulation when transcribed within a mRNA and contribute to protein-coding regions both at the transcript and the protein level, and some TE-encoded proteins have been domesticated and are part of host genes (Garcia-Perez et al., 2016). Moreover, TEs can be involved in the generation of genes and pseudogenes (Moran et al., 1999; Ohshima et al., 2003; Sayah et al., 2004) and can generate diversity through active transposition in germline cells, which can create novel insertions that are capable of being inherited, thereby generating human-specific polymorphisms. TEs also play key roles in embryogenesis (Friedli & Trono, 2015; Gerdes et al., 2016; Percharde et al., 2018), speciation (Guichard et al., 2018; Ricci et al., 2018) and possibly neurogenesis (Muotri et al., 2005; Evrony et al., 2012; Notwell et al., 2015).

Carmi and colleagues (2011) found many pairs of retrotransposons containing long clusters of G-to-A mutations that cannot be attributed to random mutagenesis and demonstrated that these clusters, which

they found across different mammalian genomes and retrotransposon families, are the hallmark of APOBEC3 activity, suggesting a potential mechanism for retrotransposon domestication (Carmi et al., 2011). Therefore, DNA editing can help to explain how some retrotransposons have acquired such a diverse collection of functions in primate genomes (Carmi et al., 2011).

**Emerging perspectives**

Located on human chromosome 22, the APOBEC3 family of deaminase proteins has a wide range of functions, from tumor progression to viruses and retrotransposons restriction.

In this review, we discussed the different mechanisms by which A3 genes inhibit retrotransposons proliferation, by inducing C-to-U or G-to-A hypermutations in the nascent DNA or by interacting with the L1 complex of RNP and ORF1 in the cell cytoplasm.

The origin of the APOBEC3 gene cluster is an extraordinary example of coevolution between a defense mechanism and its counterpart: different A3 genes appeared by duplications, fusions and rearrangements in primates, and such events happened concurrently with the invasion of some retrotransposons, most notably ERV and L1 (Figure 7.1). Indeed, a strong evolutionary arms race shaped the evolution of A3 genes and retrotransposons in primates and Homo. Diversification and functional differentiation of antiviral genes has led to the establishment of species-specific antiviral defenses, such as that of APOBEC3, which plays a pivotal role in regulating cross-species viral transmission (Revathidevi et al., 2021). In summary, the defensive roles of A3 genes are attributable to their rapid and complicated evolution, driven by retrotransposons.

Karagianni and colleagues (2022) have recently suggested that RNA editing is an emerging mechanism in disease development, displaying common and disease-specific patterns, in the context of neuropsychiatric and neurodegenerative disorders (Karagianni et al., 2022). APOBEC3-driven RNA editing is responsible for alternative splicing, regulation, degradation, and secondary structure changes that directly affect nucleic acid functions in the brain (Karagianni et al., 2022). As highlighted previously, A3s are involved in retrotransposons inhibition and, although the mechanistic

details of the functional and evolutionary impact of retrotransposons in the brain and nervous system are still unknown, an increasing bulk of data suggest that TEs play a role in the development of the CNS (reviewed in Linker et al., 2017; Suarez et al., 2018; Li & Larsen, 2021) and contribute to neurological disorders (as recently reviewed in Ochoa Thomas et al., 2020; Chesnokova et al., 2022; Ahmadi et al., 2020). Commonly edited RNAs represent potential disease-associated targets for therapeutic and diagnostic values (Karagianni et al., 2022): indeed, a recent work by Macciardi and colleagues (2022) showed that a strong dysregulation in TEs expression is associated with different stages of Alzheimer's disease development, providing clues on the use of expression profiles as potential predictors of the disease (Macciardi et al., 2022). These findings have major implications for understanding the neuroplasticity of the brain, which probably had a remarkable impact on brain evolution in mammals, especially in Hominids, and could contribute to vulnerability to neurological disorders.

During mammalian embryonic development, retrotransposons are expressed at different levels and play essential roles in embryonic stem cells (ESC) differentiation and pre-implantation embryos, as suggested by several recent publications (Garcia-Perez et al., 2016; Gerdes et al., 2016; Percharde et al., 2018; Kohlrausch et al., 2022). Moreover, it is proposed that mutator proteins such as the APOBEC superfamily may interfere with retrotransposon expression patterns to determine different levels of TEs activity in different cell types (Wissing et al., 2011; Marchetto et al., 2013; Garcia-Perez et al., 2016). Indeed, it is suggested that A3B is highly expressed in human pluripotent stem cells, making LINE-1 silencing more efficient in the early stages of cell differentiation (Marchetto et al., 2013). This is in line with experimental findings that retrotransposons (both LTR and non-LTR) are predominantly active in human embryos at the 8-cell stage and are down-regulated following whole-genome activation (Grow et al., 2015; Yin et al., 2018). Furthermore, it is reported that all APOBEC3 proteins seem to be able to act as inhibitors of LINE-1 retrotransposons (Muckenfuss et al., 2006; Siriwardena et al., 2016; Salter et al., 2016; Protasova et al., 2021), while Alu elements are particularly restricted by A3F and A3G, sometimes in macromolecular complexes (Chiu et al., 2006; Hulme et al., 2007; Khatua et al., 2010). These observations point towards an essential contribution of

APOBECs as modulators of TEs expression across embryonic developmental trajectories, although further studies are needed to elucidate the link between A3 proteins, retrotransposons, and developmental processes.

**Conclusions**

Retrotransposons are endogenous genetic elements with the ability to move around in the genome, and because of their high mutagenic potential the majority of TEs have been faced with negative selection and are counteracted by numerous mechanisms. In primates and humans, A3 genes probably arose in the context of a strong evolutionary arms race between retrotransposons and their hosts, leading to the expansion of this family of mutator proteins, which eventually became one of the strongest host defense mechanisms. The functional relationships between exogenous viral elements and the A3 family already suggested a similar association; however, several recent studies have pinpointed the positive impact of the non-coding genome on human and primate evolution through the regulation of gene expression (for example, during embryonic development). This, in turn, is paving the way for new discoveries around the evolutionary association between retrotransposons and A3 proteins, especially in the context of primate speciation. Interestingly, one of the peculiarities of primates is related to brain development, especially in the Hominoidea lineage. Indeed, retrotransposons contributed to the evolution of the CNS throughout primate phylogeny, exerting a remarkable influence on the tradeoff between brain physiology and pathological conditions in humans. In conclusion, the competition between retrotransposons and APOBEC3 genes has not only led to the development of a diversified immune defense mechanism but has also contributed to the evolutionary relationships among the primate species that are currently known.

## 8. Acknowledgements

I am particularly grateful to prof. Alessio Boattini, PhD - my supervisor - for giving me the incredible opportunity of being a PhD student and, eventually, becoming a researcher in the field of evolutionary biology: a dream that became true. I am also very thankful to him for guiding me through this journey, and for helping me whenever necessary: these have been probably the most intense and special years of my life, and they would have been very different - probably tougher - without his supervision, friendship and understanding of the problems that I have encountered before and during these years.

I am also grateful for having met very special colleagues, the most special being my very best friend Dr. Paolo Abondio, PhD: he not only teached me how to be a good computational biologist, but also how wonderful life is and what to live for.

Many thanks go also to the other professors who gave me the opportunity to analyze their data, complete my analyses and pursue my project's goals: prof. Fabio Macciardi, MD/PhD, prof. Massimo Mezzavilla, PhD, and prof. Gabriele Scorrano, PhD. Without them most of the studies presented in this thesis wouldn't have seen the light.

I want to thank my family, which always encouraged me to follow my dreams: my mom and dad, Lauretta and Graziano, and my aunts Fiorella and Nadia. I am also thankful for the best friends I could ever have: Angela, Alessia, Arianna, Vilma and Ezio. You will always be in my mind and heart.

Finally, I want to thank Barbara McClintock, discoverer of transposable elements, whose pioneering work always inspired me and thanks to whom I had and will have the opportunity to study transposons, my very obsession in these years.

## 9. References


Abondio, P., Cilli, E., & Luiselli, D. (2022). Inferring Signatures of Positive Selection in Whole-Genome Sequencing Data: An Overview of Haplotype-Based Methods. *Genes*, *13*(5), 926. https://doi.org/10.3390/genes13050926

Abrusán, G., Zhang, Y., & Szilágyi, A. (2013). Structure Prediction and Analysis of DNA Transposon and LINE Retrotransposon Proteins. *Journal of Biological Chemistry*, *288*(22), 16127–16138. https://doi.org/10.1074/jbc.M113.451500

Ahmadi, A., De Toma, I., Vilor-Tejedor, N., Eftekhariyan Ghamsari, M. R., & Sadeghi, I. (2020). Transposable elements in brain health and disease. *Ageing Research Reviews*, *64*, 101153. https://doi.org/10.1016/j.arr.2020.101153

Ahn, H. W., Worman, Z. F., Lechsinska, A., Payer, L. M., Wang, T., Malik, N., Li, W., Burns, K. H., Nath, A., & Levin, H. L. (2023). Retrotransposon insertions associated with risk of neurologic and psychiatric diseases. *EMBO Reports*, *24*(1), e55197. https://doi.org/10.15252/embr.202255197

Akiyama, M., Ishigaki, K., Sakaue, S., Momozawa, Y., Horikoshi, M., Hirata, M., Matsuda, K., Ikegawa, S., Takahashi, A., Kanai, M., Suzuki, S., Matsui, D., Naito, M., Yamaji, T., Iwasaki, M., Sawada, N., Tanno, K., Sasaki, M., Hozawa, A., … Kamatani, Y. (2019). Characterizing rare and low-frequency height-associated variants in the Japanese population. *Nature Communications*, *10*(1), 4393. https://doi.org/10.1038/s41467-019-12276-5

Aldenderfer, M. (2011). Peopling the Tibetan plateau: Insights from archaeology. *High Altitude Medicine & Biology*, *12*(2), 141–147. https://doi.org/10.1089/ham.2010.1094

Alexander, D. H., & Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics*, *12*, 246. https://doi.org/10.1186/1471-2105-12-246

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664. https://doi.org/10.1101/gr.094052.109

Ali, A., Han, K., & Liang, P. (2021). Role of Transposable Elements in Gene Regulation in the Human Genome. *Life*, *11*(2), 118. https://doi.org/10.3390/life11020118

Almarri, M. A., Bergström, A., Prado-Martinez, J., Yang, F., Fu, B., Dunham, A. S., Chen, Y., Hurles, M. E., Tyler-Smith, C., & Xue, Y. (2020). Population Structure, Stratification, and Introgression of Human Structural Variation. *Cell*, *182*(1), 189-199.e15. https://doi.org/10.1016/j.cell.2020.05.024

Amorim, C. E. G., Vai, S., Posth, C., Modi, A., Koncz, I., Hakenbeck, S., La Rocca, M. C., Mende, B., Bobo, D., Pohl, W., Baricco, L. P., Bedini, E., Francalacci, P., Giostra, C., Vida, T., Winger, D., Von Freeden, U., Ghirotto, S., Lari, M., … Veeramah, K. R. (2018). Understanding 6th-century barbarian social organization and migration through paleogenomics. *Nature Communications*, *9*(1), 3547. https://doi.org/10.1038/s41467-018-06024-4

Anwar, S. L., Wulaningsih, W., & Lehmann, U. (2017). Transposable Elements in Human Cancer: Causes and Consequences of Deregulation. *International Journal of Molecular Sciences*, *18*(5), 974. https://doi.org/10.3390/ijms18050974

Arias, J. F., Koyama, T., Kinomoto, M., & Tokunaga, K. (2012). Retroelements versus APOBEC3 family members: No great escape from the magnificent seven. *Frontiers in Microbiology*, *3*, 275. https://doi.org/10.3389/fmicb.2012.00275

Arreguin, A. J., & Colognato, H. (2020). Brain Dysfunction in LAMA2-Related Congenital Muscular Dystrophy: Lessons From Human Case Reports and Mouse Models. *Frontiers in Molecular Neuroscience*, *13*, 118. https://doi.org/10.3389/fnmol.2020.00118

Azad, P., Stobdan, T., Zhou, D., Hartley, I., Akbari, A., Bafna, V., & Haddad, G. G. (2017). High-altitude adaptation in humans: From genomics to integrative physiology. *Journal of Molecular Medicine (Berlin, Germany)*, *95*(12), 1269–1282. https://doi.org/10.1007/s00109-017-1584-7

Babaian, A., & Mager, D. L. (2016). Endogenous retroviral promoter exaptation in human cancer. *Mobile DNA*, *7*, 24. https://doi.org/10.1186/s13100-016-0080-x

Baillie, J. K., Barnett, M. W., Upton, K. R., Gerhardt, D. J., Richmond, T. A., De Sapio, F., Brennan, P. M., Rizzu, P., Smith, S., Fell, M., Talbot, R. T., Gustincich, S., Freeman, T. C., Mattick, J. S., Hume, D. A., Heutink, P., Carninci, P., Jeddeloh, J. A., & Faulkner, G. J. (2011). Somatic retrotransposition alters the genetic landscape of the human brain. *Nature*, *479*(7374), 534–537. https://doi.org/10.1038/nature10531

Barnada, S. M., Isopi, A., Tejada-Martinez, D., Goubert, C., Patoori, S., Pagliaroli, L., Tracewell, M., & Trizzino, M. (2022). Genomic features underlie the co-option of SVA transposons as cis-regulatory elements in human pluripotent stem cells. *PLOS Genetics*, *18*(6), e1010225. https://doi.org/10.1371/journal.pgen.1010225

Barreiro, L. B., Ben-Ali, M., Quach, H., Laval, G., Patin, E., Pickrell, J. K., Bouchier, C., Tichit, M., Neyrolles, O., Gicquel, B., Kidd, J. R., Kidd, K. K., Alcaïs, A., Ragimbeau, J., Pellegrini, S., Abel, L., Casanova, J.-L., & Quintana-Murci, L. (2009). Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. *PLoS Genetics*, *5*(7), e1000562. https://doi.org/10.1371/journal.pgen.1000562

Barton, A. R., Sherman, M. A., Mukamel, R. E., & Loh, P.-R. (2021). Whole-exome imputation within UK Biobank powers rare coding variant association and fine-mapping analyses. *Nature Genetics*, *53*(8), 1260–1269. https://doi.org/10.1038/s41588-021-00892-1

Batzer, M. A., & Deininger, P. L. (2002). Alu repeats and human genomic diversity. *Nature Reviews. Genetics*, *3*(5), 370–379. https://doi.org/10.1038/nrg798

Batzer, M. A., Stoneking, M., Alegria-Hartman, M., Bazan, H., Kass, D. H., Shaikh, T. H., Novick, G. E., Ioannou, P. A., Scheer, W. D., & Herrera, R. J. (1994). African origin of human-specific polymorphic Alu insertions. *Proceedings of the National Academy of Sciences of the United States of America*, *91*(25), 12288–12292. https://doi.org/10.1073/pnas.91.25.12288

Baumert, T. F., Rösler, C., Malim, M. H., & von Weizsäcker, F. (2007). Hepatitis B virus DNA is subject to extensive editing by the human deaminase APOBEC3C. *Hepatology (Baltimore, Md.)*, *46*(3), Article 3. https://doi.org/10.1002/hep.21733

Bazzan, E., Saetta, M., Turato, G., Borroni, E. M., Cancellieri, C., Baraldo, S., Savino, B., Calabrese, F., Ballarin, A., Balestro, E., Mantovani, A., Cosio, M. G., Bonecchi, R., & Locati, M. (2013). Expression of the atypical chemokine receptor D6 in human alveolar macrophages in COPD. *Chest*, *143*(1), 98–106. https://doi.org/10.1378/chest.11-3220

Beall, C. M., Cavalleri, G. L., Deng, L., Elston, R. C., Gao, Y., Knight, J., Li, C., Li, J. C., Liang, Y., McCormack, M., Montgomery, H. E., Pan, H., Robbins, P. A., Shianna, K. V., Tam, S. C., Tsering, N., Veeramah, K. R., Wang, W., Wangdui, P., … Zheng, Y. T. (2010). Natural selection on *EPAS1* ( *HIF2α*

) associated with low hemoglobin concentration in Tibetan highlanders. *Proceedings of the National Academy of Sciences*, *107*(25), 11459–11464. https://doi.org/10.1073/pnas.1002443107

Beck, C. R., Collier, P., Macfarlane, C., Malig, M., Kidd, J. M., Eichler, E. E., Badge, R. M., & Moran, J. V. (2010). LINE-1 retrotransposition activity in human genomes. *Cell*, *141*(7), Article 7. https://doi.org/10.1016/j.cell.2010.05.021

Beck, C. R., Garcia-Perez, J. L., Badge, R. M., & Moran, J. V. (2011). LINE-1 elements in structural variation and disease. *Annual Review of Genomics and Human Genetics*, *12*, 187–215. https://doi.org/10.1146/annurev-genom-082509-141802

Becker, K. G., Barnes, K. C., Bright, T. J., & Wang, S. A. (2004). The genetic association database. *Nature Genetics*, *36*(5), 431–432. https://doi.org/10.1038/ng0504-431

Belancio, V. P., Roy-Engel, A. M., & Deininger, P. (2008). The impact of multiple splice sites in human L1 elements. *Gene*, *411*(1–2), 38–45. https://doi.org/10.1016/j.gene.2007.12.022

Belshaw, R., Pereira, V., Katzourakis, A., Talbot, G., Paces, J., Burt, A., & Tristem, M. (2004). Long-term reinfection of the human genome by endogenous retroviruses. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(14), Article 14. https://doi.org/10.1073/pnas.0307800101

Bennett, R. P., Salter, J. D., & Smith, H. C. (2018). A New Class of Antiretroviral Enabling Innate Immunity by Protecting APOBEC3 from HIV Vif-Dependent Degradation. *Trends in Molecular Medicine*, *24*(5), Article 5. https://doi.org/10.1016/j.molmed.2018.03.004

Bergström, A., McCarthy, S. A., Hui, R., Almarri, M. A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J., Blanché, H., Deleuze, J.-F., Cann, H., Mallick, S., Reich, D., Sandhu, M. S., Skoglund, P., Scally, A., Xue, Y., … Tyler-Smith, C. (2020). Insights into human genetic variation and population history from 929 diverse genomes. *Science (New York, N.Y.)*, *367*(6484), eaay5012. https://doi.org/10.1126/science.aay5012

Bergstrom, E. N., Luebeck, J., Petljak, M., Khandekar, A., Barnes, M., Zhang, T., Steele, C. D., Pillay, N., Landi, M. T., Bafna, V., Mischel, P. S., Harris, R. S., & Alexandrov, L. B. (2022). Mapping clustered mutations in cancer reveals APOBEC3 mutagenesis of ecDNA. *Nature*, *602*(7897), Article 7897. https://doi.org/10.1038/s41586-022-04398-6

Beunders, G., Voorhoeve, E., Golzio, C., Pardo, L. M., Rosenfeld, J. A., Talkowski, M. E., Simonic, I., Lionel, A. C., Vergult, S., Pyatt, R. E., van de Kamp, J., Nieuwint, A., Weiss, M. M., Rizzu, P., Verwer, L. E. N. I., van Spaendonk, R. M. L., Shen, Y., Wu, B., Yu, T., … Sistermans, E. A. (2013). Exonic Deletions in AUTS2 Cause a Syndromic Form of Intellectual Disability and Suggest a Critical Role for the C Terminus. *The American Journal of Human Genetics*, *92*(2), 210–220. https://doi.org/10.1016/j.ajhg.2012.12.011

Beyter, D., Ingimundardottir, H., Oddsson, A., Eggertsson, H. P., Bjornsson, E., Jonsson, H., Atlason, B. A., Kristmundsdottir, S., Mehringer, S., Hardarson, M. T., Gudjonsson, S. A., Magnusdottir, D. N., Jonasdottir, A., Jonasdottir, A., Kristjansson, R. P., Sverrisson, S. T., Holley, G., Palsson, G., Stefansson, O. A., … Stefansson, K. (2021). Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nature Genetics*, *53*(6), 779–786. https://doi.org/10.1038/s41588-021-00865-4

Bird, C. P., Stranger, B. E., Liu, M., Thomas, D. J., Ingle, C. E., Beazley, C., Miller, W., Hurles, M. E., & Dermitzakis, E. T. (2007). Fast-evolving noncoding sequences in the human genome. *Genome Biology*, *8*(6), R118. https://doi.org/10.1186/gb-2007-8-6-r118

Boer, C. G., Narcisi, R., Ramos, Y. F., Hollander, W. D., Bomer, N., Betancourt, M. C. C., Uitterlinden, A. G., Van Osch, G., Meulenbelt, I., & Van Meurs, J. J. (2015). Genetic variants in the SUPT3H-RUNX2 locus confer susceptibility for bone and cartilage related disorders via long-range regulation of RUNX2. *Osteoarthritis and Cartilage*, *23*, A71. https://doi.org/10.1016/j.joca.2015.02.145

Bogerd, H. P., Wiegand, H. L., Doehle, B. P., Lueders, K. K., & Cullen, B. R. (2006). APOBEC3A and APOBEC3B are potent inhibitors of LTR-retrotransposon function in human cells. *Nucleic Acids Research*, *34*(1), Article 1. https://doi.org/10.1093/nar/gkj416

Boissinot, S., Entezam, A., Young, L., Munson, P. J., & Furano, A. V. (2004). The insertional history of an active family of L1 retrotransposons in humans. *Genome Research*, *14*(7), Article 7. https://doi.org/10.1101/gr.2326704

Bourgeois, Y., & Boissinot, S. (2019). On the Population Dynamics of Junk: A Review on the Population Genomics of Transposable Elements. *Genes*, *10*(6), 419. https://doi.org/10.3390/genes10060419

Bourque, G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H. L., Macfarlan, T. S., Mager, D. L., & Feschotte, C. (2018). Ten things you should know about transposable elements. *Genome Biology*, *19*(1), 199. https://doi.org/10.1186/s13059-018-1577-z

Boutej, H., Rahimian, R., Thammisetty, S. S., Béland, L.-C., Lalancette-Hébert, M., & Kriz, J. (2017). Diverging mRNA and Protein Networks in Activated Microglia Reveal SRSF3 Suppresses Translation of Highly Upregulated Innate Immune Transcripts. *Cell Reports*, *21*(11), 3220–3233. https://doi.org/10.1016/j.celrep.2017.11.058

Boyd, J. L., Skove, S. L., Rouanet, J. P., Pilaz, L.-J., Bepler, T., Gordân, R., Wray, G. A., & Silver, D. L. (2015). Human-chimpanzee differences in a FZD8 enhancer alter cell-cycle dynamics in the developing neocortex. *Current Biology: CB*, *25*(6), 772–779. https://doi.org/10.1016/j.cub.2015.01.041

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *34*(5), 525–527. https://doi.org/10.1038/nbt.3519

Brent, M. B., Simonsen, U., Thomsen, J. S., & Brüel, A. (2022). Effect of Acetazolamide and Zoledronate on Simulated High Altitude-Induced Bone Loss. *Frontiers in Endocrinology*, *13*, 831369. https://doi.org/10.3389/fendo.2022.831369

Briggs, A. W., Stenzel, U., Johnson, P. L. F., Green, R. E., Kelso, J., Prüfer, K., Meyer, M., Krause, J., Ronan, M. T., Lachmann, M., & Pääbo, S. (2007). Patterns of damage in genomic DNA sequences from a Neandertal. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(37), 14616–14621. https://doi.org/10.1073/pnas.0704665104

Brouha, B., Schustak, J., Badge, R. M., Lutz-Prigge, S., Farley, A. H., Moran, J. V., & Kazazian, H. H. (2003). Hot L1s account for the bulk of retrotransposition in the human population. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(9), 5280–5285. https://doi.org/10.1073/pnas.0831042100

Browning, S. R., & Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics*, *81*(5), 1084–1097. https://doi.org/10.1086/521987

Bundo, M., Toyoshima, M., Okada, Y., Akamatsu, W., Ueda, J., Nemoto-Miyauchi, T., Sunaga, F., Toritsuka, M., Ikawa, D., Kakita, A., Kato, M., Kasai, K., Kishimoto, T., Nawa, H., Okano, H., Yoshikawa, T., Kato, T., & Iwamoto, K. (2014). Increased L1 Retrotransposition in the Neuronal Genome in Schizophrenia. *Neuron*, *81*(2), 306–313. https://doi.org/10.1016/j.neuron.2013.10.053

Burns, K. H. (2017). Transposable elements in cancer. *Nature Reviews. Cancer*, *17*(7), 415–424. https://doi.org/10.1038/nrc.2017.35

Byrska-Bishop, M., Evani, U. S., Zhao, X., Basile, A. O., Abel, H. J., Regier, A. A., Corvelo, A., Clarke, W. E., Musunuri, R., Nagulapalli, K., Fairley, S., Runnels, A., Winterkorn, L., Lowy, E., Human Genome Structural Variation Consortium, Paul Flicek, null, Germer, S., Brand, H., Hall, I. M., … Zody, M. C. (2022). High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell*, *185*(18), 3426-3440.e19. https://doi.org/10.1016/j.cell.2022.08.004

Cai, C., Langfelder, P., Fuller, T. F., Oldham, M. C., Luo, R., van den Berg, L. H., Ophoff, R. A., & Horvath, S. (2010). Is human blood a good surrogate for brain tissue in transcriptional studies? *BMC Genomics*, *11*, 589. https://doi.org/10.1186/1471-2164-11-589

Calarco, J. A., Xing, Y., Cáceres, M., Calarco, J. P., Xiao, X., Pan, Q., Lee, C., Preuss, T. M., & Blencowe, B. J. (2007). Global analysis of alternative splicing differences between humans and chimpanzees. *Genes & Development*, *21*(22), 2963–2975. https://doi.org/10.1101/gad.1606907

Cancellieri, C., Caronni, N., Vacchini, A., Savino, B., Borroni, E. M., Locati, M., & Bonecchi, R. (2013). Review: Structure-function and biological properties of the atypical chemokine receptor D6. *Molecular Immunology*, *55*(1), 87–93. https://doi.org/10.1016/j.molimm.2012.08.003

Cao, X., Zhang, Y., Payer, L. M., Lords, H., Steranka, J. P., Burns, K. H., & Xing, J. (2020). Polymorphic mobile element insertions contribute to gene expression and alternative splicing in human tissues. *Genome Biology*, *21*(1), 185. https://doi.org/10.1186/s13059-020-02101-4

Capra, J. A., Erwin, G. D., McKinsey, G., Rubenstein, J. L. R., & Pollard, K. S. (2013). Many human accelerated regions are developmental enhancers. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *368*(1632), 20130025. https://doi.org/10.1098/rstb.2013.0025

Carbone, L., Harris, R. A., Mootnick, A. R., Milosavljevic, A., Martin, D. I. K., Rocchi, M., Capozzi, O., Archidiacono, N., Konkel, M. K., Walker, J. A., Batzer, M. A., & de Jong, P. J. (2012). Centromere remodeling in Hoolock leuconedys (Hylobatidae) by a new transposable element unique to the gibbons. *Genome Biology and Evolution*, *4*(7), Article 7. https://doi.org/10.1093/gbe/evs048

Carithers, L. J., Ardlie, K., Barcus, M., Branton, P. A., Britton, A., Buia, S. A., Compton, C. C., DeLuca, D. S., Peter-Demchok, J., Gelfand, E. T., Guan, P., Korzeniewski, G. E., Lockhart, N. C., Rabiner, C. A., Rao, A. K., Robinson, K. L., Roche, N. V., Sawyer, S. J., Segrè, A. V., … GTEx Consortium. (2015). A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. *Biopreservation and Biobanking*, *13*(5), 311–319. https://doi.org/10.1089/bio.2015.0032

Carlisle, H. J., Luong, T. N., Medina-Marino, A., Schenker, L., Khorosheva, E., Indersmitten, T., Gunapala, K. M., Steele, A. D., O'Dell, T. J., Patterson, P. H., & Kennedy, M. B. (2011). Deletion of Densin-180 Results in Abnormal Behaviors Associated with Mental Illness and Reduces mGluR5 and DISC1 in the Postsynaptic Density Fraction. *Journal of Neuroscience*, *31*(45), 16194–16207. https://doi.org/10.1523/JNEUROSCI.5877-10.2011

Carmi, S., Church, G. M., & Levanon, E. Y. (2011). Large-scale DNA editing of retrotransposons accelerates mammalian genome evolution. *Nature Communications*, *2*, 519. https://doi.org/10.1038/ncomms1525

Castillo, E., Leon, J., Mazzei, G., Abolhassani, N., Haruyama, N., Saito, T., Saido, T., Hokama, M., Iwaki, T., Ohara, T., Ninomiya, T., Kiyohara, Y., Sakumi, K., LaFerla, F. M., & Nakabeppu, Y. (2017). Comparative profiling of cortical gene expression in Alzheimer's disease patients and mouse models demonstrates a link between amyloidosis and neuroinflammation. *Scientific Reports*, *7*(1), 17762. https://doi.org/10.1038/s41598-017-17999-3

Chalopin, D., Naville, M., Plard, F., Galiana, D., & Volff, J.-N. (2015). Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. *Genome Biology and Evolution*, *7*(2), Article 2. https://doi.org/10.1093/gbe/evv005

Changeux, J.-P. (2017). Climbing Brain Levels of Organisation from Genes to Consciousness. *Trends in Cognitive Sciences*, *21*(3), 168–181. https://doi.org/10.1016/j.tics.2017.01.004

Charlesworth, B. (2009). Fundamental concepts in genetics: Effective population size and patterns of molecular evolution and variation. *Nature Reviews. Genetics*, *10*(3), 195–205. https://doi.org/10.1038/nrg2526

Chen, Y.-H., Liao, D.-L., Lai, C.-H., & Chen, C.-H. (2013). Genetic analysis of AUTS2 as a susceptibility gene of heroin dependence. *Drug and Alcohol Dependence*, *128*(3), 238–242. https://doi.org/10.1016/j.drugalcdep.2012.08.029

Chen, Y.-J., Chang, W.-A., Wu, L.-Y., Huang, C.-F., Chen, C.-H., & Kuo, P.-L. (2019). Identification of Novel Genes in Osteoarthritic Fibroblast-Like Synoviocytes Using Next-Generation Sequencing and Bioinformatics Approaches. *International Journal of Medical Sciences*, *16*(8), 1057–1071. https://doi.org/10.7150/ijms.35611

Chénais, B. (2022). Transposable Elements and Human Diseases: Mechanisms and Implication in the Response to Environmental Pollutants. *International Journal of Molecular Sciences*, *23*(5), 2551. https://doi.org/10.3390/ijms23052551

Chesnokova, E., Beletskiy, A., & Kolosov, P. (2022). The Role of Transposable Elements of the Human Genome in Neuronal Function and Pathology. *International Journal of Molecular Sciences*, *23*(10), Article 10. https://doi.org/10.3390/ijms23105847

Cheung, J. P., Tubbs, J. D., & Sham, P. C. (2022). Extended gene set analysis of human neuro-psychiatric traits shows enrichment in brain-expressed human accelerated regions across development. *Schizophrenia Research*, *246*, 148–155. https://doi.org/10.1016/j.schres.2022.06.023

Chimpanzee Sequencing and Analysis Consortium. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, *437*(7055), Article 7055. https://doi.org/10.1038/nature04072

Chiu, Y.-L., Witkowska, H. E., Hall, S. C., Santiago, M., Soros, V. B., Esnault, C., Heidmann, T., & Greene, W. C. (2006). High-molecular-mass APOBEC3G complexes restrict Alu retrotransposition. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(42), Article 42. https://doi.org/10.1073/pnas.0604524103

Christakoudi, S., Evangelou, E., Riboli, E., & Tsilidis, K. K. (2021). GWAS of allometric body-shape indices in UK Biobank identifies loci suggesting associations with morphogenesis, organogenesis,

adrenal cell renewal and cancer. *Scientific Reports*, *11*(1), 10688. https://doi.org/10.1038/s41598-021-89176-6

Chuong, E. B., Elde, N. C., & Feschotte, C. (2017). Regulatory activities of transposable elements: From conflicts to benefits. *Nature Reviews. Genetics*, *18*(2), 71–86. https://doi.org/10.1038/nrg.2016.139

Cocca, M., Barbieri, C., Concas, M. P., Robino, A., Brumat, M., Gandin, I., Trudu, M., Sala, C. F., Vuckovic, D., Girotto, G., Matullo, G., Polasek, O., Kolčić, I., Gasparini, P., Soranzo, N., Toniolo, D., & Mezzavilla, M. (2020). A bird's-eye view of Italian genomic variation through whole-genome sequencing. *European Journal of Human Genetics: EJHG*, *28*(4), 435–444. https://doi.org/10.1038/s41431-019-0551-x

Collins, R. L., Brand, H., Karczewski, K. J., Zhao, X., Alföldi, J., Francioli, L. C., Khera, A. V., Lowther, C., Gauthier, L. D., Wang, H., Watts, N. A., Solomonson, M., O'Donnell-Luria, A., Baumann, A., Munshi, R., Walker, M., Whelan, C. W., Huang, Y., Brookings, T., … Talkowski, M. E. (2020). A structural variation reference for medical and population genetics. *Nature*, *581*(7809), 444–451. https://doi.org/10.1038/s41586-020-2287-8

Colombo, A. R., Elias, H. K., & Ramsingh, G. (2018). Senescence induction universally activates transposable element expression. *Cell Cycle (Georgetown, Tex.)*, *17*(14), 1846–1857. https://doi.org/10.1080/15384101.2018.1502576

Conticello, S. G. (2008). The AID/APOBEC family of nucleic acid mutators. *Genome Biology*, *9*(6), Article 6. https://doi.org/10.1186/gb-2008-9-6-229

Conticello, S. G., Thomas, C. J. F., Petersen-Mahrt, S. K., & Neuberger, M. S. (2005). Evolution of the AID/APOBEC family of polynucleotide (deoxy)cytidine deaminases. *Molecular Biology and Evolution*, *22*(2), Article 2. https://doi.org/10.1093/molbev/msi026

Cordaux, R., & Batzer, M. A. (2009). The impact of retrotransposons on human genome evolution. *Nature Reviews. Genetics*, *10*(10), 691–703. https://doi.org/10.1038/nrg2640

Cosby, R. L., Chang, N.-C., & Feschotte, C. (2019). Host-transposon interactions: Conflict, cooperation, and cooption. *Genes & Development*, *33*(17–18), 1098–1116. https://doi.org/10.1101/gad.327312.119

Coufal, N. G., Garcia-Perez, J. L., Peng, G. E., Yeo, G. W., Mu, Y., Lovci, M. T., Morell, M., O'Shea, K. S., Moran, J. V., & Gage, F. H. (2009). L1 retrotransposition in human neural progenitor cells. *Nature*, *460*(7259), 1127–1131. https://doi.org/10.1038/nature08248

Craig, N. L. (Ed.). (2002). *Mobile DNA II*. ASM Press.

Crespi, B., Summers, K., & Dorus, S. (2007). Adaptive evolution of genes underlying schizophrenia. *Proceedings. Biological Sciences*, *274*(1627), 2801–2810. https://doi.org/10.1098/rspb.2007.0876

Crow, T. J. (1997). Is schizophrenia the price that Homo sapiens pays for language? *Schizophrenia Research*, *28*(2–3), 127–141. https://doi.org/10.1016/s0920-9964(97)00110-2

Curto, Y., Alcaide, J., Röckle, I., Hildebrandt, H., & Nacher, J. (2019). Effects of the Genetic Depletion of Polysialyltransferases on the Structure and Connectivity of Interneurons in the Adult Prefrontal Cortex. *Frontiers in Neuroanatomy*, *13*, 6. https://doi.org/10.3389/fnana.2019.00006

Damert, A. (2015). Composite non-LTR retrotransposons in hominoid primates. *Mobile Genetic Elements*, *5*(5), Article 5. https://doi.org/10.1080/2159256X.2015.1068906

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics (Oxford, England)*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

Dawkins, R., & Krebs, J. R. (1979). Arms races between and within species. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, *205*(1161), Article 1161. https://doi.org/10.1098/rspb.1979.0081

De Cecco, M., Criscione, S. W., Peterson, A. L., Neretti, N., Sedivy, J. M., & Kreiling, J. A. (2013). Transposable elements become active and mobile in the genomes of aging mammalian somatic tissues. *Aging*, *5*(12), 867–883. https://doi.org/10.18632/aging.100621

De Cecco, M., Ito, T., Petrashen, A. P., Elias, A. E., Skvir, N. J., Criscione, S. W., Caligiana, A., Brocculi, G., Adney, E. M., Boeke, J. D., Le, O., Beauséjour, C., Ambati, J., Ambati, K., Simon, M., Seluanov,

A., Gorbunova, V., Slagboom, P. E., Helfand, S. L., … Sedivy, J. M. (2019). L1 drives IFN in senescent cells and promotes age-associated inflammation. *Nature*, *566*(7742), 73–78. https://doi.org/10.1038/s41586-018-0784-9

de Koning, A. P. J., Gu, W., Castoe, T. A., Batzer, M. A., & Pollock, D. D. (2011). Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. *PLoS Genetics*, *7*(12), e1002384. https://doi.org/10.1371/journal.pgen.1002384

de Yebra, L., Adroer, R., de Gregorio-Rocasolano, N., Blesa, R., Trullas, R., & Mahy, N. (2004). Reduced KIAA0471 mRNA expression in Alzheimer's patients: A new candidate gene product linked to the disease? *Human Molecular Genetics*, *13*(21), 2607–2612. https://doi.org/10.1093/hmg/ddh293

DeGiorgio, M., Lohmueller, K. E., & Nielsen, R. (2014). A Model-Based Approach for Identifying Signatures of Ancient Balancing Selection in Genetic Data. *PLoS Genetics*, *10*(8), e1004561. https://doi.org/10.1371/journal.pgen.1004561

Deininger, P. (2011). Alu elements: Know the SINEs. *Genome Biology*, *12*(12), 236. https://doi.org/10.1186/gb-2011-12-12-236

Delaneau, O., Zagury, J.-F., & Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods*, *10*(1), 5–6. https://doi.org/10.1038/nmeth.2307

Deng, L., Zhang, C., Yuan, K., Gao, Y., Pan, Y., Ge, X., He, Y., Yuan, Y., Lu, Y., Zhang, X., Chen, H., Lou, H., Wang, X., Lu, D., Liu, J., Tian, L., Feng, Q., Khan, A., Yang, Y., … Xu, S. (2019). Prioritizing natural-selection signals from the deep-sequencing genomic data suggests multi-variant adaptation in Tibetan highlanders. *National Science Review*, *6*(6), 1201–1222. https://doi.org/10.1093/nsr/nwz108

Deniz, Ö., Ahmed, M., Todd, C. D., Rio-Machin, A., Dawson, M. A., & Branco, M. R. (2020). Endogenous retroviruses are a source of enhancers with oncogenic potential in acute myeloid leukaemia. *Nature Communications*, *11*(1), 3506. https://doi.org/10.1038/s41467-020-17206-4

DeRosa, H., Richter, T., Wilkinson, C., & Hunter, R. G. (2022). Bridging the Gap Between Environmental Adversity and Neuropsychiatric Disorders: The Role of Transposable Elements. *Frontiers in Genetics*, *13*, 813510. https://doi.org/10.3389/fgene.2022.813510

Desrosiers, R. C., Lifson, J. D., Gibbs, J. S., Czajak, S. C., Howe, A. Y., Arthur, L. O., & Johnson, R. P. (1998). Identification of highly attenuated mutants of simian immunodeficiency virus. *Journal of Virology*, *72*(2), Article 2. https://doi.org/10.1128/JVI.72.2.1431-1437.1998

Destro Bisol, G., Anagnostou, P., Batini, C., Battaggia, C., Bertoncini, S., Boattini, A., Caciagli, L., Caló, M. C., Capelli, C., Capocasa, M., Castrí, L., Ciani, G., Coia, V., Corrias, L., Crivellaro, F., Ghiani, M. E., Luiselli, D., Mela, C., Melis, A., … Pettener, D. (2008). Italian isolates today: Geographic and linguistic factors shaping human biodiversity. *Journal of Anthropological Sciences = Rivista Di Antropologia: JASS*, *86*, 179–188.

Devine, S. E. (2023). Emerging Opportunities to Study Mobile Element Insertions and Their Source Elements in an Expanding Universe of Sequenced Human Genomes. *Genes*, *14*(10), 1923. https://doi.org/10.3390/genes14101923

DeWeerd, R. A., Németh, E., Póti, Á., Petryk, N., Chen, C.-L., Hyrien, O., Szüts, D., & Green, A. M. (2022). Prospectively defined patterns of APOBEC3A mutagenesis are prevalent in human cancers. *Cell Reports*, *38*(12), Article 12. https://doi.org/10.1016/j.celrep.2022.110555

Divers, J., Palmer, N. D., Langefeld, C. D., Brown, W. M., Lu, L., Hicks, P. J., Smith, S. C., Xu, J., Terry, J. G., Register, T. C., Wagenknecht, L. E., Parks, J. S., Ma, L., Chan, G. C., Buxbaum, S. G., Correa, A., Musani, S., Wilson, J. G., Taylor, H. A., … Freedman, B. I. (2017). Genome-wide association study of coronary artery calcified atherosclerotic plaque in African Americans with type 2 diabetes. *BMC Genetics*, *18*(1), 105. https://doi.org/10.1186/s12863-017-0572-9

Doan, R. N., Bae, B.-I., Cubelos, B., Chang, C., Hossain, A. A., Al-Saad, S., Mukaddes, N. M., Oner, O., Al-Saffar, M., Balkhy, S., Gascon, G. G., Homozygosity Mapping Consortium for Autism, Nieto, M., & Walsh, C. A. (2016). Mutations in Human Accelerated Regions Disrupt Cognition and Social Behavior. *Cell*, *167*(2), 341-354.e12. https://doi.org/10.1016/j.cell.2016.08.071

Dombroski, B. A., Mathias, S. L., Nanthakumar, E., Scott, A. F., & Kazazian, H. H. (1991). Isolation of an active human transposable element. *Science (New York, N.Y.)*, *254*(5039), Article 5039. https://doi.org/10.1126/science.1662412

Donahue, J. P., Vetter, M. L., Mukhtar, N. A., & D'Aquila, R. T. (2008). The HIV-1 Vif PPLP motif is necessary for human APOBEC3G binding and degradation. *Virology*, *377*(1), Article 1. https://doi.org/10.1016/j.virol.2008.04.017

Doyle, G. A., Crist, R. C., Karatas, E. T., Hammond, M. J., Ewing, A. D., Ferraro, T. N., Hahn, C.-G., & Berrettini, W. H. (2017). Analysis of LINE-1 Elements in DNA from Postmortem Brains of Individuals with Schizophrenia. *Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology*, *42*(13), 2602–2611. https://doi.org/10.1038/npp.2017.115

Duan, S., Wang, S., Song, Y., Gao, N., Meng, L., Gai, Y., Zhang, Y., Wang, S., Wang, C., Yu, B., Wu, J., & Yu, X. (2020). A novel HIV-1 inhibitor that blocks viral replication and rescues APOBEC3s by interrupting vif/CBFβ interaction. *The Journal of Biological Chemistry*, *295*(43), Article 43. https://doi.org/10.1074/jbc.RA120.013404

Duke, D., Wohlgemuth, E., Adams, K. R., Armstrong-Ingram, A., Rice, S. K., & Young, D. C. (2021). Earliest evidence for human use of tobacco in the Pleistocene Americas. *Nature Human Behaviour*, *6*(2), 183–192. https://doi.org/10.1038/s41562-021-01202-9

Dumas, G., Malesys, S., & Bourgeron, T. (2021). Systematic detection of brain protein-coding genes under positive selection during primate evolution and their roles in cognition. *Genome Research*, *31*(3), 484–496. https://doi.org/10.1101/gr.262113.120

Dvorak, P., Hlavac, V., & Soucek, P. (2020). 5' Untranslated Region Elements Show High Abundance and Great Variability in Homologous ABCA Subfamily Genes. *International Journal of Molecular Sciences*, *21*(22), 8878. https://doi.org/10.3390/ijms21228878

Ebenesersdóttir, S. S., Sandoval-Velasco, M., Gunnarsdóttir, E. D., Jagadeesan, A., Guðmundsdóttir, V. B., Thordardóttir, E. L., Einarsdóttir, M. S., Moore, K. H. S., Sigurðsson, Á., Magnúsdóttir, D. N., Jónsson, H., Snorradóttir, S., Hovig, E., Møller, P., Kockum, I., Olsson, T., Alfredsson, L., Hansen, T. F., Werge, T., … Helgason, A. (2018). Ancient genomes from Iceland reveal the making of a human population. *Science*, *360*(6392), 1028–1032. https://doi.org/10.1126/science.aar2625

Edgar, R. D., Jones, M. J., Meaney, M. J., Turecki, G., & Kobor, M. S. (2017). BECon: A tool for interpreting DNA methylation findings from blood in the context of brain. *Translational Psychiatry*, *7*(8), e1187. https://doi.org/10.1038/tp.2017.171

Emera, D., & Wagner, G. P. (2012). Transformation of a transposon into a derived prolactin promoter with function during human pregnancy. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(28), 11246–11251. https://doi.org/10.1073/pnas.1118566109

Enriquez-Gasca, R., Gould, P. A., & Rowe, H. M. (2020). Host Gene Regulation by Transposable Elements: The New, the Old and the Ugly. *Viruses*, *12*(10), 1089. https://doi.org/10.3390/v12101089

Erady, C., Amin, K., Onilogbo, T. O. A. E., Tomasik, J., Jukes-Jones, R., Umrania, Y., Bahn, S., & Prabakaran, S. (2022). Novel open reading frames in human accelerated regions and transposable elements reveal new leads to understand schizophrenia and bipolar disorder. *Molecular Psychiatry*, *27*(3), 1455–1468. https://doi.org/10.1038/s41380-021-01405-6

Erwin, J. A., Marchetto, M. C., & Gage, F. H. (2014). Mobile DNA elements in the generation of diversity and complexity in the brain. *Nature Reviews. Neuroscience*, *15*(8), 497–506. https://doi.org/10.1038/nrn3730

Esko, T., Mezzavilla, M., Nelis, M., Borel, C., Debniak, T., Jakkula, E., Julia, A., Karachanak, S., Khrunin, A., Kisfali, P., Krulisova, V., Aušrelé Kučinskiené, Z., Rehnström, K., Traglia, M., Nikitina-Zake, L., Zimprich, F., Antonarakis, S. E., Estivill, X., Glavač, D., … D'Adamo, P. (2013). Genetic characterization of northeastern Italian population isolates in the context of broader European genetic diversity. *European Journal of Human Genetics: EJHG*, *21*(6), 659–665. https://doi.org/10.1038/ejhg.2012.229

Esnault, C., Heidmann, O., Delebecque, F., Dewannieux, M., Ribet, D., Hance, A. J., Heidmann, T., & Schwartz, O. (2005). APOBEC3G cytidine deaminase inhibits retrotransposition of endogenous retroviruses. *Nature*, *433*(7024), Article 7024. https://doi.org/10.1038/nature03238

Evrony, G. D., Cai, X., Lee, E., Hills, L. B., Elhosary, P. C., Lehmann, H. S., Parker, J. J., Atabay, K. D., Gilmore, E. C., Poduri, A., Park, P. J., & Walsh, C. A. (2012). Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell*, *151*(3), Article 3. https://doi.org/10.1016/j.cell.2012.09.035

Fagny, M., Patin, E., Enard, D., Barreiro, L. B., Quintana-Murci, L., & Laval, G. (2014). Exploring the Occurrence of Classic Selective Sweeps in Humans Using Whole-Genome Sequencing Data Sets. *Molecular Biology and Evolution*, *31*(7), 1850–1868. https://doi.org/10.1093/molbev/msu118

Faraonio, R. (2022). Oxidative Stress and Cell Senescence Process. *Antioxidants (Basel, Switzerland)*, *11*(9), 1718. https://doi.org/10.3390/antiox11091718

Farbu, E. H., Höper, A. C., Reierth, E., Nilsson, T., & Skandfer, M. (2022). Cold exposure and musculoskeletal conditions; A scoping review. *Frontiers in Physiology*, *13*, 934163. https://doi.org/10.3389/fphys.2022.934163

Fedoroff, N. V. (2012). Transposable Elements, Epigenetics, and Genome Evolution. *Science*, *338*(6108), 758–767. https://doi.org/10.1126/science.338.6108.758

Feng, Q., Moran, J. V., Kazazian, H. H., & Boeke, J. D. (1996). Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell*, *87*(5), Article 5. https://doi.org/10.1016/s0092-8674(00)81997-2

Feng, Z., Duren, Z., Xiong, Z., Wang, S., Liu, F., Wong, W. H., & Wang, Y. (2021). HReg-CNCC reconstructs a regulatory network in human cranial neural crest cells and annotates variants in a developmental context. *Communications Biology*, *4*(1), 442. https://doi.org/10.1038/s42003-021-01970-0

Ferrari, R., Grandi, N., Tramontano, E., & Dieci, G. (2021). Retrotransposons as Drivers of Mammalian Brain Evolution. *Life (Basel, Switzerland)*, *11*(5), 376. https://doi.org/10.3390/life11050376

Ferrer-Admetlla, A., Liang, M., Korneliussen, T., & Nielsen, R. (2014). On Detecting Incomplete Soft or Hard Selective Sweeps Using Haplotype Structure. *Molecular Biology and Evolution*, *31*(5), 1275–1291. https://doi.org/10.1093/molbev/msu077

Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. *Nature Reviews Genetics*, *9*(5), 397–405. https://doi.org/10.1038/nrg2337

Feschotte, C., & Pritham, E. J. (2005). Non-mammalian c-integrases are encoded by giant transposable elements. *Trends in Genetics: TIG*, *21*(10), Article 10. https://doi.org/10.1016/j.tig.2005.07.007

Festari, M. F., Trajtenberg, F., Berois, N., Pantano, S., Revoredo, L., Kong, Y., Solari-Saquieres, P., Narimatsu, Y., Freire, T., Bay, S., Robello, C., Bénard, J., Gerken, T. A., Clausen, H., & Osinaga, E. (2017). Revisiting the human polypeptide GalNAc-T1 and T13 paralogs. *Glycobiology*, *27*(2), 140–153. https://doi.org/10.1093/glycob/cww111

Fiandaca, M. S., Zhong, X., Cheema, A. K., Orquiza, M. H., Chidambaram, S., Tan, M. T., Gresenz, C. R., FitzGerald, K. T., Nalls, M. A., Singleton, A. B., Mapstone, M., & Federoff, H. J. (2015). Plasma 24-metabolite Panel Predicts Preclinical Transition to Clinical Stages of Alzheimer's Disease. *Frontiers in Neurology*, *6*, 237. https://doi.org/10.3389/fneur.2015.00237

Fisher, A. G., Ensoli, B., Ivanoff, L., Chamberlain, M., Petteway, S., Ratner, L., Gallo, R. C., & Wong-Staal, F. (1987). The sor gene of HIV-1 is required for efficient virus transmission in vitro. *Science (New York, N.Y.)*, *237*(4817), Article 4817. https://doi.org/10.1126/science.3497453

Franchini, L. F., & Pollard, K. S. (2017). Human evolution: The non-coding revolution. *BMC Biology*, *15*(1), 89. https://doi.org/10.1186/s12915-017-0428-9

Friedli, M., & Trono, D. (2015). The developmental control of transposable elements and the evolution of higher species. *Annual Review of Cell and Developmental Biology*, *31*, 429–451. https://doi.org/10.1146/annurev-cellbio-100814-125514

Frost, B., Hemberg, M., Lewis, J., & Feany, M. B. (2014). Tau promotes neurodegeneration through global chromatin relaxation. *Nature Neuroscience*, *17*(3), 357–366. https://doi.org/10.1038/nn.3639

Fu, Q., Li, H., Moorjani, P., Jay, F., Slepchenko, S. M., Bondarev, A. A., Johnson, P. L. F., Aximu-Petri, A., Prüfer, K., De Filippo, C., Meyer, M., Zwyns, N., Salazar-García, D. C., Kuzmin, Y. V., Keates, S. G., Kosintsev, P. A., Razhev, D. I., Richards, M. P., Peristov, N. V., … Pääbo, S. (2014). Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature*, *514*(7523), 445–449. https://doi.org/10.1038/nature13810

Fueyo, R., Judd, J., Feschotte, C., & Wysocka, J. (2022). Roles of transposable elements in the regulation of mammalian transcription. *Nature Reviews Molecular Cell Biology*, *23*(7), 481–497. https://doi.org/10.1038/s41580-022-00457-y

Gabuzda, D. H., Lawrence, K., Langhoff, E., Terwilliger, E., Dorfman, T., Haseltine, W. A., & Sodroski, J. (1992). Role of vif in replication of human immunodeficiency virus type 1 in CD4+ T lymphocytes. *Journal of Virology*, *66*(11), Article 11. https://doi.org/10.1128/JVI.66.11.6489-6495.1992

Gaillard, S., Dellasanta, P., Loutan, L., & Kayser, B. (2004). Awareness, prevalence, medication use, and risk factors of acute mountain sickness in tourists trekking around the Annapurnas in Nepal: A

12-year follow-up. *High Altitude Medicine & Biology*, *5*(4), 410–419. https://doi.org/10.1089/ham.2004.5.410

Garcia-Perez, J. L., Widmann, T. J., & Adams, I. R. (2016). The impact of transposable elements on mammalian development. *Development (Cambridge, England)*, *143*(22), Article 22. https://doi.org/10.1242/dev.132639

Gardner, E. J., Lam, V. K., Harris, D. N., Chuang, N. T., Scott, E. C., Pittard, W. S., Mills, R. E., 1000 Genomes Project Consortium, & Devine, S. E. (2017). The Mobile Element Locator Tool (MELT): Population-scale mobile element discovery and biology. *Genome Research*, *27*(11), 1916–1929. https://doi.org/10.1101/gr.218032.116

Garud, N. R., Messer, P. W., Buzbas, E. O., & Petrov, D. A. (2015). Recent Selective Sweeps in North American Drosophila melanogaster Show Signatures of Soft Sweeps. *PLOS Genetics*, *11*(2), e1005004. https://doi.org/10.1371/journal.pgen.1005004

Garud, N. R., Messer, P. W., & Petrov, D. A. (2021). Detection of hard and soft selective sweeps from Drosophila melanogaster population genomic data. *PLOS Genetics*, *17*(2), e1009373. https://doi.org/10.1371/journal.pgen.1009373

Gebrie, A. (2023). Transposable elements as essential elements in the control of gene expression. *Mobile DNA*, *14*(1), 9. https://doi.org/10.1186/s13100-023-00297-3

Gerdes, P., Richardson, S. R., Mager, D. L., & Faulkner, G. J. (2016). Transposable elements in the mammalian embryo: Pioneers surviving through stealth and service. *Genome Biology*, *17*, 100. https://doi.org/10.1186/s13059-016-0965-5

Gialluisi, A., Visconti, A., Willcutt, E. G., Smith, S. D., Pennington, B. F., Falchi, M., DeFries, J. C., Olson, R. K., Francks, C., & Fisher, S. E. (2016). Investigating the effects of copy number variants on reading and language performance. *Journal of Neurodevelopmental Disorders*, *8*(1), 17. https://doi.org/10.1186/s11689-016-9147-8

Gianfrancesco, O., Bubb, V. J., & Quinn, J. P. (2017). SVA retrotransposons as potential modulators of neuropeptide gene expression. *Neuropeptides*, *64*, 3–7. https://doi.org/10.1016/j.npep.2016.09.006

Gireud-Goss, M., Reyes, S., Tewari, R., Patrizz, A., Howe, M. D., Kofler, J., Waxham, M. N., McCullough, L. D., & Bean, A. J. (2020). The ubiquitin ligase UBE4B regulates amyloid precursor protein

ubiquitination, endosomal trafficking, and amyloid β42 generation and secretion. *Molecular and Cellular Neurosciences*, *108*, 103542. https://doi.org/10.1016/j.mcn.2020.103542

Girskis, K. M., Stergachis, A. B., DeGennaro, E. M., Doan, R. N., Qian, X., Johnson, M. B., Wang, P. P., Sejourne, G. M., Nagy, M. A., Pollina, E. A., Sousa, A. M. M., Shin, T., Kenny, C. J., Scotellaro, J. L., Debo, B. M., Gonzalez, D. M., Rento, L. M., Yeh, R. C., Song, J. H. T., … Walsh, C. A. (2021). Rewiring of human neurodevelopmental gene regulatory programs by human accelerated regions. *Neuron*, *109*(20), 3239-3251.e7. https://doi.org/10.1016/j.neuron.2021.08.005

Gittelman, R. M., Hun, E., Ay, F., Madeoy, J., Pennacchio, L., Noble, W. S., Hawkins, R. D., & Akey, J. M. (2015). Comprehensive identification and analysis of human accelerated regulatory DNA. *Genome Research*, *25*(9), 1245–1255. https://doi.org/10.1101/gr.192591.115

Gnecchi-Ruscone, G. A., Abondio, P., De Fanti, S., Sarno, S., Sherpa, M. G., Sherpa, P. T., Marinelli, G., Natali, L., Di Marcello, M., Peluzzi, D., Luiselli, D., Pettener, D., & Sazzini, M. (2018). Evidence of polygenic adaptation to high altitude from Tibetan and Sherpa genomes. *Genome Biology and Evolution*. https://doi.org/10.1093/gbe/evy233

Goes, F. S., McGrath, J., Avramopoulos, D., Wolyniec, P., Pirooznia, M., Ruczinski, I., Nestadt, G., Kenny, E. E., Vacic, V., Peters, I., Lencz, T., Darvasi, A., Mulle, J. G., Warren, S. T., & Pulver, A. E. (2015). Genome-wide association study of schizophrenia in Ashkenazi Jews. *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics: The Official Publication of the International Society of Psychiatric Genetics*, *168*(8), 649–659. https://doi.org/10.1002/ajmg.b.32349

González-Fortes, G., Jones, E. R., Lightfoot, E., Bonsall, C., Lazar, C., Grandal-d'Anglade, A., Garralda, M. D., Drak, L., Siska, V., Simalcsik, A., Boroneanţ, A., Vidal Romaní, J. R., Vaqueiro Rodríguez, M., Arias, P., Pinhasi, R., Manica, A., & Hofreiter, M. (2017). Paleogenomic Evidence for Multi-generational Mixing between Neolithic Farmers and Mesolithic Hunter-Gatherers in the Lower Danube Basin. *Current Biology*, *27*(12), 1801-1810.e10. https://doi.org/10.1016/j.cub.2017.05.023

González-Peñas, J., de Hoyos, L., Díaz-Caneja, C. M., Andreu-Bernabeu, Á., Stella, C., Gurriarán, X., Fañanás, L., Bobes, J., González-Pinto, A., Crespo-Facorro, B., Martorell, L., Vilella, E., Muntané, G., Molto, M. D., Gonzalez-Piqueras, J. C., Parellada, M., Arango, C., & Costas, J. (2023). Recent

natural selection conferred protection against schizophrenia by non-antagonistic pleiotropy. *Scientific Reports*, *13*(1), 15500. https://doi.org/10.1038/s41598-023-42578-0

Goodier, J. L. (2016). Restricting retrotransposons: A review. *Mobile DNA*, *7*, 16. https://doi.org/10.1186/s13100-016-0070-z

Goubert, C., Thomas, J., Payer, L. M., Kidd, J. M., Feusier, J., Watkins, W. S., Burns, K. H., Jorde, L. B., & Feschotte, C. (2020). TypeTE: A tool to genotype mobile element insertions from whole genome resequencing data. *Nucleic Acids Research*, *48*(6), e36. https://doi.org/10.1093/nar/gkaa074

Goubert, C., Zevallos, N. A., & Feschotte, C. (2020). Contribution of unfixed transposable element insertions to human regulatory variation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1795), 20190331. https://doi.org/10.1098/rstb.2019.0331

Gouy, A., Daub, J. T., & Excoffier, L. (2017). Detecting gene subnetworks under selection in biological pathways. *Nucleic Acids Research*, *45*(16), e149–e149. https://doi.org/10.1093/nar/gkx626

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–652. https://doi.org/10.1038/nbt.1883

Grandi, N., Pisano, M. P., Demurtas, M., Blomberg, J., Magiorkinis, G., Mayer, J., & Tramontano, E. (2020). Identification and characterization of ERV-W-like sequences in Platyrrhini species provides new insights into the evolutionary history of ERV-W in primates. *Mobile DNA*, *11*, 6. https://doi.org/10.1186/s13100-020-0203-2

Green, A. M., Landry, S., Budagyan, K., Avgousti, D. C., Shalhout, S., Bhagwat, A. S., & Weitzman, M. D. (2016). APOBEC3A damages the cellular genome during DNA replication. *Cell Cycle (Georgetown, Tex.)*, *15*(7), Article 7. https://doi.org/10.1080/15384101.2016.1152426

Green, A. M., & Weitzman, M. D. (2019). The spectrum of APOBEC3 activity: From anti-viral agents to anti-cancer opportunities. *DNA Repair*, *83*, 102700. https://doi.org/10.1016/j.dnarep.2019.102700

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H.-Y., Hansen, N. F., Durand, E. Y., Malaspinas, A.-S., Jensen, J. D., Marques-Bonet, T.,

Alkan, C., Prüfer, K., Meyer, M., Burbano, H. A., … Pääbo, S. (2010). A Draft Sequence of the Neandertal Genome. *Science*, *328*(5979), 710–722. https://doi.org/10.1126/science.1188021

Griffiths, D. J. (2001). Endogenous retroviruses in the human genome sequence. *Genome Biology*, *2*(6), Article 6. https://doi.org/10.1186/gb-2001-2-6-reviews1017

Grillo, M. J., Jones, K. F. M., Carpenter, M. A., Harris, R. S., & Harki, D. A. (2022). The current toolbox for APOBEC drug discovery. *Trends in Pharmacological Sciences*, *43*(5), Article 5. https://doi.org/10.1016/j.tips.2022.02.007

Grow, E. J., Flynn, R. A., Chavez, S. L., Bayless, N. L., Wossidlo, M., Wesche, D. J., Martin, L., Ware, C. B., Blish, C. A., Chang, H. Y., Pera, R. A. R., & Wysocka, J. (2015). Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature*, *522*(7555), 221–225. https://doi.org/10.1038/nature14308

Grundy, E. E., Diab, N., & Chiappinelli, K. B. (2022). Transposable element regulation and expression in cancer. *The FEBS Journal*, *289*(5), 1160–1179. https://doi.org/10.1111/febs.15722

GTEx Consortium. (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, *45*(6), 580–585. https://doi.org/10.1038/ng.2653

Guffanti, G., Bartlett, A., Klengel, T., Klengel, C., Hunter, R., Glinsky, G., & Macciardi, F. (2018). Novel Bioinformatics Approach Identifies Transcriptional Profiles of Lineage-Specific Transposable Elements at Distinct Loci in the Human Dorsolateral Prefrontal Cortex. *Molecular Biology and Evolution*, *35*(10), 2435–2453. https://doi.org/10.1093/molbev/msy143

Guffanti, G., Gaudi, S., Fallon, J. H., Sobell, J., Potkin, S. G., Pato, C., & Macciardi, F. (2014). Transposable elements and psychiatric disorders. *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics: The Official Publication of the International Society of Psychiatric Genetics*, *165B*(3), 201–216. https://doi.org/10.1002/ajmg.b.32225

Guffanti, G., Gaudi, S., Klengel, T., Fallon, J. H., Mangalam, H., Madduri, R., Rodriguez, A., DeCrescenzo, P., Glovienka, E., Sobell, J., Klengel, C., Pato, M., Ressler, K. J., Pato, C., & Macciardi, F. (2016). LINE1 insertions as a genomic risk factor for schizophrenia: Preliminary evidence from an affected family. *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics: The*

*Official Publication of the International Society of Psychiatric Genetics*, *171*(4), 534–545. https://doi.org/10.1002/ajmg.b.32437

Guichard, E., Peona, V., Malagoli Tagliazucchi, G., Abitante, L., Jagoda, E., Musella, M., Ricci, M., Rubio-Roldán, A., Sarno, S., Luiselli, D., Pettener, D., Taccioli, C., Pagani, L., Garcia-Perez, J. L., & Boattini, A. (2018). Impact of non-LTR retrotransposons in the differentiation and evolution of anatomically modern humans. *Mobile DNA*, *9*, 28. https://doi.org/10.1186/s13100-018-0133-4

Guio, L., & González, J. (2019). New Insights on the Evolution of Genome Content: Population Dynamics of Transposable Elements in Flies and Humans. In M. Anisimova (Ed.), *Evolutionary Genomics* (Vol. 1910, pp. 505–530). Springer New York. https://doi.org/10.1007/978-1-4939-9074-0_16

Günther, T., Malmström, H., Svensson, E. M., Omrak, A., Sánchez-Quinto, F., Kılınç, G. M., Krzewińska, M., Eriksson, G., Fraser, M., Edlund, H., Munters, A. R., Coutinho, A., Simões, L. G., Vicente, M., Sjölander, A., Jansen Sellevold, B., Jørgensen, R., Claes, P., Shriver, M. D., … Jakobsson, M. (2018). Population genomics of Mesolithic Scandinavia: Investigating early postglacial migration routes and high-latitude adaptation. *PLOS Biology*, *16*(1), e2003703. https://doi.org/10.1371/journal.pbio.2003703

Guo, C., Jeong, H.-H., Hsieh, Y.-C., Klein, H.-U., Bennett, D. A., De Jager, P. L., Liu, Z., & Shulman, J. M. (2018). Tau Activates Transposable Elements in Alzheimer's Disease. *Cell Reports*, *23*(10), 2874–2880. https://doi.org/10.1016/j.celrep.2018.05.004

Guo, H., Zhu, L., Huang, L., Sun, Z., Zhang, H., Nong, B., & Xiong, Y. (2022). APOBEC Alteration Contributes to Tumor Growth and Immune Escape in Pan-Cancer. *Cancers*, *14*(12), Article 12. https://doi.org/10.3390/cancers14122827

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., MacManes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C. N., … Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, *8*(8), 1494–1512. https://doi.org/10.1038/nprot.2013.084

Han, S., Basting, P. J., Dias, G. B., Luhur, A., Zelhof, A. C., & Bergman, C. M. (2021). Transposable element profiles reveal cell line identity and loss of heterozygosity in *Drosophila* cell culture. *Genetics*, *219*(2), iyab113. https://doi.org/10.1093/genetics/iyab113

Hata, K., & Sakaki, Y. (1997). Identification of critical CpG sites for repression of L1 transcription by DNA methylation. *Gene*, *189*(2), 227–234. https://doi.org/10.1016/s0378-1119(96)00856-6

Hathy, E., Szabó, E., Varga, N., Erdei, Z., Tordai, C., Czehlár, B., Baradits, M., Jezsó, B., Koller, J., Nagy, L., Molnár, M. J., Homolya, L., Nemoda, Z., Apáti, Á., & Réthelyi, J. M. (2020). Investigation of de novo mutations in a schizophrenia case-parent trio by induced pluripotent stem cell-based in vitro disease modeling: Convergence of schizophrenia- and autism-related cellular phenotypes. *Stem Cell Research & Therapy*, *11*(1), 504. https://doi.org/10.1186/s13287-020-01980-5

Hathy, E., Szabó, E., Vincze, K., Haltrich, I., Kiss, E., Varga, N., Erdei, Z., Várady, G., Homolya, L., Apáti, Á., & Réthelyi, J. M. (2021). Generation of multiple iPSC clones from a male schizophrenia patient carrying de novo mutations in genes KHSRP, LRRC7, and KIR2DL1, and his parents. *Stem Cell Research*, *51*, 102140. https://doi.org/10.1016/j.scr.2020.102140

Hatzikotoulas, K., Gilly, A., & Zeggini, E. (2014). Using population isolates in genetic association studies. *Briefings in Functional Genomics*, *13*(5), 371–377. https://doi.org/10.1093/bfgp/elu022

Hezroni, H., Koppstein, D., Schwartz, M. G., Avrutin, A., Bartel, D. P., & Ulitsky, I. (2015). Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell Reports*, *11*(7), 1110–1122. https://doi.org/10.1016/j.celrep.2015.04.023

Hohjoh, H., & Singer, M. F. (1997). Sequence-specific single-strand RNA binding protein encoded by the human LINE-1 retrotransposon. *The EMBO Journal*, *16*(19), Article 19. https://doi.org/10.1093/emboj/16.19.6034

Horn, A. V., Klawitter, S., Held, U., Berger, A., Vasudevan, A. A. J., Bock, A., Hofmann, H., Hanschmann, K.-M. O., Trösemeier, J.-H., Flory, E., Jabulowsky, R. A., Han, J. S., Löwer, J., Löwer, R., Münk, C., & Schumann, G. G. (2014). Human LINE-1 restriction by APOBEC3C is deaminase independent and mediated by an ORF1p interaction that affects LINE reverse transcriptase activity. *Nucleic Acids Research*, *42*(1), Article 1. https://doi.org/10.1093/nar/gkt898

Hoyt, S. J., Storer, J. M., Hartley, G. A., Grady, P. G. S., Gershman, A., de Lima, L. G., Limouse, C., Halabian, R., Wojenski, L., Rodriguez, M., Altemose, N., Rhie, A., Core, L. J., Gerton, J. L., Makalowski, W., Olson, D., Rosen, J., Smit, A. F. A., Straight, A. F., … O'Neill, R. J. (2022). From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. *Science (New York, N.Y.)*, *376*(6588), eabk3112. https://doi.org/10.1126/science.abk3112

Hsieh, P., Vollger, M. R., Dang, V., Porubsky, D., Baker, C., Cantsilieris, S., Hoekzema, K., Lewis, A. P., Munson, K. M., Sorensen, M., Kronenberg, Z. N., Murali, S., Nelson, B. J., Chiatante, G., Maggiolini, F. A. M., Blanché, H., Underwood, J. G., Antonacci, F., Deleuze, J.-F., & Eichler, E. E. (2019). Adaptive archaic introgression of copy number variants and the discovery of previously unknown human genes. *Science (New York, N.Y.)*, *366*(6463), eaax2083. https://doi.org/10.1126/science.aax2083

Hu, H., Petousi, N., Glusman, G., Yu, Y., Bohlender, R., Tashi, T., Downie, J. M., Roach, J. C., Cole, A. M., Lorenzo, F. R., Rogers, A. R., Brunkow, M. E., Cavalleri, G., Hood, L., Alpatty, S. M., Prchal, J. T., Jorde, L. B., Robbins, P. A., Simonson, T. S., & Huff, C. D. (2017). Evolutionary history of Tibetans inferred from whole-genome sequencing. *PLoS Genetics*, *13*(4), e1006675. https://doi.org/10.1371/journal.pgen.1006675

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, *4*(1), 44–57. https://doi.org/10.1038/nprot.2008.211

Huang, S., Tao, X., Yuan, S., Zhang, Y., Li, P., Beilinson, H. A., Zhang, Y., Yu, W., Pontarotti, P., Escriva, H., Le Petillon, Y., Liu, X., Chen, S., Schatz, D. G., & Xu, A. (2016). Discovery of an Active RAG Transposon Illuminates the Origins of V(D)J Recombination. *Cell*, *166*(1), Article 1. https://doi.org/10.1016/j.cell.2016.05.032

Hubisz, M. J., & Pollard, K. S. (2014). Exploring the genesis and functions of Human Accelerated Regions sheds light on their role in human evolution. *Current Opinion in Genetics & Development*, *29*, 15–21. https://doi.org/10.1016/j.gde.2014.07.005

Hublin, J.-J., Ben-Ncer, A., Bailey, S. E., Freidline, S. E., Neubauer, S., Skinner, M. M., Bergmann, I., Le Cabec, A., Benazzi, S., Harvati, K., & Gunz, P. (2017). New fossils from Jebel Irhoud, Morocco and

the pan-African origin of Homo sapiens. *Nature*, *546*(7657), 289–292. https://doi.org/10.1038/nature22336

Huerta-Sánchez, E., Jin, X., Asan, Bianba, Z., Peter, B. M., Vinckenbosch, N., Liang, Y., Yi, X., He, M., Somel, M., Ni, P., Wang, B., Ou, X., Huasang, Luosang, J., Cuo, Z. X. P., Li, K., Gao, G., Yin, Y., … Nielsen, R. (2014). Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature*, *512*(7513), 194–197. https://doi.org/10.1038/nature13408

Hulme, A. E., Bogerd, H. P., Cullen, B. R., & Moran, J. V. (2007). Selective inhibition of Alu retrotransposition by APOBEC3G. *Gene*, *390*(1–2), Article 1–2. https://doi.org/10.1016/j.gene.2006.08.032

Hultquist, J. F., Lengyel, J. A., Refsland, E. W., LaRue, R. S., Lackey, L., Brown, W. L., & Harris, R. S. (2011). Human and rhesus APOBEC3D, APOBEC3F, APOBEC3G, and APOBEC3H demonstrate a conserved capacity to restrict Vif-deficient HIV-1. *Journal of Virology*, *85*(21), Article 21. https://doi.org/10.1128/JVI.05238-11

Hummel, M., Edelmann, D., & Kopp-Schneider, A. (2017). Clustering of samples and variables with mixed-type data. *PloS One*, *12*(11), e0188274. https://doi.org/10.1371/journal.pone.0188274

Hunter, R. G., Gagnidze, K., McEwen, B. S., & Pfaff, D. W. (2015). Stress and the dynamic genome: Steroids, epigenetics, and the transposome. *Proceedings of the National Academy of Sciences*, *112*(22), 6828–6833. https://doi.org/10.1073/pnas.1411260111

Hunyara, J. L., Daly, K. M., Torres, K., Yurgel, M. E., Komal, R., Hattar, S., & Kolodkin, A. L. (2023). Teneurin-3 regulates the generation of non-image-forming visual circuitry and responsiveness to light in the suprachiasmatic nucleus. *PLoS Biology*, *21*(12), e3002412. https://doi.org/10.1371/journal.pbio.3002412

Ianc, B., Ochis, C., Persch, R., Popescu, O., & Damert, A. (2014). Hominoid composite non-LTR retrotransposons-variety, assembly, evolution, and structural determinants of mobilization. *Molecular Biology and Evolution*, *31*(11), Article 11. https://doi.org/10.1093/molbev/mst256

Ito, J., Gifford, R. J., & Sato, K. (2020). Retroviruses drive the rapid evolution of mammalian APOBEC3 genes. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(1), Article 1. https://doi.org/10.1073/pnas.1914183116

Iwatani, Y., Chan, D. S. B., Wang, F., Stewart-Maynard, K., Sugiura, W., Gronenborn, A. M., Rouzina, I., Williams, M. C., Musier-Forsyth, K., & Levin, J. G. (2007). Deaminase-independent inhibition of HIV-1 reverse transcription by APOBEC3G. *Nucleic Acids Research*, *35*(21), Article 21. https://doi.org/10.1093/nar/gkm750

Jain, P. R., Yates, M., de Celis, C. R., Drineas, P., Jahanshad, N., Thompson, P., & Paschou, P. (2023). Multiomic approach and Mendelian randomization analysis identify causal associations between blood biomarkers and subcortical brain structure volumes. *NeuroImage*, *284*, 120466. https://doi.org/10.1016/j.neuroimage.2023.120466

Jakobsdottir, G. M., Brewer, D. S., Cooper, C., Green, C., & Wedge, D. C. (2022). APOBEC3 mutational signatures are associated with extensive and diverse genomic instability across multiple tumour types. *BMC Biology*, *20*(1), Article 1. https://doi.org/10.1186/s12915-022-01316-0

Jarmuz, A., Chester, A., Bayliss, J., Gisbourne, J., Dunham, I., Scott, J., & Navaratnam, N. (2002). An anthropoid-specific locus of orphan C to U RNA-editing enzymes on chromosome 22. *Genomics*, *79*(3), Article 3. https://doi.org/10.1006/geno.2002.6718

Jelassi, A., Slimani, A., Rabès, J. P., Jguirim, I., Abifadel, M., Boileau, C., Najah, M., M'rabet, S., Mzid, J., Slimane, M. N., & Varret, M. (2012). Genomic characterization of two deletions in the LDLR gene in Tunisian patients with familial hypercholesterolemia. *Clinica Chimica Acta; International Journal of Clinical Chemistry*, *414*, 146–151. https://doi.org/10.1016/j.cca.2012.08.002

Jeong, C., Alkorta-Aranburu, G., Basnyat, B., Neupane, M., Witonsky, D. B., Pritchard, J. K., Beall, C. M., & Di Rienzo, A. (2014). Admixture facilitates genetic adaptations to high altitude in Tibet. *Nature Communications*, *5*(1), 3281. https://doi.org/10.1038/ncomms4281

Jeong, C., Ozga, A. T., Witonsky, D. B., Malmström, H., Edlund, H., Hofman, C. A., Hagan, R. W., Jakobsson, M., Lewis, C. M., Aldenderfer, M. S., Di Rienzo, A., & Warinner, C. (2016). Long-term genetic stability and a high-altitude East Asian origin for the peoples of the high valleys of the Himalayan arc. *Proceedings of the National Academy of Sciences*, *113*(27), 7485–7490. https://doi.org/10.1073/pnas.1520844113

Johnson, W. E. (2019). Origins and evolutionary consequences of ancient endogenous retroviruses. *Nature Reviews. Microbiology*, *17*(6), Article 6. https://doi.org/10.1038/s41579-019-0189-2

Joilin, G., Leigh, P. N., Newbury, S. F., & Hafezparast, M. (2019). An Overview of MicroRNAs as Biomarkers of ALS. *Frontiers in Neurology*, *10*, 186. https://doi.org/10.3389/fneur.2019.00186

Jones, E. R., Gonzalez-Fortes, G., Connell, S., Siska, V., Eriksson, A., Martiniano, R., McLaughlin, R. L., Gallego Llorente, M., Cassidy, L. M., Gamba, C., Meshveliani, T., Bar-Yosef, O., Müller, W., Belfer-Cohen, A., Matskevich, Z., Jakeli, N., Higham, T. F. G., Currat, M., Lordkipanidze, D., … Bradley, D. G. (2015). Upper Palaeolithic genomes reveal deep roots of modern Eurasians. *Nature Communications*, *6*(1), 8912. https://doi.org/10.1038/ncomms9912

Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. F., & Orlando, L. (2013). mapDamage2.0: Fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics*, *29*(13), 1682–1684. https://doi.org/10.1093/bioinformatics/btt193

Jönsson, M. E., Garza, R., Johansson, P. A., & Jakobsson, J. (2020). Transposable Elements: A Common Feature of Neurodevelopmental and Neurodegenerative Disorders. *Trends in Genetics: TIG*, *36*(8), 610–623. https://doi.org/10.1016/j.tig.2020.05.004

Jun, G., Wing, M. K., Abecasis, G. R., & Kang, H. M. (2015). An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Research*, *25*(6), 918–925. https://doi.org/10.1101/gr.176552.114

Kanzawa-Kiriyama, H., Jinam, T. A., Kawai, Y., Sato, T., Hosomichi, K., Tajima, A., Adachi, N., Matsumura, H., Kryukov, K., Saitou, N., & Shinoda, K.-I. (2019). Late Jomon male and female genome sequences from the Funadomari site in Hokkaido, Japan. *Anthropological Science*, *127*(2), 83–108. https://doi.org/10.1537/ase.190415

Kapitonov, V. V., & Jurka, J. (2005). RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. *PLoS Biology*, *3*(6), Article 6. https://doi.org/10.1371/journal.pbio.0030181

Kapitonov, V. V., & Jurka, J. (2006). Self-synthesizing DNA transposons in eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(12), Article 12. https://doi.org/10.1073/pnas.0600833103

Kapitonov, V. V., & Jurka, J. (2008). A universal classification of eukaryotic transposable elements implemented in Repbase. *Nature Reviews. Genetics*, *9*(5), Article 5. https://doi.org/10.1038/nrg2165-c1

Karagianni, K., Pettas, S., Christoforidou, G., Kanata, E., Bekas, N., Xanthopoulos, K., Dafou, D., & Sklaviadis, T. (2022). A Systematic Review of Common and Brain-Disease-Specific RNA Editing Alterations Providing Novel Insights into Neurological and Neurodegenerative Disease Manifestations. *Biomolecules*, *12*(3), Article 3. https://doi.org/10.3390/biom12030465

Karlsson Linnér, R., Biroli, P., Kong, E., Meddens, S. F. W., Wedow, R., Fontana, M. A., Lebreton, M., Tino, S. P., Abdellaoui, A., Hammerschlag, A. R., Nivard, M. G., Okbay, A., Rietveld, C. A., Timshel, P. N., Trzaskowski, M., Vlaming, R. de, Zünd, C. L., Bao, Y., Buzdugan, L., … Beauchamp, J. P. (2019). Genome-wide association analyses of risk tolerance and risky behaviors in over 1 million individuals identify hundreds of loci and shared genetic influences. *Nature Genetics*, *51*(2), 245–257. https://doi.org/10.1038/s41588-018-0309-3

Kazazian, H. H., Wong, C., Youssoufian, H., Scott, A. F., Phillips, D. G., & Antonarakis, S. E. (1988). Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature*, *332*(6160), Article 6160. https://doi.org/10.1038/332164a0

Keller, A., Graefen, A., Ball, M., Matzas, M., Boisguerin, V., Maixner, F., Leidinger, P., Backes, C., Khairat, R., Forster, M., Stade, B., Franke, A., Mayer, J., Spangler, J., McLaughlin, S., Shah, M., Lee, C., Harkins, T. T., Sartori, A., … Zink, A. (2012). New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. *Nature Communications*, *3*(1), 698. https://doi.org/10.1038/ncomms1701

Khan, H., Smit, A., & Boissinot, S. (2006). Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. *Genome Research*, *16*(1), Article 1. https://doi.org/10.1101/gr.4001406

Khatua, A. K., Taylor, H. E., Hildreth, J. E. K., & Popik, W. (2010). Inhibition of LINE-1 and Alu retrotransposition by exosomes encapsidating APOBEC3G and APOBEC3F. *Virology*, *400*(1), Article 1. https://doi.org/10.1016/j.virol.2010.01.021

Kim, D. S., & Hahn, Y. (2011). Identification of human-specific transcript variants induced by DNA insertions in the human genome. *Bioinformatics (Oxford, England)*, *27*(1), 14–21. https://doi.org/10.1093/bioinformatics/btq612

Kim, H. R., Kim, D. H., An, J. Y., Kang, D., Park, J. W., Hwang, E. M., Seo, E. J., Jang, I. H., Ha, C. M., & Lee, B. J. (2020). NELL2 Function in Axon Development of Hippocampal Neurons. *Molecules and Cells*, *43*(6), 581–589. https://doi.org/10.14348/molcells.2020.0032

Kim, T. H., Barrera, L. O., Zheng, M., Qu, C., Singer, M. A., Richmond, T. A., Wu, Y., Green, R. D., & Ren, B. (2005). A high-resolution map of active promoters in the human genome. *Nature*, *436*(7052), 876–880. https://doi.org/10.1038/nature03877

Knisbacher, B. A., Gerber, D., & Levanon, E. Y. (2016). DNA Editing by APOBECs: A Genomic Preserver and Transformer. *Trends in Genetics: TIG*, *32*(1), Article 1. https://doi.org/10.1016/j.tig.2015.10.005

Kohlrausch, F. B., Berteli, T. S., Wang, F., Navarro, P. A., & Keefe, D. L. (2022). Control of LINE-1 Expression Maintains Genome Integrity in Germline and Early Embryo Development. *Reproductive Sciences (Thousand Oaks, Calif.)*, *29*(2), Article 2. https://doi.org/10.1007/s43032-021-00461-1

Kõks, S., Pfaff, A. L., Singleton, L. M., Bubb, V. J., & Quinn, J. P. (2022). Non-reference genome transposable elements (TEs) have a significant impact on the progression of the Parkinson's disease. *Experimental Biology and Medicine (Maywood, N.J.)*, *247*(18), 1680–1690. https://doi.org/10.1177/15353702221117147

Kolle, G., Georgas, K., Holmes, G. P., Little, M. H., & Yamada, T. (2000). CRIM1, a novel gene encoding a cysteine-rich repeat protein, is developmentally regulated and implicated in vertebrate CNS development and organogenesis. *Mechanisms of Development*, *90*(2), 181–193. https://doi.org/10.1016/s0925-4773(99)00248-8

Konkel, M. K., Walker, J. A., & Batzer, M. A. (2010). LINEs and SINEs of primate evolution. *Evolutionary Anthropology*, *19*(6), Article 6. https://doi.org/10.1002/evan.20283

Koonin, E. V., & Krupovic, M. (2015). Evolution of adaptive immunity from transposable elements combined with innate immune systems. *Nature Reviews. Genetics*, *16*(3), Article 3. https://doi.org/10.1038/nrg3859

Kriegs, J. O., Churakov, G., Jurka, J., Brosius, J., & Schmitz, J. (2007). Evolutionary history of 7SL RNA-derived SINEs in Supraprimates. *Trends in Genetics: TIG*, *23*(4), Article 4. https://doi.org/10.1016/j.tig.2007.02.002

Kröcher, T., Röckle, I., Diederichs, U., Weinhold, B., Burkhardt, H., Yanagawa, Y., Gerardy-Schahn, R., & Hildebrandt, H. (2014). A crucial role for polysialic acid in developmental interneuron migration and the establishment of interneuron densities in the mouse prefrontal cortex. *Development (Cambridge, England)*, *141*(15), 3022–3032. https://doi.org/10.1242/dev.111773

Kuhn, M. (2008). Building Predictive Models in *R* Using the **caret** Package. *Journal of Statistical Software*, *28*(5). https://doi.org/10.18637/jss.v028.i05

Kulminski, A. M., Loiko, E., Loika, Y., & Culminskaya, I. (2022). Pleiotropic predisposition to Alzheimer's disease and educational attainment: Insights from the summary statistics analysis. *GeroScience*, *44*(1), 265–280. https://doi.org/10.1007/s11357-021-00484-1

Kulpa, D. A., & Moran, J. V. (2005). Ribonucleoprotein particle formation is necessary but not sufficient for LINE-1 retrotransposition. *Human Molecular Genetics*, *14*(21), Article 21. https://doi.org/10.1093/hmg/ddi354

Kursa, M. B., & Rudnicki, W. R. (2010). Feature Selection with the **Boruta** Package. *Journal of Statistical Software*, *36*(11). https://doi.org/10.18637/jss.v036.i11

Kyriakou, K., Lederer, C. W., Kleanthous, M., Drousiotou, A., & Malekkou, A. (2020). Acid Ceramidase Depletion Impairs Neuronal Survival and Induces Morphological Defects in Neurites Associated with Altered Gene Transcription and Sphingolipid Content. *International Journal of Molecular Sciences*, *21*(5), 1607. https://doi.org/10.3390/ijms21051607

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., … International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. https://doi.org/10.1038/35057062

Langmüller, A. M., Nolte, V., Dolezal, M., & Schlötterer, C. (2023). The genomic distribution of transposable elements is driven by spatially variable purifying selection. *Nucleic Acids Research*, *51*(17), 9203–9213. https://doi.org/10.1093/nar/gkad635

LaRocca, T. J., Cavalier, A. N., & Wahl, D. (2020). Repetitive elements as a transcriptomic marker of aging: Evidence in multiple datasets and models. *Aging Cell*, *19*(7), e13167. https://doi.org/10.1111/acel.13167

Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., Sudmant, P. H., Schraiber, J. G., Castellano, S., Lipson, M., Berger, B., Economou, C., Bollongino, R., Fu, Q., Bos, K. I., Nordenfelt, S., Li, H., De Filippo, C., Prüfer, K., … Krause, J. (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*, *513*(7518), 409–413. https://doi.org/10.1038/nature13673

Lee, H.-E., Ayarpadikannan, S., & Kim, H.-S. (2015). Role of transposable elements in genomic rearrangement, evolution, gene regulation and epigenetics in primates. *Genes & Genetic Systems*, *90*(5), Article 5. https://doi.org/10.1266/ggs.15-00016

Lee, J. Y., Ji, Z., & Tian, B. (2008). Phylogenetic analysis of mRNA polyadenylation sites reveals a role of transposable elements in evolution of the 3'-end of genes. *Nucleic Acids Research*, *36*(17), 5581–5590. https://doi.org/10.1093/nar/gkn540

Lee, Y. N., & Bieniasz, P. D. (2007). Reconstitution of an infectious human endogenous retrovirus. *PLoS Pathogens*, *3*(1), Article 1. https://doi.org/10.1371/journal.ppat.0030010

Lefebvre, V. (2019). Roles and regulation of SOX transcription factors in skeletogenesis. In *Current Topics in Developmental Biology* (Vol. 133, pp. 171–193). Elsevier. https://doi.org/10.1016/bs.ctdb.2019.01.007

Lei, L., Chen, H., Xue, W., Yang, B., Hu, B., Wei, J., Wang, L., Cui, Y., Li, W., Wang, J., Yan, L., Shang, W., Gao, J., Sha, J., Zhuang, M., Huang, X., Shen, B., Yang, L., & Chen, J. (2018). APOBEC3 induces mutations during repair of CRISPR-Cas9-generated DNA breaks. *Nature Structural & Molecular Biology*, *25*(1), Article 1. https://doi.org/10.1038/s41594-017-0004-6

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., Tukiainen, T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., … Exome Aggregation Consortium. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), 285–291. https://doi.org/10.1038/nature19057

Levchenko, A., Kanapin, A., Samsonova, A., & Gainetdinov, R. R. (2018). Human Accelerated Regions and Other Human-Specific Sequence Variations in the Context of Evolution and Their Relevance for Brain Development. *Genome Biology and Evolution*, *10*(1), 166–188. https://doi.org/10.1093/gbe/evx240

Levin, H. L., & Moran, J. V. (2011). Dynamic interactions between transposable elements and their hosts. *Nature Reviews. Genetics*, *12*(9), Article 9. https://doi.org/10.1038/nrg3030

Lewis, C. M. (2002). Genetic association studies: Design, analysis and interpretation. *Briefings in Bioinformatics*, *3*(2), 146–153. https://doi.org/10.1093/bib/3.2.146

Lewis, L. S., Huffman, K. M., Smith, I. J., Donahue, M. P., Slentz, C. A., Houmard, J. A., Hubal, M. J., Hoffman, E. P., Hauser, E. R., Siegler, I. C., & Kraus, W. E. (2018). Genetic Variation in Acid Ceramidase Predicts Non-completion of an Exercise Intervention. *Frontiers in Physiology*, *9*, 781. https://doi.org/10.3389/fphys.2018.00781

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987–2993. https://doi.org/10.1093/bioinformatics/btr509

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, J., Yoshikawa, A., Brennan, M. D., Ramsey, T. L., & Meltzer, H. Y. (2018). Genetic predictors of antipsychotic response to lurasidone identified in a genome wide association study and by schizophrenia risk genes. *Schizophrenia Research*, *192*, 194–204. https://doi.org/10.1016/j.schres.2017.04.009

Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., Cann, H. M., Barsh, G. S., Feldman, M., Cavalli-Sforza, L. L., & Myers, R. M. (2008). Worldwide human relationships

inferred from genome-wide patterns of variation. *Science (New York, N.Y.)*, *319*(5866), 1100–1104. https://doi.org/10.1126/science.1153717

Li, M., & Larsen, P. A. (2021). Primate-specific retrotransposons and the evolution of circadian networks in the human brain. *Neuroscience and Biobehavioral Reviews*, *131*, 988–1004. https://doi.org/10.1016/j.neubiorev.2021.09.049

Liang, W., Xu, J., Yuan, W., Song, X., Zhang, J., Wei, W., Yu, X.-F., & Yang, Y. (2016). APOBEC3DE Inhibits LINE-1 Retrotransposition by Interacting with ORF1p and Influencing LINE Reverse Transcriptase Activity. *PloS One*, *11*(7), Article 7. https://doi.org/10.1371/journal.pone.0157220

Liao, W.-W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J. K., Monlong, J., Abel, H. J., Buonaiuto, S., Chang, X. H., Cheng, H., Chu, J., Colonna, V., Eizenga, J. M., Feng, X., Fischer, C., Fulton, R. S., … Paten, B. (2023). A draft human pangenome reference. *Nature*, *617*(7960), 312–324. https://doi.org/10.1038/s41586-023-05896-x

Liddament, M. T., Brown, W. L., Schumacher, A. J., & Harris, R. S. (2004). APOBEC3F properties and hypermutation preferences indicate activity against HIV-1 in vivo. *Current Biology: CB*, *14*(15), Article 15. https://doi.org/10.1016/j.cub.2004.06.050

Liguori, I., Russo, G., Curcio, F., Bulli, G., Aran, L., Della-Morte, D., Gargiulo, G., Testa, G., Cacciatore, F., Bonaduce, D., & Abete, P. (2018). Oxidative stress, aging, and diseases. *Clinical Interventions in Aging*, *13*, 757–772. https://doi.org/10.2147/CIA.S158513

Lin, B. Y., Lin, W.-D., Huang, C.-K., Hsin, M.-C., Lin, W.-Y., & Pryor, A. D. (2019). Changes of gut microbiota between different weight reduction programs. *Surgery for Obesity and Related Diseases: Official Journal of the American Society for Bariatric Surgery*, *15*(5), 749–758. https://doi.org/10.1016/j.soard.2019.01.026

Lindgreen, S. (2012). AdapterRemoval: Easy cleaning of next-generation sequencing reads. *BMC Research Notes*, *5*(1), 337. https://doi.org/10.1186/1756-0500-5-337

Linker, S. B., Marchetto, M. C., Narvaiza, I., Denli, A. M., & Gage, F. H. (2017). Examining non-LTR retrotransposons in the context of the evolving primate brain. *BMC Biology*, *15*(1), Article 1. https://doi.org/10.1186/s12915-017-0409-z

Liu, E. Y., Russ, J., Cali, C. P., Phan, J. M., Amlie-Wolf, A., & Lee, E. B. (2019). Loss of Nuclear TDP-43 Is Associated with Decondensation of LINE Retrotransposons. *Cell Reports*, *27*(5), 1409-1421.e6. https://doi.org/10.1016/j.celrep.2019.04.003

Liu, W., Deng, Y., Li, Z., Chen, Y., Zhu, X., Tan, X., & Cao, G. (2021). Cancer Evo-Dev: A Theory of Inflammation-Induced Oncogenesis. *Frontiers in Immunology*, *12*, 768098. https://doi.org/10.3389/fimmu.2021.768098

Locke, D. P., Hillier, L. W., Warren, W. C., Worley, K. C., Nazareth, L. V., Muzny, D. M., Yang, S.-P., Wang, Z., Chinwalla, A. T., Minx, P., Mitreva, M., Cook, L., Delehaunty, K. D., Fronick, C., Schmidt, H., Fulton, L. A., Fulton, R. S., Nelson, J. O., Magrini, V., … Wilson, R. K. (2011). Comparative and demographic analysis of orang-utan genomes. *Nature*, *469*(7331), Article 7331. https://doi.org/10.1038/nature09687

Lockstone, H. E., Harris, L. W., Swatton, J. E., Wayland, M. T., Holland, A. J., & Bahn, S. (2007). Gene expression profiling in the adult Down syndrome brain. *Genomics*, *90*(6), 647–660. https://doi.org/10.1016/j.ygeno.2007.08.005

Lombardo, M. V. (2021). Ribosomal protein genes in post-mortem cortical tissue and iPSC-derived neural progenitor cells are commonly upregulated in expression in autism. *Molecular Psychiatry*, *26*(5), 1432–1435. https://doi.org/10.1038/s41380-020-0773-x

Lou, H., Lu, Y., Lu, D., Fu, R., Wang, X., Feng, Q., Wu, S., Yang, Y., Li, S., Kang, L., Guan, Y., Hoh, B.-P., Chung, Y.-J., Jin, L., Su, B., & Xu, S. (2015). A 3.4-kb Copy-Number Deletion near EPAS1 Is Significantly Enriched in High-Altitude Tibetans but Absent from the Denisovan Sequence. *The American Journal of Human Genetics*, *97*(1), 54–66. https://doi.org/10.1016/j.ajhg.2015.05.005

Löwer, R., Löwer, J., & Kurth, R. (1996). The viruses in all of us: Characteristics and biological significance of human endogenous retrovirus sequences. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(11), Article 11. https://doi.org/10.1073/pnas.93.11.5177

Lu, D., Lou, H., Yuan, K., Wang, X., Wang, Y., Zhang, C., Lu, Y., Yang, X., Deng, L., Zhou, Y., Feng, Q., Hu, Y., Ding, Q., Yang, Y., Li, S., Jin, L., Guan, Y., Su, B., Kang, L., & Xu, S. (2016). Ancestral Origins and Genetic History of Tibetan Highlanders. *The American Journal of Human Genetics*, *99*(3), 580–594. https://doi.org/10.1016/j.ajhg.2016.07.002

Lu, X., Sachs, F., Ramsay, L., Jacques, P.-É., Göke, J., Bourque, G., & Ng, H.-H. (2014). The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. *Nature Structural & Molecular Biology*, *21*(4), 423–425. https://doi.org/10.1038/nsmb.2799

Lupan, I., Bulzu, P., Popescu, O., & Damert, A. (2015). Lineage specific evolution of the VNTR composite retrotransposon central domain and its role in retrotransposition of gibbon LAVA elements. *BMC Genomics*, *16*, 389. https://doi.org/10.1186/s12864-015-1543-z

Lynch, V. J., Nnamani, M. C., Kapusta, A., Brayer, K., Plaza, S. L., Mazur, E. C., Emera, D., Sheikh, S. Z., Grützner, F., Bauersachs, S., Graf, A., Young, S. L., Lieb, J. D., DeMayo, F. J., Feschotte, C., & Wagner, G. P. (2015). Ancient transposable elements transformed the uterine regulatory landscape and transcriptome during the evolution of mammalian pregnancy. *Cell Reports*, *10*(4), 551–561. https://doi.org/10.1016/j.celrep.2014.12.052

Macciardi, F., Giulia Bacalini, M., Miramontes, R., Boattini, A., Taccioli, C., Modenini, G., Malhas, R., Anderlucci, L., Gusev, Y., Gross, T. J., Padilla, R. M., Fiandaca, M. S., Head, E., Guffanti, G., Federoff, H. J., & Mapstone, M. (2022). A retrotransposon storm marks clinical phenoconversion to late-onset Alzheimer's disease. *GeroScience*, *44*(3), 1525–1550. https://doi.org/10.1007/s11357-022-00580-w

Mafessoni, F., Grote, S., De Filippo, C., Slon, V., Kolobova, K. A., Viola, B., Markin, S. V., Chintalapati, M., Peyrégne, S., Skov, L., Skoglund, P., Krivoshapkin, A. I., Derevianko, A. P., Meyer, M., Kelso, J., Peter, B., Prüfer, K., & Pääbo, S. (2020). A high-coverage Neandertal genome from Chagyrskaya Cave. *Proceedings of the National Academy of Sciences*, *117*(26), 15132–15136. https://doi.org/10.1073/pnas.2004944117

Maiti, A., Hou, S., Schiffer, C. A., & Matsuo, H. (2021). Interactions of APOBEC3s with DNA and RNA. *Current Opinion in Structural Biology*, *67*, 195–204. https://doi.org/10.1016/j.sbi.2020.12.004

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., Skoglund, P., Lazaridis, I., Sankararaman, S., Fu, Q., Rohland, N., Renaud, G., Erlich, Y., Willems, T., Gallo, C., … Reich, D. (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, *538*(7624), 201–206. https://doi.org/10.1038/nature18964

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., … Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*(7265), 747–753. https://doi.org/10.1038/nature08494

Mapstone, M., Cheema, A. K., Fiandaca, M. S., Zhong, X., Mhyre, T. R., MacArthur, L. H., Hall, W. J., Fisher, S. G., Peterson, D. R., Haley, J. M., Nazar, M. D., Rich, S. A., Berlau, D. J., Peltz, C. B., Tan, M. T., Kawas, C. H., & Federoff, H. J. (2014). Plasma phospholipids identify antecedent memory impairment in older adults. *Nature Medicine*, *20*(4), 415–418. https://doi.org/10.1038/nm.3466

Marchetto, M. C. N., Narvaiza, I., Denli, A. M., Benner, C., Lazzarini, T. A., Nathanson, J. L., Paquola, A. C. M., Desai, K. N., Herai, R. H., Weitzman, M. D., Yeo, G. W., Muotri, A. R., & Gage, F. H. (2013). Differential L1 regulation in pluripotent stem cells of humans and apes. *Nature*, *503*(7477), Article 7477. https://doi.org/10.1038/nature12686

Martinez, E., Palhan, V. B., Tjernberg, A., Lymar, E. S., Gamper, A. M., Kundu, T. K., Chait, B. T., & Roeder, R. G. (2001). Human STAGA complex is a chromatin-acetylating transcription coactivator that interacts with pre-mRNA splicing and DNA damage-binding factors in vivo. *Molecular and Cellular Biology*, *21*(20), 6782–6795. https://doi.org/10.1128/MCB.21.20.6782-6795.2001

Mas-Ponte, D., & Supek, F. (2020). DNA mismatch repair promotes APOBEC3-mediated diffuse hypermutation in human cancers. *Nature Genetics*, *52*(9), Article 9. https://doi.org/10.1038/s41588-020-0674-6

Massarat, A., Gymrek, M., McStay, B., & Jónsson, H. (2023). Human pangenome supports analysis of complex genomic regions. *Nature*, *617*(7960), 256–258. https://doi.org/10.1038/d41586-023-01490-3

Mathias, S. L., Scott, A. F., Kazazian, H. H., Boeke, J. D., & Gabriel, A. (1991). Reverse transcriptase encoded by a human transposable element. *Science (New York, N.Y.)*, *254*(5039), Article 5039. https://doi.org/10.1126/science.1722352

Matsuhashi, S., Noji, S., Koyama, E., Myokai, F., Ohuchi, H., Taniguchi, S., & Hori, K. (1995). New gene, nel, encoding a M(r) 93 K protein with EGF-like repeats is strongly expressed in neural tissues of

early stage chick embryos. *Developmental Dynamics: An Official Publication of the American Association of Anatomists*, *203*(2), 212–222. https://doi.org/10.1002/aja.1002030209

McCarthy, D. J., Chen, Y., & Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, *40*(10), 4288–4297. https://doi.org/10.1093/nar/gks042

McClintock, B. (1950). The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences*, *36*(6), Article 6. https://doi.org/10.1073/pnas.36.6.344

McGue, M., Zhang, Y., Miller, M. B., Basu, S., Vrieze, S., Hicks, B., Malone, S., Oetting, W. S., & Iacono, W. G. (2013). A genome-wide association study of behavioral disinhibition. *Behavior Genetics*, *43*(5), 363–373. https://doi.org/10.1007/s10519-013-9606-x

McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., Wenger, A. M., & Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology*, *28*(5), 495–501. https://doi.org/10.1038/nbt.1630

Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., Schraiber, J. G., Jay, F., Prüfer, K., de Filippo, C., Sudmant, P. H., Alkan, C., Fu, Q., Do, R., Rohland, N., Tandon, A., Siebauer, M., Green, R. E., Bryc, K., … Pääbo, S. (2012). A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science*, *338*(6104), 222–226. https://doi.org/10.1126/science.1224344

Mezzavilla, M., Cocca, M., Guidolin, F., & Gasparini, P. (2020). A population-based approach for gene prioritization in understanding complex traits. *Human Genetics*, *139*(5), 647–655. https://doi.org/10.1007/s00439-020-02152-4

Mignone, F., Gissi, C., Liuni, S., & Pesole, G. (2002). Untranslated regions of mRNAs. *Genome Biology*, *3*(3), REVIEWS0004. https://doi.org/10.1186/gb-2002-3-3-reviews0004

Mingione, A., Pivari, F., Plotegher, N., Dei Cas, M., Zulueta, A., Bocci, T., Trinchera, M., Albi, E., Maglione, V., Caretti, A., Bubacco, L., Paroni, R., Bottai, D., Ghidoni, R., & Signorelli, P. (2021). Inhibition of Ceramide Synthesis Reduces α-Synuclein Proteinopathy in a Cellular Model of Parkinson's Disease. *International Journal of Molecular Sciences*, *22*(12), 6469. https://doi.org/10.3390/ijms22126469

Mir, G., Meer, S., Cottrell, D., McMillan, D., House, A., & Kanter, J. W. (2015). Adapted behavioural activation for the treatment of depression in Muslims. *Journal of Affective Disorders*, *180*, 190–199. https://doi.org/10.1016/j.jad.2015.03.060

Misiak, B., Ricceri, L., & Sąsiadek, M. M. (2019). Transposable Elements and Their Epigenetic Regulation in Mental Disorders: Current Evidence in the Field. *Frontiers in Genetics*, *10*, 580. https://doi.org/10.3389/fgene.2019.00580

Modenini, G., Abondio, P., & Boattini, A. (2022). The coevolution between APOBEC3 and retrotransposons in primates. *Mobile DNA*, *13*(1), 27. https://doi.org/10.1186/s13100-022-00283-1

Modenini, G., Abondio, P., Guffanti, G., Boattini, A., & Macciardi, F. (2023). Evolutionarily recent retrotransposons contribute to schizophrenia. *Translational Psychiatry*, *13*(1), 181. https://doi.org/10.1038/s41398-023-02472-9

Modenini, G., Abondio, P., Sazzini, M., & Boattini, A. (2024). Polymorphic transposable elements provide new insights on high-altitude adaptation in the Tibetan Plateau. *Genomics*, *116*(3), 110854. https://doi.org/10.1016/j.ygeno.2024.110854

Moran, J. V., DeBerardinis, R. J., & Kazazian, H. H. (1999). Exon shuffling by L1 retrotransposition. *Science (New York, N.Y.)*, *283*(5407), Article 5407. https://doi.org/10.1126/science.283.5407.1530

Moreno-Mayar, J. V., Potter, B. A., Vinner, L., Steinrücken, M., Rasmussen, S., Terhorst, J., Kamm, J. A., Albrechtsen, A., Malaspinas, A.-S., Sikora, M., Reuther, J. D., Irish, J. D., Malhi, R. S., Orlando, L., Song, Y. S., Nielsen, R., Meltzer, D. J., & Willerslev, E. (2018). Terminal Pleistocene Alaskan genome reveals first founding population of Native Americans. *Nature*, *553*(7687), 203–207. https://doi.org/10.1038/nature25173

Moreno-Mayar, J. V., Vinner, L., De Barros Damgaard, P., De La Fuente, C., Chan, J., Spence, J. P., Allentoft, M. E., Vimala, T., Racimo, F., Pinotti, T., Rasmussen, S., Margaryan, A., Iraeta Orbegozo, M., Mylopotamitaki, D., Wooller, M., Bataille, C., Becerra-Valdivia, L., Chivall, D., Comeskey, D., … Willerslev, E. (2018). Early human dispersals within the Americas. *Science*, *362*(6419), eaav2621. https://doi.org/10.1126/science.aav2621

Muckenfuss, H., Hamdorf, M., Held, U., Perković, M., Löwer, J., Cichutek, K., Flory, E., Schumann, G. G., & Münk, C. (2006). APOBEC3 Proteins Inhibit Human LINE-1 Retrotransposition. *Journal of Biological Chemistry*, *281*(31), Article 31. https://doi.org/10.1074/jbc.M601716200

Münk, C., Beck, T., Zielonka, J., Hotz-Wagenblatt, A., Chareza, S., Battenberg, M., Thielebein, J., Cichutek, K., Bravo, I. G., O'Brien, S. J., Löchelt, M., & Yuhki, N. (2008). Functions, structure, and read-through alternative splicing of feline APOBEC3 genes. *Genome Biology*, *9*(3), Article 3. https://doi.org/10.1186/gb-2008-9-3-r48

Münk, C., Willemsen, A., & Bravo, I. G. (2012). An ancient history of gene duplications, fusions and losses in the evolution of APOBEC3 mutators in mammals. *BMC Evolutionary Biology*, *12*, 71. https://doi.org/10.1186/1471-2148-12-71

Munk, R., Anerillas, C., Rossi, M., Tsitsipatis, D., Martindale, J. L., Herman, A. B., Yang, J.-H., Roberts, J. A., Varma, V. R., Pandey, P. R., Thambisetty, M., Gorospe, M., & Abdelmohsen, K. (2021). Acid ceramidase promotes senescent cell survival. *Aging*, *13*(12), 15750–15769. https://doi.org/10.18632/aging.203170

Muotri, A. R., Chu, V. T., Marchetto, M. C. N., Deng, W., Moran, J. V., & Gage, F. H. (2005). Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature*, *435*(7044), Article 7044. https://doi.org/10.1038/nature03663

Nakamura, Y., Murata, M., Takagi, Y., Kozuka, T., Nakata, Y., Hasebe, R., Takagi, A., Kitazawa, J., Shima, M., & Kojima, T. (2015). SVA retrotransposition in exon 6 of the coagulation factor IX gene causing severe hemophilia B. *International Journal of Hematology*, *102*(1), 134–139. https://doi.org/10.1007/s12185-015-1765-5

Narvaiza, I., Linfesty, D. C., Greener, B. N., Hakata, Y., Pintel, D. J., Logue, E., Landau, N. R., & Weitzman, M. D. (2009). Deaminase-independent inhibition of parvoviruses by the APOBEC3A cytidine deaminase. *PLoS Pathogens*, *5*(5), Article 5. https://doi.org/10.1371/journal.ppat.1000439

Neph, S., Kuehn, M. S., Reynolds, A. P., Haugen, E., Thurman, R. E., Johnson, A. K., Rynes, E., Maurano, M. T., Vierstra, J., Thomas, S., Sandstrom, R., Humbert, R., & Stamatoyannopoulos, J. A. (2012). BEDOPS: High-performance genomic feature operations. *Bioinformatics (Oxford, England)*, *28*(14), 1919–1920. https://doi.org/10.1093/bioinformatics/bts277

Nibbs, R. J., Wylie, S. M., Yang, J., Landau, N. R., & Graham, G. J. (1997). Cloning and characterization of a novel promiscuous human beta-chemokine receptor D6. *The Journal of Biological Chemistry*, *272*(51), 32078–32083. https://doi.org/10.1074/jbc.272.51.32078

Nogimori, K., Hori, T., Kawaguchi, K., Fukui, T., Mii, S., Nakada, H., Matsumoto, Y., Yamauchi, Y., Takahashi, M., Furukawa, K., Tetsuya, O., Yokoi, K., Hasegawa, Y., & Furukawa, K. (2016). Increased expression levels of ppGalNAc-T13 in lung cancers: Significance in the prognostic diagnosis. *International Journal of Oncology*, *49*(4), 1369–1376. https://doi.org/10.3892/ijo.2016.3638

Notwell, J. H., Chung, T., Heavner, W., & Bejerano, G. (2015). A family of transposable elements co-opted into developmental enhancers in the mouse neocortex. *Nature Communications*, *6*, 6644. https://doi.org/10.1038/ncomms7644

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., … Phillippy, A. M. (2022). The complete sequence of a human genome. *Science (New York, N.Y.)*, *376*(6588), 44–53. https://doi.org/10.1126/science.abj6987

Ochoa Thomas, E., Zuniga, G., Sun, W., & Frost, B. (2020). Awakening the dark side: Retrotransposon activation in neurodegenerative disorders. *Current Opinion in Neurobiology*, *61*, 65–72. https://doi.org/10.1016/j.conb.2020.01.012

Ohshima, K., Hattori, M., Yada, T., Gojobori, T., Sakaki, Y., & Okada, N. (2003). Whole-genome screening indicates a possible burst of formation of processed pseudogenes and Alu repeats by particular L1 subfamilies in ancestral primates. *Genome Biology*, *4*(11), Article 11. https://doi.org/10.1186/gb-2003-4-11-r74

Okada, A., & Iwatani, Y. (2016). APOBEC3G-Mediated G-to-A Hypermutation of the HIV-1 Genome: The Missing Link in Antiviral Molecular Mechanisms. *Frontiers in Microbiology*, *7*, 2027. https://doi.org/10.3389/fmicb.2016.02027

Oksenberg, N., Stevison, L., Wall, J. D., & Ahituv, N. (2013). Function and regulation of AUTS2, a gene implicated in autism and human evolution. *PLoS Genetics*, *9*(1), e1003221. https://doi.org/10.1371/journal.pgen.1003221

Oliver, K. R., & Greene, W. K. (2009). Transposable elements: Powerful facilitators of evolution. *BioEssays*, *31*(7), 703–714. https://doi.org/10.1002/bies.200800219

Orlando, L., Allaby, R., Skoglund, P., Der Sarkissian, C., Stockhammer, P. W., Ávila-Arcos, M. C., Fu, Q., Krause, J., Willerslev, E., Stone, A. C., & Warinner, C. (2021). Ancient DNA analysis. *Nature Reviews Methods Primers*, *1*(1), 14. https://doi.org/10.1038/s43586-020-00011-0

Otowa, T., Maher, B. S., Aggen, S. H., McClay, J. L., van den Oord, E. J., & Hettema, J. M. (2014). Genome-wide and gene-based association studies of anxiety disorders in European and African American samples. *PloS One*, *9*(11), e112559. https://doi.org/10.1371/journal.pone.0112559

Oyasu, M., Kuroda, S., Nakashita, M., Fujimiya, M., Kikkawa, U., & Saito, N. (2000). Immunocytochemical localization of a neuron-specific thrombospondin-1-like protein, NELL2: Light and electron microscopic studies in the rat brain. *Brain Research. Molecular Brain Research*, *76*(1), 151–160. https://doi.org/10.1016/s0169-328x(99)00342-3

Ozsoy, F., Karakus, N. B., Yigit, S., & Kulu, M. (2020). Effect of AUTS2 gene rs6943555 variant in male patients with schizophrenia in a Turkish population. *Gene*, *756*, 144913. https://doi.org/10.1016/j.gene.2020.144913

Pantazopoulos, H., Katsel, P., Haroutunian, V., Chelini, G., Klengel, T., & Berretta, S. (2021). Molecular signature of extracellular matrix pathology in schizophrenia. *The European Journal of Neuroscience*, *53*(12), 3960–3987. https://doi.org/10.1111/ejn.15009

Park, L. (2011). Effective population size of current human population. *Genetics Research*, *93*(2), Article 2. https://doi.org/10.1017/S0016672310000558

Payer, L. M., & Burns, K. H. (2019). Transposable elements in human genetic disease. *Nature Reviews. Genetics*, *20*(12), 760–772. https://doi.org/10.1038/s41576-019-0165-8

Payer, L. M., Steranka, J. P., Kryatova, M. S., Grillo, G., Lupien, M., Rocha, P. P., & Burns, K. H. (2021). *Alu* insertion variants alter gene transcript levels. *Genome Research*, *31*(12), 2236–2248. https://doi.org/10.1101/gr.261305.120

Payer, L. M., Steranka, J. P., Yang, W. R., Kryatova, M., Medabalimi, S., Ardeljan, D., Liu, C., Boeke, J. D., Avramopoulos, D., & Burns, K. H. (2017). Structural variants caused by *Alu* insertions are associated with risks for many human diseases. *Proceedings of the National Academy of Sciences*, *114*(20). https://doi.org/10.1073/pnas.1704117114

Pehrsson, E. C., Choudhary, M. N. K., Sundaram, V., & Wang, T. (2019). The epigenomic landscape of transposable elements across normal human development and anatomy. *Nature Communications*, *10*(1), 5640. https://doi.org/10.1038/s41467-019-13555-x

Pena, E., El Alam, S., Siques, P., & Brito, J. (2022). Oxidative Stress and Diseases Associated with High-Altitude Exposure. *Antioxidants (Basel, Switzerland)*, *11*(2), 267. https://doi.org/10.3390/antiox11020267

Percharde, M., Lin, C.-J., Yin, Y., Guan, J., Peixoto, G. A., Bulut-Karslioglu, A., Biechele, S., Huang, B., Shen, X., & Ramalho-Santos, M. (2018). A LINE1-Nucleolin Partnership Regulates Early Development and ESC Identity. *Cell*, *174*(2), Article 2. https://doi.org/10.1016/j.cell.2018.05.043

Perkins, J. M., Subramanian, S. V., Davey Smith, G., & Özaltin, E. (2016). Adult height, nutrition, and population health. *Nutrition Reviews*, *74*(3), 149–165. https://doi.org/10.1093/nutrit/nuv105

Perna, N. T., Batzer, M. A., Deininger, P. L., & Stoneking, M. (1992). Alu insertion polymorphism: A new type of marker for human population studies. *Human Biology*, *64*(5), 641–648.

Pertea, M., Kim, D., Pertea, G. M., Leek, J. T., & Salzberg, S. L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature Protocols*, *11*(9), 1650–1667. https://doi.org/10.1038/nprot.2016.095

Pesaresi, S., Galdenzi, D., Biondi, E., & Casavecchia, S. (2014). Bioclimate of Italy: Application of the worldwide bioclimatic classification system. *Journal of Maps*, *10*(4), 538–553. https://doi.org/10.1080/17445647.2014.891472

Petljak, M., Dananberg, A., Chu, K., Bergstrom, E. N., Striepen, J., von Morgen, P., Chen, Y., Shah, H., Sale, J. E., Alexandrov, L. B., Stratton, M. R., & Maciejowski, J. (2022). Mechanisms of APOBEC3 mutagenesis in human cancer cells. *Nature*, *607*(7920), Article 7920. https://doi.org/10.1038/s41586-022-04972-y

Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S., & Goldstein, D. B. (2013). Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genetics*, *9*(8), e1003709. https://doi.org/10.1371/journal.pgen.1003709

Peze-Heidsieck, E., Bonnifet, T., Znaidi, R., Ravel-Godreuil, C., Massiani-Beaudoin, O., Joshi, R. L., & Fuchs, J. (2022). Retrotransposons as a Source of DNA Damage in Neurodegeneration. *Frontiers in Aging Neuroscience*, *13*, 786897. https://doi.org/10.3389/fnagi.2021.786897

Piras, I. S., De Montis, A., Calò, C. M., Marini, M., Atzori, M., Corrias, L., Sazzini, M., Boattini, A., Vona, G., & Contu, L. (2012). Genome-wide scan with nearly 700 000 SNPs in two Sardinian sub-populations suggests some regions as candidate targets for positive selection. *European Journal of Human Genetics*, *20*(11), 1155–1161. https://doi.org/10.1038/ejhg.2012.65

Platt, R. N., Vandewege, M. W., & Ray, D. A. (2018). Mammalian transposable elements and their impacts on genome evolution. *Chromosome Research: An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology*, *26*(1–2), 25–43. https://doi.org/10.1007/s10577-017-9570-z

Platzer, K., & Lemke, J. R. (1993). GRIN2B-Related Neurodevelopmental Disorder. In M. P. Adam, J. Feldman, G. M. Mirzaa, R. A. Pagon, S. E. Wallace, L. J. Bean, K. W. Gripp, & A. Amemiya (Eds.), *GeneReviews®*. University of Washington, Seattle. http://www.ncbi.nlm.nih.gov/books/NBK501979/

Polimanti, R., & Gelernter, J. (2017). Widespread signatures of positive selection in common risk alleles associated to autism spectrum disorder. *PLOS Genetics*, *13*(2), e1006618. https://doi.org/10.1371/journal.pgen.1006618

Pollard, K. S., Salama, S. R., King, B., Kern, A. D., Dreszer, T., Katzman, S., Siepel, A., Pedersen, J. S., Bejerano, G., Baertsch, R., Rosenbloom, K. R., Kent, J., & Haussler, D. (2006). Forces shaping the fastest evolving regions in the human genome. *PLoS Genetics*, *2*(10), e168. https://doi.org/10.1371/journal.pgen.0020168

Pollard, K. S., Salama, S. R., Lambert, N., Lambot, M.-A., Coppens, S., Pedersen, J. S., Katzman, S., King, B., Onodera, C., Siepel, A., Kern, A. D., Dehay, C., Igel, H., Ares, M., Vanderhaeghen, P., & Haussler, D. (2006). An RNA gene expressed during cortical development evolved rapidly in humans. *Nature*, *443*(7108), 167–172. https://doi.org/10.1038/nature05113

Ponomarev, I., Rau, V., Eger, E. I., Harris, R. A., & Fanselow, M. S. (2010). Amygdala transcriptome and cellular mechanisms underlying stress-enhanced fear learning in a rat model of posttraumatic stress disorder. *Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology*, *35*(6), 1402–1411. https://doi.org/10.1038/npp.2010.10

Pontis, J., Planet, E., Offner, S., Turelli, P., Duc, J., Coudray, A., Theunissen, T. W., Jaenisch, R., & Trono, D. (2019). Hominoid-Specific Transposable Elements and KZFPs Facilitate Human Embryonic Genome Activation and Control Transcription in Naive Human ESCs. *Cell Stem Cell*, *24*(5), 724-735.e5. https://doi.org/10.1016/j.stem.2019.03.012

Posth, C., Wißing, C., Kitagawa, K., Pagani, L., Van Holstein, L., Racimo, F., Wehrberger, K., Conard, N. J., Kind, C. J., Bocherens, H., & Krause, J. (2017). Deeply divergent archaic mitochondrial genome provides lower time boundary for African gene flow into Neanderthals. *Nature Communications*, *8*(1), 16046. https://doi.org/10.1038/ncomms16046

Prabhakar, S., Noonan, J. P., Pääbo, S., & Rubin, E. M. (2006). Accelerated evolution of conserved noncoding sequences in humans. *Science (New York, N.Y.)*, *314*(5800), 786. https://doi.org/10.1126/science.1130738

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904–909. https://doi.org/10.1038/ng1847

Pritham, E. J., Putliwala, T., & Feschotte, C. (2007). Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. *Gene*, *390*(1–2), Article 1–2. https://doi.org/10.1016/j.gene.2006.08.008

Protasova, M. S., Andreeva, T. V., & Rogaev, E. I. (2021). Factors Regulating the Activity of LINE1 Retrotransposons. *Genes*, *12*(10), Article 10. https://doi.org/10.3390/genes12101562

Protasova, M. S., Gusev, F. E., Grigorenko, A. P., Kuznetsova, I. L., Rogaev, E. I., & Andreeva, T. V. (2017). Quantitative Analysis of L1-Retrotransposons in Alzheimer's Disease and Aging. *Biochemistry. Biokhimiia*, *82*(8), 962–971. https://doi.org/10.1134/S0006297917080120

Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P. H., de Filippo, C., Li, H., Mallick, S., Dannemann, M., Fu, Q., Kircher, M., Kuhlwilm,

M., Lachmann, M., Meyer, M., Ongyerth, M., … Pääbo, S. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, *505*(7481), 43–49. https://doi.org/10.1038/nature12886

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Qiu, X., Hill, A., Packer, J., Lin, D., Ma, Y.-A., & Trapnell, C. (2017). Single-cell mRNA quantification and differential analysis with Census. *Nature Methods*, *14*(3), 309–315. https://doi.org/10.1038/nmeth.4150

Quan, C., Li, Y., Liu, X., Wang, Y., Ping, J., Lu, Y., & Zhou, G. (2021). Characterization of structural variation in Tibetans reveals new evidence of high-altitude adaptation and introgression. *Genome Biology*, *22*(1), 159. https://doi.org/10.1186/s13059-021-02382-3

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Racimo, F., Sankararaman, S., Nielsen, R., & Huerta-Sánchez, E. (2015). Evidence for archaic adaptive introgression in humans. *Nature Reviews. Genetics*, *16*(6), 359–371. https://doi.org/10.1038/nrg3936

Raghavan, S., Huang, J., Tcheandjieu, C., Huffman, J. E., Litkowski, E., Liu, C., Ho, Y.-L. A., Hunter-Zinck, H., Zhao, H., Marouli, E., North, K. E., the VA Million Veteran Program, Lange, E., Lange, L. A., Voight, B. F., Gaziano, J. M., Pyarajan, S., Hauser, E. R., Tsao, P. S., … Assimes, T. L. (2022). A multi-population phenome-wide association study of genetically-predicted height in the Million Veteran Program. *PLOS Genetics*, *18*(6), e1010193. https://doi.org/10.1371/journal.pgen.1010193

Ramirez, G., Bittle, P. A., Colice, G. L., Herrera, R., Agosti, S. J., & Foulis, P. R. (1991). The effect of cigarette smoking upon hematological adaptations to moderately high altitude living. *Journal of Wilderness Medicine*, *2*(4), 274–286. https://doi.org/10.1580/0953-9859-2.4.274

Rangwala, S. H., Zhang, L., & Kazazian, H. H. (2009). Many LINE1 elements contribute to the transcriptome of human somatic cells. *Genome Biology*, *10*(9), R100. https://doi.org/10.1186/gb-2009-10-9-r100

Rasmussen, M., Li, Y., Lindgreen, S., Pedersen, J. S., Albrechtsen, A., Moltke, I., Metspalu, M., Metspalu, E., Kivisild, T., Gupta, R., Bertalan, M., Nielsen, K., Gilbert, M. T. P., Wang, Y., Raghavan, M., Campos, P. F., Kamp, H. M., Wilson, A. S., Gledhill, A., … Willerslev, E. (2010). Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature*, *463*(7282), 757–762. https://doi.org/10.1038/nature08835

Ratcliff, J., & Simmonds, P. (2021). Potential APOBEC-mediated RNA editing of the genomes of SARS-CoV-2 and other coronaviruses and its impact on their longer term evolution. *Virology*, *556*, 62–72. https://doi.org/10.1016/j.virol.2020.12.018

Rau, V., & Fanselow, M. S. (2009). Exposure to a stressor produces a long lasting enhancement of fear learning in rats. *Stress (Amsterdam, Netherlands)*, *12*(2), 125–133. https://doi.org/10.1080/10253890802137320

Ravel‐Godreuil, C., Znaidi, R., Bonnifet, T., Joshi, R. L., & Fuchs, J. (2021). Transposable elements as new players in neurodegenerative diseases. *FEBS Letters*, *595*(22), 2733–2755. https://doi.org/10.1002/1873-3468.14205

Ray, D. A., & Batzer, M. A. (2011). Reading TE leaves: New approaches to the identification of transposable element insertions. *Genome Research*, *21*(6), 813–820. https://doi.org/10.1101/gr.110528.110

Ray, D. A., Xing, J., Salem, A.-H., & Batzer, M. A. (2006). SINEs of a nearly perfect character. *Systematic Biology*, *55*(6), 928–935. https://doi.org/10.1080/10635150600865419

Reich, D., Green, R. E., Kircher, M., Krause, J., Patterson, N., Durand, E. Y., Viola, B., Briggs, A. W., Stenzel, U., Johnson, P. L. F., Maricic, T., Good, J. M., Marques-Bonet, T., Alkan, C., Fu, Q., Mallick, S., Li, H., Meyer, M., Eichler, E. E., … Pääbo, S. (2010). Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, *468*(7327), 1053–1060. https://doi.org/10.1038/nature09710

Reich, D., Patterson, N., Kircher, M., Delfin, F., Nandineni, M. R., Pugach, I., Ko, A. M.-S., Ko, Y.-C., Jinam, T. A., Phipps, M. E., Saitou, N., Wollstein, A., Kayser, M., Pääbo, S., & Stoneking, M. (2011).

Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. *American Journal of Human Genetics*, *89*(4), 516–528. https://doi.org/10.1016/j.ajhg.2011.09.005

Reilly, M. T., Faulkner, G. J., Dubnau, J., Ponomarev, I., & Gage, F. H. (2013). The role of transposable elements in health and diseases of the central nervous system. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *33*(45), 17577–17586. https://doi.org/10.1523/JNEUROSCI.3369-13.2013

Reilly, S. K., Yin, J., Ayoub, A. E., Emera, D., Leng, J., Cotney, J., Sarro, R., Rakic, P., & Noonan, J. P. (2015). Evolutionary genomics. Evolutionary changes in promoter and enhancer activity during human corticogenesis. *Science (New York, N.Y.)*, *347*(6226), 1155–1159. https://doi.org/10.1126/science.1260943

Revathidevi, S., Murugan, A. K., Nakaoka, H., Inoue, I., & Munirajan, A. K. (2021). APOBEC: A molecular driver in cervical cancer pathogenesis. *Cancer Letters*, *496*, 104–116. https://doi.org/10.1016/j.canlet.2020.10.004

Ricci, M., Peona, V., Guichard, E., Taccioli, C., & Boattini, A. (2018). Transposable Elements Activity is Positively Related to Rate of Speciation in Mammals. *Journal of Molecular Evolution*, *86*(5), Article 5. https://doi.org/10.1007/s00239-018-9847-7

Richardson, S. R., Doucet, A. J., Kopera, H. C., Moldovan, J. B., Garcia-Perez, J. L., & Moran, J. V. (2015). The Influence of LINE-1 and SINE Retrotransposons on Mammalian Genomes. *Microbiology Spectrum*, *3*(2), Article 2. https://doi.org/10.1128/microbiolspec.MDNA3-0061-2014

Richardson, S. R., Narvaiza, I., Planegger, R. A., Weitzman, M. D., & Moran, J. V. (2014). APOBEC3A deaminates transiently exposed single-strand DNA during LINE-1 retrotransposition. *ELife*, *3*, e02008. https://doi.org/10.7554/eLife.02008

Rishishwar, L., Mariño-Ramírez, L., & Jordan, I. K. (2017). Benchmarking computational tools for polymorphic transposable element detection. *Briefings in Bioinformatics*, *18*(6), Article 6. https://doi.org/10.1093/bib/bbw072

Rishishwar, L., Tellez Villa, C. E., & Jordan, I. K. (2015). Transposable element polymorphisms recapitulate human evolution. *Mobile DNA*, *6*, 21. https://doi.org/10.1186/s13100-015-0052-6

Rishishwar, L., Wang, L., Clayton, E. A., Mariño-Ramírez, L., McDonald, J. F., & Jordan, I. K. (2017). Population and clinical genetics of human transposable elements in the (post) genomic era. *Mobile Genetic Elements*, *7*(1), Article 1. https://doi.org/10.1080/2159256X.2017.1280116

Rishishwar, L., Wang, L., Wang, J., Yi, S. V., Lachance, J., & Jordan, I. K. (2018). Evidence for positive selection on recent human transposable element insertions. *Gene*, *675*, 69–79. https://doi.org/10.1016/j.gene.2018.06.077

Roberts, E. (2006). GABAergic malfunction in the limbic system resulting from an aboriginal genetic defect in voltage-gated Na+-channel SCN5A is proposed to give rise to susceptibility to schizophrenia. *Advances in Pharmacology (San Diego, Calif.)*, *54*, 119–145. https://doi.org/10.1016/s1054-3589(06)54006-2

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*, 77. https://doi.org/10.1186/1471-2105-12-77

Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, *11*(3), R25. https://doi.org/10.1186/gb-2010-11-3-r25

Roden, D. M. (2014). The Brugada ECG and Schizophrenia. *Circulation: Arrhythmia and Electrophysiology*, *7*(3), 365–367. https://doi.org/10.1161/CIRCEP.114.001641

Rogozin, I. B., Iyer, L. M., Liang, L., Glazko, G. V., Liston, V. G., Pavlov, Y. I., Aravind, L., & Pancer, Z. (2007). Evolution and diversification of lamprey antigen receptors: Evidence for involvement of an AID-APOBEC family cytosine deaminase. *Nature Immunology*, *8*(6), Article 6. https://doi.org/10.1038/ni1463

Rosenstock, E., Ebert, J., Martin, R., Hicketier, A., Walter, P., & Groß, M. (2019). Human stature in the Near East and Europe ca. 10,000–1000 BC: Its spatiotemporal development in a Bayesian errors-in-variables model. *Archaeological and Anthropological Sciences*, *11*(10), 5657–5690. https://doi.org/10.1007/s12520-019-00850-3

Rotival, M., & Quintana-Murci, L. (2020). Functional consequences of archaic introgression and their impact on fitness. *Genome Biology*, *21*(1), 3, s13059-019-1920-z. https://doi.org/10.1186/s13059-019-1920-z

Sadeghpour, S., Khodaee, S., Rahnama, M., Rahimi, H., & Ebrahimi, D. (2021). Human APOBEC3 Variations and Viral Infection. *Viruses*, *13*(7), Article 7. https://doi.org/10.3390/v13071366

Sakaue, S., Kanai, M., Tanigawa, Y., Karjalainen, J., Kurki, M., Koshiba, S., Narita, A., Konuma, T., Yamamoto, K., Akiyama, M., Ishigaki, K., Suzuki, A., Suzuki, K., Obara, W., Yamaji, K., Takahashi, K., Asai, S., Takahashi, Y., Suzuki, T., … Okada, Y. (2021). A cross-population atlas of genetic associations for 220 human phenotypes. *Nature Genetics*, *53*(10), 1415–1424. https://doi.org/10.1038/s41588-021-00931-x

Salter, J. D., Bennett, R. P., & Smith, H. C. (2016). The APOBEC Protein Family: United by Structure, Divergent in Function. *Trends in Biochemical Sciences*, *41*(7), Article 7. https://doi.org/10.1016/j.tibs.2016.05.001

Sánchez-Quinto, F., Malmström, H., Fraser, M., Girdland-Flink, L., Svensson, E. M., Simões, L. G., George, R., Hollfelder, N., Burenhult, G., Noble, G., Britton, K., Talamo, S., Curtis, N., Brzobohata, H., Sumberova, R., Götherström, A., Storå, J., & Jakobsson, M. (2019). Megalithic tombs in western and northern Neolithic Europe were linked to a kindred society. *Proceedings of the National Academy of Sciences*, *116*(19), 9469–9474. https://doi.org/10.1073/pnas.1818037116

Sánchez-Valle, J., Tejero, H., Ibáñez, K., Portero, J. L., Krallinger, M., Al-Shahrour, F., Tabarés-Seisdedos, R., Baudot, A., & Valencia, A. (2017). A molecular hypothesis to explain direct and inverse co-morbidities between Alzheimer's Disease, Glioblastoma and Lung cancer. *Scientific Reports*, *7*(1), 4474. https://doi.org/10.1038/s41598-017-04400-6

Santoni, F. A., Guerra, J., & Luban, J. (2012). HERV-H RNA is abundant in human embryonic stem cells and a precise marker for pluripotency. *Retrovirology*, *9*, 111. https://doi.org/10.1186/1742-4690-9-111

Saunders, G. R. B., Wang, X., Chen, F., Jang, S.-K., Liu, M., Wang, C., Gao, S., Jiang, Y., Khunsriraksakul, C., Otto, J. M., Addison, C., Akiyama, M., Albert, C. M., Aliev, F., Alonso, A., Arnett, D. K., Ashley-Koch, A. E., Ashrani, A. A., Barnes, K. C., … Vrieze, S. (2022). Genetic diversity fuels gene discovery for tobacco and alcohol use. *Nature*, *612*(7941), 720–724. https://doi.org/10.1038/s41586-022-05477-4

Sawyer, S. L., Emerman, M., & Malik, H. S. (2004). Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G. *PLoS Biology*, *2*(9), Article 9. https://doi.org/10.1371/journal.pbio.0020275

Sayah, D. M., Sokolskaja, E., Berthoux, L., & Luban, J. (2004). Cyclophilin A retrotransposition into TRIM5 explains owl monkey resistance to HIV-1. *Nature*, *430*(6999), Article 6999. https://doi.org/10.1038/nature02777

Sazzini, M., Gnecchi Ruscone, G. A., Giuliani, C., Sarno, S., Quagliariello, A., De Fanti, S., Boattini, A., Gentilini, D., Fiorito, G., Catanoso, M., Boiardi, L., Croci, S., Macchioni, P., Mantovani, V., Di Blasio, A. M., Matullo, G., Salvarani, C., Franceschi, C., Pettener, D., … Luiselli, D. (2016). Complex interplay between neutral and adaptive evolution shaped differential genomic background and disease susceptibility along the Italian peninsula. *Scientific Reports*, *6*, 32513. https://doi.org/10.1038/srep32513

Schiffels, S., Haak, W., Paajanen, P., Llamas, B., Popescu, E., Loe, L., Clarke, R., Lyons, A., Mortimer, R., Sayer, D., Tyler-Smith, C., Cooper, A., & Durbin, R. (2016). Iron Age and Anglo-Saxon genomes from East England reveal British migration history. *Nature Communications*, *7*(1), 10408. https://doi.org/10.1038/ncomms10408

Schizophrenia Working Group of the Psychiatric Genomics Consortium. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, *511*(7510), 421–427. https://doi.org/10.1038/nature13595

Schlebusch, C. M., Malmström, H., Günther, T., Sjödin, P., Coutinho, A., Edlund, H., Munters, A. R., Vicente, M., Steyn, M., Soodyall, H., Lombard, M., & Jakobsson, M. (2017). Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. *Science (New York, N.Y.)*, *358*(6363), 652–655. https://doi.org/10.1126/science.aao6266

Schoeler, T., Speed, D., Porcu, E., Pirastu, N., Pingault, J.-B., & Kutalik, Z. (2023). Participation bias in the UK Biobank distorts genetic associations and downstream analyses. *Nature Human Behaviour*, *7*(7), 1216–1227. https://doi.org/10.1038/s41562-023-01579-9

Schumann, G., Coin, L. J., Lourdusamy, A., Charoen, P., Berger, K. H., Stacey, D., Desrivières, S., Aliev, F. A., Khan, A. A., Amin, N., Aulchenko, Y. S., Bakalkin, G., Bakker, S. J., Balkau, B., Beulens, J. W.,

Bilbao, A., de Boer, R. A., Beury, D., Bots, M. L., … Elliott, P. (2011). Genome-wide association and genetic functional studies identify autism susceptibility candidate 2 gene (AUTS2) in the regulation of alcohol consumption. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(17), 7119–7124. https://doi.org/10.1073/pnas.1017288108

Scott, A. F., Schmeckpeper, B. J., Abdelrazik, M., Comey, C. T., O'Hara, B., Rossiter, J. P., Cooley, T., Heath, P., Smith, K. D., & Margolet, L. (1987). Origin of the human L1 elements: Proposed progenitor genes deduced from a consensus DNA sequence. *Genomics*, *1*(2), Article 2. https://doi.org/10.1016/0888-7543(87)90003-6

Sedivy, J. M., Kreiling, J. A., Neretti, N., De Cecco, M., Criscione, S. W., Hofmann, J. W., Zhao, X., Ito, T., & Peterson, A. L. (2013). Death by transposition—The enemy within? *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, *35*(12), 1035–1043. https://doi.org/10.1002/bies.201300097

Seguin-Orlando, A., Korneliussen, T. S., Sikora, M., Malaspinas, A.-S., Manica, A., Moltke, I., Albrechtsen, A., Ko, A., Margaryan, A., Moiseyev, V., Goebel, T., Westaway, M., Lambert, D., Khartanovich, V., Wall, J. D., Nigst, P. R., Foley, R. A., Lahr, M. M., Nielsen, R., … Willerslev, E. (2014). Genomic structure in Europeans dating back at least 36,200 years. *Science*, *346*(6213), 1113–1118. https://doi.org/10.1126/science.aaa0114

Sehgal, D., Mondal, S., Crespo-Herrera, L., Velu, G., Juliana, P., Huerta-Espino, J., Shrestha, S., Poland, J., Singh, R., & Dreisigacker, S. (2020). Haplotype-Based, Genome-Wide Association Study Reveals Stable Genomic Regions for Grain Yield in CIMMYT Spring Bread Wheat. *Frontiers in Genetics*, *11*, 589490. https://doi.org/10.3389/fgene.2020.589490

Sekar, A., Bialas, A. R., de Rivera, H., Davis, A., Hammond, T. R., Kamitaki, N., Tooley, K., Presumey, J., Baum, M., Van Doren, V., Genovese, G., Rose, S. A., Handsaker, R. E., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Daly, M. J., Carroll, M. C., Stevens, B., & McCarroll, S. A. (2016). Schizophrenia risk from complex variation of complement component 4. *Nature*, *530*(7589), 177–183. https://doi.org/10.1038/nature16549

Setter, D., Mousset, S., Cheng, X., Nielsen, R., DeGiorgio, M., & Hermisson, J. (2020). VolcanoFinder: Genomic scans for adaptive introgression. *PLOS Genetics*, *16*(6), e1008867. https://doi.org/10.1371/journal.pgen.1008867

Sheehy, A. M., Gaddis, N. C., Choi, J. D., & Malim, M. H. (2002). Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. *Nature*, *418*(6898), Article 6898. https://doi.org/10.1038/nature00939

Shen, M. R., Batzer, M. A., & Deininger, P. L. (1991). Evolution of the master Alu gene(s). *Journal of Molecular Evolution*, *33*(4), Article 4. https://doi.org/10.1007/BF02102862

Sherman, B. T., Hao, M., Qiu, J., Jiao, X., Baseler, M. W., Lane, H. C., Imamichi, T., & Chang, W. (2022). DAVID: A web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Research*, *50*(W1), W216–W221. https://doi.org/10.1093/nar/gkac194

Sherva, R., Zhu, C., Wetherill, L., Edenberg, H. J., Johnson, E., Degenhardt, L., Agrawal, A., Martin, N. G., Nelson, E., Kranzler, H. R., Gelernter, J., & Farrer, L. A. (2021). Genome-wide association study of phenotypes measuring progression from first cocaine or opioid use to dependence reveals novel risk genes. *Exploration of Medicine*, *2*, 60–73. https://doi.org/10.37349/emed.2021.00032

Sies, H., & Jones, D. P. (2020). Reactive oxygen species (ROS) as pleiotropic physiological signalling agents. *Nature Reviews. Molecular Cell Biology*, *21*(7), 363–383. https://doi.org/10.1038/s41580-020-0230-3

Sikela, J. M., & Searles Quick, V. B. (2018). Genomic trade-offs: Are autism and schizophrenia the steep price of the human brain? *Human Genetics*, *137*(1), 1–13. https://doi.org/10.1007/s00439-017-1865-9

Sikora, M., Seguin-Orlando, A., Sousa, V. C., Albrechtsen, A., Korneliussen, T., Ko, A., Rasmussen, S., Dupanloup, I., Nigst, P. R., Bosch, M. D., Renaud, G., Allentoft, M. E., Margaryan, A., Vasilyev, S. V., Veselovskaya, E. V., Borutskaya, S. B., Deviese, T., Comeskey, D., Higham, T., … Willerslev, E. (2017). Ancient genomes show social and reproductive behavior of early Upper Paleolithic foragers. *Science (New York, N.Y.)*, *358*(6363), 659–662. https://doi.org/10.1126/science.aao1807

Simon, J. H., Miller, D. L., Fouchier, R. A., Soares, M. A., Peden, K. W., & Malim, M. H. (1998). The regulation of primate immunodeficiency virus infectivity by Vif is cell species restricted: A role for

Vif in determining virus host range and cross-species transmission. *The EMBO Journal*, *17*(5), Article 5. https://doi.org/10.1093/emboj/17.5.1259

Simonson, T. S., Yang, Y., Huff, C. D., Yun, H., Qin, G., Witherspoon, D. J., Bai, Z., Lorenzo, F. R., Xing, J., Jorde, L. B., Prchal, J. T., & Ge, R. (2010). Genetic evidence for high-altitude adaptation in Tibet. *Science (New York, N.Y.)*, *329*(5987), 72–75. https://doi.org/10.1126/science.1189406

Siriwardena, S. U., Chen, K., & Bhagwat, A. S. (2016). Functions and Malfunctions of Mammalian DNA-Cytosine Deaminases. *Chemical Reviews*, *116*(20), Article 20. https://doi.org/10.1021/acs.chemrev.6b00296

Sjöstedt, E., Zhong, W., Fagerberg, L., Karlsson, M., Mitsios, N., Adori, C., Oksvold, P., Edfors, F., Limiszewska, A., Hikmet, F., Huang, J., Du, Y., Lin, L., Dong, Z., Yang, L., Liu, X., Jiang, H., Xu, X., Wang, J., … Mulder, J. (2020). An atlas of the protein-coding genes in the human, pig, and mouse brain. *Science (New York, N.Y.)*, *367*(6482), eaay5947. https://doi.org/10.1126/science.aay5947

Smit, A. F., Tóth, G., Riggs, A. D., & Jurka, J. (1995). Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. *Journal of Molecular Biology*, *246*(3), Article 3. https://doi.org/10.1006/jmbi.1994.0095

Smith, N. J., & Fenton, T. R. (2019). The APOBEC3 genes and their role in cancer: Insights from human papillomavirus. *Journal of Molecular Endocrinology*, *62*(4), Article 4. https://doi.org/10.1530/JME-19-0011

Soleimani Zakeri, N. S., Pashazadeh, S., & MotieGhader, H. (2020). Gene biomarker discovery at different stages of Alzheimer using gene co-expression network approach. *Scientific Reports*, *10*(1), 12210. https://doi.org/10.1038/s41598-020-69249-8

Song, P., Zhang, J.-H., Qin, J., Gao, X.-B., Yu, J., Tang, X.-G., Tang, C.-F., & Huang, L. (2014). Smoking is associated with the incidence of AMS: A large-sample cohort study. *Military Medical Research*, *1*, 16. https://doi.org/10.1186/2054-9369-1-16

Sotero-Caio, C. G., Platt, R. N., Suh, A., & Ray, D. A. (2017). Evolution and Diversity of Transposable Elements in Vertebrate Genomes. *Genome Biology and Evolution*, *9*(1), Article 1. https://doi.org/10.1093/gbe/evw264

Southam, L., Gilly, A., Süveges, D., Farmaki, A.-E., Schwartzentruber, J., Tachmazidou, I., Matchan, A., Rayner, N. W., Tsafantakis, E., Karaleftheri, M., Xue, Y., Dedoussis, G., & Zeggini, E. (2017). Whole genome sequencing and imputation in isolated populations identify genetic associations with medically-relevant complex traits. *Nature Communications*, *8*, 15606. https://doi.org/10.1038/ncomms15606

Spellmann, I., Reinhard, M. A., Veverka, D., Zill, P., Obermeier, M., Dehning, S., Schennach, R., Müller, N., Möller, H.-J., Riedel, M., & Musil, R. (2018). QTc prolongation in short-term treatment of schizophrenia patients: Effects of different antipsychotics and genetic factors. *European Archives of Psychiatry and Clinical Neuroscience*, *268*(4), 383–390. https://doi.org/10.1007/s00406-018-0880-8

Srinivasan, S., Bettella, F., Mattingsdal, M., Wang, Y., Witoelar, A., Schork, A. J., Thompson, W. K., Zuber, V., Schizophrenia Working Group of the Psychiatric Genomics Consortium, The International Headache Genetics Consortium, Winsvold, B. S., Zwart, J.-A., Collier, D. A., Desikan, R. S., Melle, I., Werge, T., Dale, A. M., Djurovic, S., & Andreassen, O. A. (2016). Genetic Markers of Human Evolution Are Enriched in Schizophrenia. *Biological Psychiatry*, *80*(4), 284–292. https://doi.org/10.1016/j.biopsych.2015.10.009

Stavrou, S., Crawford, D., Blouch, K., Browne, E. P., Kohli, R. M., & Ross, S. R. (2014). Different Modes of Retrovirus Restriction by Human APOBEC3A and APOBEC3G In Vivo. *PLoS Pathogens*, *10*(5), Article 5. https://doi.org/10.1371/journal.ppat.1004145

Stearrett, N., Dawson, T., Rahnavard, A., Bachali, P., Bendall, M. L., Zeng, C., Caricchio, R., Pérez-Losada, M., Grammer, A. C., Lipsky, P. E., & Crandall, K. A. (2021). Expression of Human Endogenous Retroviruses in Systemic Lupus Erythematosus: Multiomic Integration With Gene Expression. *Frontiers in Immunology*, *12*, 661437. https://doi.org/10.3389/fimmu.2021.661437

Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., Stein, T. I., Nudel, R., Lieder, I., Mazor, Y., Kaplan, S., Dahary, D., Warshawsky, D., Guan-Golan, Y., Kohn, A., Rappaport, N., Safran, M., & Lancet, D. (2016). The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Current Protocols in Bioinformatics*, *54*, 1.30.1-1.30.33. https://doi.org/10.1002/cpbi.5

Stenglein, M. D., & Harris, R. S. (2006). APOBEC3B and APOBEC3F inhibit L1 retrotransposition by a DNA deamination-independent mechanism. *The Journal of Biological Chemistry*, *281*(25), Article 25. https://doi.org/10.1074/jbc.M602367200

Strebel, K., Daugherty, D., Clouse, K., Cohen, D., Folks, T., & Martin, M. A. (1987). The HIV 'A' (sor) gene product is essential for virus infectivity. *Nature*, *328*(6132), Article 6132. https://doi.org/10.1038/328728a0

Su, M., Han, D., Boyd-Kirkup, J., Yu, X., & Han, J.-D. J. (2014). Evolution of Alu elements toward enhancers. *Cell Reports*, *7*(2), 376–385. https://doi.org/10.1016/j.celrep.2014.03.011

Suarez, N. A., Macia, A., & Muotri, A. R. (2018). LINE-1 retrotransposons in healthy and diseased human brain. *Developmental Neurobiology*, *78*(5), 434–455. https://doi.org/10.1002/dneu.22567

Subbiah, V., Dumbrava, E. I., Jiang, Y., Thein, K. Z., Naing, A., Hong, D. S., Fu, S., Piha-Paul, S. A., Tsimberidou, A. M., Janku, F., Meric-Bernstam, F., Kurzrock, R., & Falchook, G. (2020). Dual EGFR blockade with cetuximab and erlotinib combined with anti-VEGF antibody bevacizumab in advanced solid tumors: A phase 1 dose escalation triplet combination trial. *Experimental Hematology & Oncology*, *9*(1), 7. https://doi.org/10.1186/s40164-020-00159-1

Sugano, E., Edwards, G., Saha, S., Wilmott, L. A., Grambergs, R. C., Mondal, K., Qi, H., Stiles, M., Tomita, H., & Mandal, N. (2019). Overexpression of acid ceramidase (ASAH1) protects retinal cells (ARPE19) from oxidative stress. *Journal of Lipid Research*, *60*(1), 30–43. https://doi.org/10.1194/jlr.M082198

Sultana, R., Yu, C.-E., Yu, J., Munson, J., Chen, D., Hua, W., Estes, A., Cortes, F., de la Barra, F., Yu, D., Haider, S. T., Trask, B. J., Green, E. D., Raskind, W. H., Disteche, C. M., Wijsman, E., Dawson, G., Storm, D. R., Schellenberg, G. D., & Villacres, E. C. (2002). Identification of a novel gene on chromosome 7q11.2 interrupted by a translocation breakpoint in a pair of autistic twins. *Genomics*, *80*(2), 129–134. https://doi.org/10.1006/geno.2002.6810

Sun, W., Samimi, H., Gamez, M., Zare, H., & Frost, B. (2018). Pathogenic tau-induced piRNA depletion promotes neuronal death through transposable element dysregulation in neurodegenerative tauopathies. *Nature Neuroscience*, *21*(8), 1038–1048. https://doi.org/10.1038/s41593-018-0194-1

Suntsova, M., Garazha, A., Ivanova, A., Kaminsky, D., Zhavoronkov, A., & Buzdin, A. (2015). Molecular functions of human endogenous retroviruses in health and disease. *Cellular and Molecular Life Sciences: CMLS*, *72*(19), 3653–3675. https://doi.org/10.1007/s00018-015-1947-6

Suntsova, M., Gogvadze, E. V., Salozhin, S., Gaifullin, N., Eroshkin, F., Dmitriev, S. E., Martynova, N., Kulikov, K., Malakhova, G., Tukhbatova, G., Bolshakov, A. P., Ghilarov, D., Garazha, A., Aliper, A., Cantor, C. R., Solokhin, Y., Roumiantsev, S., Balaban, P., Zhavoronkov, A., & Buzdin, A. (2013). Human-specific endogenous retroviral insert serves as an enhancer for the schizophrenia-linked gene PRODH. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(48), 19472–19477. https://doi.org/10.1073/pnas.1318172110

Suspène, R., Aynaud, M.-M., Koch, S., Pasdeloup, D., Labetoulle, M., Gaertner, B., Vartanian, J.-P., Meyerhans, A., & Wain-Hobson, S. (2011). Genetic editing of herpes simplex virus 1 and Epstein-Barr herpesvirus genomes by human APOBEC3 cytidine deaminases in culture and in vivo. *Journal of Virology*, *85*(15), Article 15. https://doi.org/10.1128/JVI.00290-11

Swanton, C., McGranahan, N., Starrett, G. J., & Harris, R. S. (2015). APOBEC Enzymes: Mutagenic Fuel for Cancer Evolution and Heterogeneity. *Cancer Discovery*, *5*(7), Article 7. https://doi.org/10.1158/2159-8290.CD-15-0344

Swergold, G. D. (1990). Identification, characterization, and cell specificity of a human LINE-1 promoter. *Molecular and Cellular Biology*, *10*(12), Article 12. https://doi.org/10.1128/mcb.10.12.6718-6729.1990

Takahashi, I., Hama, Y., Matsushima, M., Hirotani, M., Kano, T., Hohzen, H., Yabe, I., Utsumi, J., & Sasaki, H. (2015). Identification of plasma microRNAs as a biomarker of sporadic Amyotrophic Lateral Sclerosis. *Molecular Brain*, *8*(1), 67. https://doi.org/10.1186/s13041-015-0161-7

Talkowski, M. E., Rosenfeld, J. A., Blumenthal, I., Pillalamarri, V., Chiang, C., Heilbut, A., Ernst, C., Hanscom, C., Rossin, E., Lindgren, A. M., Pereira, S., Ruderfer, D., Kirby, A., Ripke, S., Harris, D. J., Lee, J.-H., Ha, K., Kim, H.-G., Solomon, B. D., … Gusella, J. F. (2012). Sequencing chromosomal abnormalities reveals neurodevelopmental loci that confer risk across diagnostic boundaries. *Cell*, *149*(3), 525–537. https://doi.org/10.1016/j.cell.2012.03.028

Teffer, K., & Semendeferi, K. (2012). Human prefrontal cortex: Evolution, development, and pathology. *Progress in Brain Research*, *195*, 191–218. https://doi.org/10.1016/B978-0-444-53860-4.00009-X

Terry, D. M., & Devine, S. E. (2019). Aberrantly High Levels of Somatic LINE-1 Expression and Retrotransposition in Human Neurological Disorders. *Frontiers in Genetics*, *10*, 1244. https://doi.org/10.3389/fgene.2019.01244

The 1000 Genomes Project Consortium, Corresponding authors, Auton, A., Abecasis, G. R., Steering committee, Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., … Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Tuller, T., Kupiec, M., & Ruppin, E. (2008). Evolutionary Rate and Gene Expression Across Different Brain Regions. *Genome Biology*, *9*(9), R142. https://doi.org/10.1186/gb-2008-9-9-r142

Ullu, E., & Tschudi, C. (1984). Alu sequences are processed 7SL RNA genes. *Nature*, *312*(5990), Article 5990. https://doi.org/10.1038/312171a0

Uriu, K., Kosugi, Y., Suzuki, N., Ito, J., & Sato, K. (2021). Elucidation of the Complicated Scenario of Primate APOBEC3 Gene Evolution. *Journal of Virology*, *95*(12), Article 12. https://doi.org/10.1128/JVI.00144-21

van den Heuvel, M. P., Scholtens, L. H., de Lange, S. C., Pijnenburg, R., Cahn, W., van Haren, N. E. M., Sommer, I. E., Bozzali, M., Koch, K., Boks, M. P., Repple, J., Pievani, M., Li, L., Preuss, T. M., & Rilling, J. K. (2019). Evolutionary modifications in human brain connectivity associated with schizophrenia. *Brain: A Journal of Neurology*, *142*(12), 3991–4002. https://doi.org/10.1093/brain/awz330

Vanderhaeghen, P., & Polleux, F. (2023). Developmental mechanisms underlying the evolution of human cortical circuits. *Nature Reviews Neuroscience*, *24*(4), 213–232. https://doi.org/10.1038/s41583-023-00675-z

Vartanian, J.-P., Guétard, D., Henry, M., & Wain-Hobson, S. (2008). Evidence for editing of human papillomavirus DNA by APOBEC3 in benign and precancerous lesions. *Science (New York, N.Y.)*, *320*(5873), Article 5873. https://doi.org/10.1126/science.1153201

Vega-Hinojosa, O., Cardiel, M. H., & Ochoa-Miranda, P. (2018). Prevalence of musculoskeletal manifestations and related disabilities in a Peruvian urban population living at high altitude. COPCORD Study. Stage I. *Reumatologia Clinica*, *14*(5), 278–284. https://doi.org/10.1016/j.reuma.2017.01.011

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer New York. https://doi.org/10.1007/978-0-387-21706-2

Villanueva‐Cañas, J. L., Rech, G. E., de Cara, M. A. R., & González, J. (2017). Beyond SNP s: How to detect selection on transposable element insertions. *Methods in Ecology and Evolution*, *8*(6), 728–737. https://doi.org/10.1111/2041-210X.12781

Vinnikov, D., Blanc, P. D., & Steinmaus, C. (2016). Is Smoking a Predictor for Acute Mountain Sickness? Findings From a Meta-Analysis. *Nicotine & Tobacco Research: Official Journal of the Society for Research on Nicotine and Tobacco*, *18*(6), 1509–1516. https://doi.org/10.1093/ntr/ntv218

Vinnikov, D., Brimkulov, N., & Blanc, P. D. (2015). Smoking increases the risk of acute mountain sickness. *Wilderness & Environmental Medicine*, *26*(2), 164–172. https://doi.org/10.1016/j.wem.2014.10.006

Volk, D. W., Edelson, J. R., & Lewis, D. A. (2016). Altered expression of developmental regulators of parvalbumin and somatostatin neurons in the prefrontal cortex in schizophrenia. *Schizophrenia Research*, *177*(1–3), 3–9. https://doi.org/10.1016/j.schres.2016.03.001

von Schwedler, U., Song, J., Aiken, C., & Trono, D. (1993). Vif is crucial for human immunodeficiency virus type 1 proviral DNA synthesis in infected cells. *Journal of Virology*, *67*(8), Article 8. https://doi.org/10.1128/JVI.67.8.4945-4955.1993

Wang, H., Xing, J., Grover, D., Hedges, D. J., Han, K., Walker, J. A., & Batzer, M. A. (2005). SVA elements: A hominid-specific retroposon family. *Journal of Molecular Biology*, *354*(4), Article 4. https://doi.org/10.1016/j.jmb.2005.09.085

Wang, J., Shaban, N. M., Land, A. M., Brown, W. L., & Harris, R. S. (2018). Simian Immunodeficiency Virus Vif and Human APOBEC3B Interactions Resemble Those between HIV-1 Vif and Human APOBEC3G. *Journal of Virology*, *92*(12), Article 12. https://doi.org/10.1128/JVI.00447-18

Wang, J., Xie, G., Singh, M., Ghanbarian, A. T., Raskó, T., Szvetnik, A., Cai, H., Besser, D., Prigione, A., Fuchs, N. V., Schumann, G. G., Chen, W., Lorincz, M. C., Ivics, Z., Hurst, L. D., & Izsvák, Z. (2014).

Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. *Nature*, *516*(7531), 405–409. https://doi.org/10.1038/nature13804

Wang, K., Li, W.-D., Zhang, C. K., Wang, Z., Glessner, J. T., Grant, S. F. A., Zhao, H., Hakonarson, H., & Price, R. A. (2011). A genome-wide association study on obesity and obesity-related traits. *PloS One*, *6*(4), e18939. https://doi.org/10.1371/journal.pone.0018939

Wang, L., Norris, E. T., & Jordan, I. K. (2017). Human Retrotransposon Insertion Polymorphisms Are Associated with Health and Disease via Gene Regulatory Phenotypes. *Frontiers in Microbiology*, *8*, 1418. https://doi.org/10.3389/fmicb.2017.01418

Wang, X., Liu, J., Wang, Q., & Chen, Q. (2023). The transcriptomic and epigenetic alterations in type 2 diabetes mellitus patients of Chinese Tibetan and Han populations. *Frontiers in Endocrinology*, *14*, 1122047. https://doi.org/10.3389/fendo.2023.1122047

Wang, Y., Zhao, B., Choi, J., & Lee, E. A. (2021). Genomic approaches to trace the history of human brain evolution with an emerging opportunity for transposon profiling of ancient humans. *Mobile DNA*, *12*(1), 22. https://doi.org/10.1186/s13100-021-00250-2

Wang, Z.-H., Liu, P., Liu, X., Manfredsson, F. P., Sandoval, I. M., Yu, S. P., Wang, J.-Z., & Ye, K. (2017). Delta-Secretase Phosphorylation by SRPK2 Enhances Its Enzymatic Activity, Provoking Pathogenesis in Alzheimer's Disease. *Molecular Cell*, *67*(5), 812-825.e5. https://doi.org/10.1016/j.molcel.2017.07.018

Warren, C. J., Santiago, M. L., & Pyeon, D. (2022). APOBEC3: Friend or Foe in Human Papillomavirus Infection and Oncogenesis? *Annual Review of Virology*. https://doi.org/10.1146/annurev-virology-092920-030354

Warren, I. A., Naville, M., Chalopin, D., Levin, P., Berger, C. S., Galiana, D., & Volff, J.-N. (2015). Evolutionary impact of transposable elements on genomic diversity and lineage-specific innovation in vertebrates. *Chromosome Research: An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology*, *23*(3), Article 3. https://doi.org/10.1007/s10577-015-9493-5

Watkins, W. S., Feusier, J. E., Thomas, J., Goubert, C., Mallick, S., & Jorde, L. B. (2020). The Simons Genome Diversity Project: A Global Analysis of Mobile Element Diversity. *Genome Biology and Evolution*, *12*(6), 779–794. https://doi.org/10.1093/gbe/evaa086

Watkins, W. S., Rogers, A. R., Ostler, C. T., Wooding, S., Bamshad, M. J., Brassington, A.-M. E., Carroll, M. L., Nguyen, S. V., Walker, J. A., Prasad, B. V. R., Reddy, P. G., Das, P. K., Batzer, M. A., & Jorde, L. B. (2003). Genetic variation among world populations: Inferences from 100 Alu insertion polymorphisms. *Genome Research*, *13*(7), 1607–1618. https://doi.org/10.1101/gr.894603

Watson, C. M., Crinnion, L. A., Lindsay, H., Mitchell, R., Camm, N., Robinson, R., Joyce, C., Tanteles, G. A., Halloran, D. J. O., Pena, S. D. J., Carr, I. M., & Bonthron, D. T. (2021). Assessing the utility of long-read nanopore sequencing for rapid and efficient characterization of mobile element insertions. *Laboratory Investigation; a Journal of Technical Methods and Pathology*, *101*(4), 442–449. https://doi.org/10.1038/s41374-020-00489-y

Watson, J. D., & Crick, F. H. C. (1953). Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, *171*(4356), 737–738. https://doi.org/10.1038/171737a0

Wei, Y., de Lange, S. C., Scholtens, L. H., Watanabe, K., Ardesch, D. J., Jansen, P. R., Savage, J. E., Li, L., Preuss, T. M., Rilling, J. K., Posthuma, D., & van den Heuvel, M. P. (2019). Genetic mapping and evolutionary analysis of human-expanded cognitive networks. *Nature Communications*, *10*(1), 4839. https://doi.org/10.1038/s41467-019-12764-8

Weiner, A. M. (2002). SINEs and LINEs: The art of biting the hand that feeds you. *Current Opinion in Cell Biology*, *14*(3), 343–350. https://doi.org/10.1016/s0955-0674(02)00338-1

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, *38*(6), 1358. https://doi.org/10.2307/2408641

West, S., & Proudfoot, N. J. (2009). Transcriptional termination enhances protein expression in human cells. *Molecular Cell*, *33*(3), 354–364. https://doi.org/10.1016/j.molcel.2009.01.008

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., & Schulman, A. H. (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews. Genetics*, *8*(12), Article 12. https://doi.org/10.1038/nrg2165

Wissing, S., Montano, M., Garcia-Perez, J. L., Moran, J. V., & Greene, W. C. (2011). Endogenous APOBEC3B restricts LINE-1 retrotransposition in transformed cells and human embryonic stem cells. *The Journal of Biological Chemistry*, *286*(42), Article 42. https://doi.org/10.1074/jbc.M111.251058

Wright, M. N., & Ziegler, A. (2017). **ranger**: A Fast Implementation of Random Forests for High Dimensional Data in *C++* and *R*. *Journal of Statistical Software*, *77*(1). https://doi.org/10.18637/jss.v077.i01

Wu, T.-Y., Ding, S.-Q., Liu, J.-L., Jia, J.-H., Chai, Z.-C., Dai, R.-C., Zhao, J.-Z., Tang, Q. D., & Kayser, B. (2012). Smoking, acute mountain sickness and altitude acclimatisation: A cohort study. *Thorax*, *67*(10), 914–919. https://doi.org/10.1136/thoraxjnl-2011-200623

Wu, X., Huai, C., Shen, L., Li, M., Yang, C., Zhang, J., Chen, L., Zhu, W., Fan, L., Zhou, W., Xing, Q., He, L., Wan, C., & Qin, S. (2021). Genome-wide study of copy number variation implicates multiple novel loci for schizophrenia risk in Han Chinese family trios. *IScience*, *24*(8), 102894. https://doi.org/10.1016/j.isci.2021.102894

Xiang, K., Ouzhuluobu, Peng, Y., Yang, Z., Zhang, X., Cui, C., Zhang, H., Li, M., Zhang, Y., Bianba, Gonggalanzi, Basang, Ciwangsangbu, Wu, T., Chen, H., Shi, H., Qi, X., & Su, B. (2013). Identification of a Tibetan-Specific Mutation in the Hypoxic Gene EGLN1 and Its Contribution to High-Altitude Adaptation. *Molecular Biology and Evolution*, *30*(8), 1889–1898. https://doi.org/10.1093/molbev/mst090

Xu, B., Ionita-Laza, I., Roos, J. L., Boone, B., Woodrick, S., Sun, Y., Levy, S., Gogos, J. A., & Karayiorgou, M. (2012). De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nature Genetics*, *44*(12), 1365–1369. https://doi.org/10.1038/ng.2446

Xu, S., Li, S., Yang, Y., Tan, J., Lou, H., Jin, W., Yang, L., Pan, X., Wang, J., Shen, Y., Wu, B., Wang, H., & Jin, L. (2011). A genome-wide search for signals of high-altitude adaptation in Tibetans. *Molecular Biology and Evolution*, *28*(2), 1003–1011. https://doi.org/10.1093/molbev/msq277

Xue, Y., Mezzavilla, M., Haber, M., McCarthy, S., Chen, Y., Narasimhan, V., Gilly, A., Ayub, Q., Colonna, V., Southam, L., Finan, C., Massaia, A., Chheda, H., Palta, P., Ritchie, G., Asimit, J., Dedoussis, G., Gasparini, P., Palotie, A., … Zeggini, E. (2017). Enrichment of low-frequency functional variants

revealed by whole-genome sequencing of multiple isolated European populations. *Nature Communications*, *8*, 15927. https://doi.org/10.1038/ncomms15927

Yang, J., Jin, Z.-B., Chen, J., Huang, X.-F., Li, X.-M., Liang, Y.-B., Mao, J.-Y., Chen, X., Zheng, Z., Bakshi, A., Zheng, D.-D., Zheng, M.-Q., Wray, N. R., Visscher, P. M., Lu, F., & Qu, J. (2017). Genetic signatures of high-altitude adaptation in Tibetans. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(16), 4189–4194. https://doi.org/10.1073/pnas.1617042114

Yao, Y., Chu, X., Ma, M., Ye, J., Wen, Y., Li, P., Cheng, B., Cheng, S., Zhang, L., Liu, L., Qi, X., Liang, C., Kafle, O. P., Wu, C., Wang, S., Wang, X., Ning, Y., & Zhang, F. (2021). Evaluate the effects of serum urate level on bone mineral density: A genome-wide gene-environment interaction analysis in UK Biobank cohort. *Endocrine*, *73*(3), 702–711. https://doi.org/10.1007/s12020-021-02760-8

Yao, Y., Yang, J., Xie, Y., Liao, H., Yang, B., Xu, Q., & Rao, S. (2020). No Evidence for Widespread Positive Selection Signatures in Common Risk Alleles Associated with Schizophrenia. *Schizophrenia Bulletin*, *46*(3), 603–611. https://doi.org/10.1093/schbul/sbz048

Yengo, L., Vedantam, S., Marouli, E., Sidorenko, J., Bartell, E., Sakaue, S., Graff, M., Eliasen, A. U., Jiang, Y., Raghavan, S., Miao, J., Arias, J. D., Graham, S. E., Mukamel, R. E., Spracklen, C. N., Yin, X., Chen, S.-H., Ferreira, T., Highland, H. H., … Hirschhorn, J. N. (2022). A saturated map of common genetic variants associated with human height. *Nature*, *610*(7933), 704–712. https://doi.org/10.1038/s41586-022-05275-y

Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z. X. P., Pool, J. E., Xu, X., Jiang, H., Vinckenbosch, N., Korneliussen, T. S., Zheng, H., Liu, T., He, W., Li, K., Luo, R., Nie, X., Wu, H., Zhao, M., Cao, H., … Wang, J. (2010). Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude. *Science*, *329*(5987), 75–78. https://doi.org/10.1126/science.1190371

Yin, L., Zhang, H., Tang, Z., Xu, J., Yin, D., Zhang, Z., Yuan, X., Zhu, M., Zhao, S., Li, X., & Liu, X. (2021). rMVP: A Memory-efficient, Visualization-enhanced, and Parallel-accelerated Tool for Genome-wide Association Study. *Genomics, Proteomics & Bioinformatics*, *19*(4), 619–628. https://doi.org/10.1016/j.gpb.2020.10.007

Yin, Y., Zhou, L., & Yuan, S. (2018). Enigma of Retrotransposon Biology in Mammalian Early Embryos and Embryonic Stem Cells. *Stem Cells International*, *2018*, 6239245. https://doi.org/10.1155/2018/6239245

You, H., Li, X., Pei, T., Huang, Q., Liu, F., & Gao, Y. (2012). Predictive value of basal exhaled nitric oxide and carbon monoxide for acute mountain sickness. *Wilderness & Environmental Medicine*, *23*(4), 316–324. https://doi.org/10.1016/j.wem.2012.04.001

Yurdakul, E., Barlas, Y., & Ulgen, K. O. (2023). Circadian clock crosstalks with autism. *Brain and Behavior*, *13*(12), e3273. https://doi.org/10.1002/brb3.3273

Zhang, B., Xu, Y.-H., Wei, S.-G., Zhang, H.-B., Fu, D.-K., Feng, Z.-F., Guan, F.-L., Zhu, Y.-S., & Li, S.-B. (2014). Association study identifying a new susceptibility gene (AUTS2) for schizophrenia. *International Journal of Molecular Sciences*, *15*(11), 19406–19416. https://doi.org/10.3390/ijms151119406

Zhang, J., & Webb, D. M. (2004). Rapid evolution of primate antiviral enzyme APOBEC3G. *Human Molecular Genetics*, *13*(16), Article 16. https://doi.org/10.1093/hmg/ddh183

Zhang, L., Meng, J., Li, H., Tang, M., Zhou, Z., Zhou, X., Feng, L., Li, X., Guo, Y., He, Y., He, W., & Huang, X. (2022). Hippocampal adaptation to high altitude: A neuroanatomic profile of hippocampal subfields in Tibetans and acclimatized Han Chinese residents. *Frontiers in Neuroanatomy*, *16*, 999033. https://doi.org/10.3389/fnana.2022.999033

Zhang, X., Lin, P.-Y., Liakath-Ali, K., & Südhof, T. C. (2022). Teneurins assemble into presynaptic nanoclusters that promote synapse formation via postsynaptic non-teneurin ligands. *Nature Communications*, *13*(1), 2297. https://doi.org/10.1038/s41467-022-29751-1

Zhang, X., Witt, K. E., Bañuelos, M. M., Ko, A., Yuan, K., Xu, S., Nielsen, R., & Huerta-Sanchez, E. (2021). The history and evolution of the Denisovan-EPAS1 haplotype in Tibetans. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(22), e2020803118. https://doi.org/10.1073/pnas.2020803118

Zhang, X., Xie, W., Du, W., Liu, Y., Lin, J., Yin, W., Yang, L., Yuan, F., Zhang, R., Liu, H., Ma, H., & Zhang, J. (2023). Consistent differences in brain structure and functional connectivity in high-altitude

native Tibetans and immigrants. *Brain Imaging and Behavior*, *17*(3), 271–281. https://doi.org/10.1007/s11682-023-00759-5

Zhang, Y., Chen, X., Cao, Y., & Yang, Z. (2021). Roles of APOBEC3 in hepatitis B virus (HBV) infection and hepatocarcinogenesis. *Bioengineered*, *12*(1), Article 1. https://doi.org/10.1080/21655979.2021.1931640

Zhang, Y., Iwasaki, H., Wang, H., Kudo, T., Kalka, T. B., Hennet, T., Kubota, T., Cheng, L., Inaba, N., Gotoh, M., Togayachi, A., Guo, J., Hisatomi, H., Nakajima, K., Nishihara, S., Nakamura, M., Marth, J. D., & Narimatsu, H. (2003). Cloning and characterization of a new human UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase, designated pp-GalNAc-T13, that is specifically expressed in neurons and synthesizes GalNAc alpha-serine/threonine antigen. *The Journal of Biological Chemistry*, *278*(1), 573–584. https://doi.org/10.1074/jbc.M203094200

Zheng, Q., Li, G., Wang, S., Zhou, Y., Liu, K., Gao, Y., Zhou, Y., Zheng, L., Zhu, L., Deng, Q., Wu, M., Di, A., Zhang, L., Zhao, Y., Zhang, H., Sun, H., Dong, C., Xu, H., & Wang, X. (2021). Trisomy 21-induced dysregulation of microglial homeostasis in Alzheimer's brains is mediated by USP25. *Science Advances*, *7*(1), eabe1340. https://doi.org/10.1126/sciadv.abe1340

Zheng, W., He, Y., Guo, Y., Yue, T., Zhang, H., Li, J., Zhou, B., Zeng, X., Li, L., Wang, B., Cao, J., Chen, L., Li, C., Li, H., Cui, C., Bai, C., Baimakangzhuo, null, Qi, X., Ouzhuluobu, null, & Su, B. (2023). Large-scale genome sequencing redefines the genetic footprints of high-altitude adaptation in Tibetans. *Genome Biology*, *24*(1), 73. https://doi.org/10.1186/s13059-023-02912-1

Zheng, X., Zheng, D., Zhang, C., Guo, H., Zhang, Y., Xue, X., Shi, Z., Zhang, X., Zeng, X., Wu, Y., & Gao, W. (2023). A cuproptosis-related lncRNA signature predicts the prognosis and immune cell status in head and neck squamous cell carcinoma. *Frontiers in Oncology*, *13*, 1055717. https://doi.org/10.3389/fonc.2023.1055717

Zhou, H., Cheng, Z., Bass, N., Krystal, J. H., Farrer, L. A., Kranzler, H. R., & Gelernter, J. (2018). Genome-wide association study identifies glutamate ionotropic receptor GRIA4 as a risk gene for comorbid nicotine dependence and major depression. *Translational Psychiatry*, *8*(1), 208. https://doi.org/10.1038/s41398-018-0258-8

Zhou, X., & Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, *44*(7), 821–824. https://doi.org/10.1038/ng.2310

Zhou, X., & Stephens, M. (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature Methods*, *11*(4), 407–409. https://doi.org/10.1038/nmeth.2848

Ziffra, R. S., Kim, C. N., Ross, J. M., Wilfert, A., Turner, T. N., Haeussler, M., Casella, A. M., Przytycki, P. F., Keough, K. C., Shin, D., Bogdanoff, D., Kreimer, A., Pollard, K. S., Ament, S. A., Eichler, E. E., Ahituv, N., & Nowakowski, T. J. (2021). Single-cell epigenomics reveals mechanisms of human cortical development. *Nature*, *598*(7879), 205–213. https://doi.org/10.1038/s41586-021-03209-8

Zuo, H., Zheng, T., Wu, K., Yang, T., Wang, L., Nima, Q., Bai, H., Dong, K., Fan, Z., Huang, S., Luo, R., Wu, J., Zhou, J., Xu, H., Zhang, Y., Feng, S., Zeng, P., Xiao, X., Guo, B., … China Multi-Ethnic Cohort (CMEC). (2022). High-altitude exposure decreases bone mineral density and its relationship with gut microbiota: Results from the China multi-ethnic cohort (CMEC) study. *Environmental Research*, *215*(Pt 2), 114206. https://doi.org/10.1016/j.envres.2022.114206

## Supplementary Material

| ID | dataSource | country | groupLabel |
|---|---|---|---|
| 12884A | Schiffels NatureCommunications 2016 | UK | Britain IronAge |
| 15570A | Schiffels NatureCommunications 2016 | UK | Britain Medieval Saxon |
| AHUR 2064 | MorenoMayar Science 2018 | USA | USA SpiritCave |
| AltaiNea | Prufer Nature 2014 | Russia | Siberia Neanderthal |
| Andaman | MorenoMayar Science 2018 | India | India Historical GreatAndaman |
| baa | Schlebusch Science 2017 | South Africa | SouthAfrica Neolithic |
| Bichon | Jones NatureCommunications 2015 | Switzerland | Switzerland Mesolithic |
| Denisova | Meyer Science 2012 | Russia | Siberia Denisova |
| Funadomari 23 | KanzawaKiriyama AnthropScience 2019 | Japan | Japan Jomon |
| HSJ-A-1 | Ebenesersdottir Science 2018 | Iceland | Iceland VikingAge |
| LBK | Lazaridis Nature 2014 | Germany | Germany Neolithic |
| Loschbour | Lazaridis Nature 2014 | Luxembourg | Luxembourg Mesolithic |
| Lovelock2 | MorenoMayar Science 2018 | USA | USA LovelockCave |
| Lovelock3 | MorenoMayar Science 2018 | USA | USA LovelockCave |
| mfo | Schlebusch Science 2017 | South Africa | SouthAfrica IronAge |
| new | Schlebusch Science 2017 | South Africa | SouthAfrica IronAge |
| prs009 | SanchezQuinto PNAS 2019 | Ireland | Ireland Neolithic Megalithic |
| prs016 | SanchezQuinto PNAS 2019 | Ireland | Ireland Neolithic Megalithic |
| Saqqaq | Rasmussen Nature 2010 | Greenland | Greenland PaleoEskimo |
| SBT-A-1 | Ebenesersdottir Science 2018 | Iceland | Iceland VikingAge |
| sf12 | Gunther PLoSBiology 2018 | Sweden | Sweden Mesolithic |
| SIII | Sikora Science 2017 | Russia | Russia UpperPaleolithic Sunghir |
| SSG-A-2 | Ebenesersdottir Science 2018 | Iceland | Iceland VikingAge |
| SSG-A-4 | Ebenesersdottir Science 2018 | Iceland | Iceland VikingAge |
| STT-A-2 | Ebenesersdottir Science 2018 | Iceland | Iceland VikingAge |
| SZ1 | Amorim NatureCommunications 2018 | Hungary | Hungary Medieval Avar |
| SZ11 | Amorim NatureCommunications 2018 | Hungary | Hungary Medieval Langobard |
| SZ15 | Amorim NatureCommunications 2018 | Hungary | Hungary Medieval Langobard |
| SZ2 | Amorim NatureCommunications 2018 | Hungary | Hungary Medieval Langobard |
| SZ3 | Amorim NatureCommunications 2018 | Hungary | Hungary Medieval Langobard |
| SZ4 | Amorim NatureCommunications 2018 | Hungary | Hungary Medieval Langobard |
| SZ43 | Amorim NatureCommunications 2018 | Hungary | Hungary Medieval Langobard |
| SZ45 | Amorim NatureCommunications 2018 | Hungary | Hungary Medieval Langobard |
| USR1 | MorenoMayar Nature 2017 | USA | USA Beringia |
| UstIshim | Fu Nature 2014 | Russia | Siberia UpperPaleolithic UstIshim |

**Supplementary Table S1.** List of the 35 ancient *Homo* samples analyzed with *ngs te mapper 2*. "ID" = name/label of the sample; "dataSource" = relative publication; "country" = modern country in which the sample has been discovered; "groupLabel" = age group of the sample.

| Chr | ngs_pos | MELT_pos | altai_ngs | altai_MELT | den_ngs | den_MELT | ustishim_ngs | ustishim_MELT |
|---|---|---|---|---|---|---|---|---|
| 1 | 184689643 | 184689638 | ? | Abs | ? | Abs | ? | Pres |
| 1 | 189084959 | 189084961 | ? | ./. | ? | Pres | ? | ./. |
| 1 | 28247807 | 28247800 | ? | Pres | ? | Pres | ? | Pres |
| 1 | 46207290 | 46207287 | ? | Pres | ? | Pres | ? | Pres |
| 1 | 65909639 | 65909638 | Pres | ./. | ? | ./. | ? | ./. |
| 1 | 69943724 | 69943716 | ? | Abs | ? | Abs | ? | Pres |
| 1 | 91914108 | 91914111 | ? | Pres | ? | ./. | ? | ./. |
| 2 | 100510932 | 100510927 | ? | Pres | ? | Pres | ? | Pres |
| **2** | **109302500** | **109302490** | ? | Pres | ? | Pres | Pres | Pres |
| **2** | **120912220** | **120912219** | Pres | Pres | Pres | Pres | ? | Pres |
| 2 | 142788904 | 142788902 | ? | Abs | ? | Pres | ? | ./. |
| 2 | 144556642 | 144556642 | ? | Pres | ? | Pres | ? | Pres |
| **2** | **149062350** | **149062341** | ? | Pres | Pres | Pres | ? | Pres |
| 2 | 151420592 | 151420592 | ? | Pres | ? | Pres | ? | Pres |
| 2 | 152271950 | 152271951 | ? | Pres | ? | Pres | ? | Pres |
| 2 | 171527988 | 171527980 | ? | Pres | ? | Pres | ? | Pres |
| **2** | **188301987** | **188301982** | ? | Pres | ? | Pres | Pres | Pres |
| 2 | 196051831 | 196051831 | ? | ./. | ? | ./. | ? | ./. |
| 2 | 208605697 | 208605691 | ? | Pres | ? | Pres | ? | Pres |
| **2** | **219225932** | **219225932** | Pres | Pres | ? | Pres | ? | Pres |
| **2** | **242669740** | **242669730** | Pres | Pres | ? | Pres | ? | Pres |
| 2 | 2459618 | 2459610 | ? | Pres | ? | Pres | ? | Pres |
| 2 | 27598057 | 27598057 | ? | Pres | ? | Pres | ? | Pres |
| 2 | 33000259 | 33000253 | ? | Pres | ? | Pres | ? | Pres |
| 2 | 33636435 | 33636429 | ? | Pres | ? | Pres | ? | Pres |
| 2 | 34482756 | 34482756 | ? | Pres | ? | Pres | ? | Pres |
| **2** | **42025275** | **42025266** | Pres | Pres | ? | Pres | ? | Abs |
| **2** | **45535778** | **45535768** | Pres | Pres | ? | Pres | ? | Pres |
| **3** | **108696734** | **108696732** | ? | Pres | ? | Pres | Pres | Pres |
| **3** | **126763510** | **126763500** | Pres | Pres | ? | Pres | ? | Pres |
| 3 | 12983820 | 12983815 | ? | Pres | ? | Pres | ? | Pres |
| 3 | 183920892 | 183920884 | ? | Pres | ? | Pres | ? | Pres |
| 3 | 187593081 | 187593073 | ? | Pres | ? | Pres | ? | Pres |
| **3** | **191009246** | **191009246** | Pres | Pres | ? | Pres | ? | Pres |
| 3 | 196069566 | 196069558 | ? | Pres | ? | Pres | ? | Pres |
| 3 | 29533202 | 29533202 | ? | Pres | ? | Pres | ? | Pres |
| **3** | **48087212** | **48087212** | Pres | Pres | ? | Pres | ? | Pres |
| 3 | 57007041 | 57007031 | ? | Pres | ? | Pres | ? | Abs |
| 3 | 60402537 | 60402530 | ? | Pres | ? | Pres | ? | Pres |
| 3 | 64030624 | 64030614 | ? | Pres | ? | Pres | ? | Pres |
| 3 | 65493796 | 65493787 | ? | Pres | ? | Pres | ? | Pres |
| 3 | 69845641 | 69845631 | ? | Pres | ? | Pres | ? | Pres |
| 3 | 7347033 | 7347029 | ? | ./. | ? | Pres | ? | ./. |
| **3** | **82936657** | **82936658** | Pres | Pres | ? | Pres | ? | ./. |
| 3 | 95291394 | 95291388 | ? | Pres | ? | Pres | ? | Pres |
| 3 | 97473536 | 97473527 | ? | Pres | ? | Pres | ? | Pres |
| **4** | **107442392** | **107442392** | Pres | Pres | ? | Pres | ? | Pres |
| 4 | 111653615 | 111653605 | ? | Pres | ? | Pres | ? | Pres |
| **4** | **128889095** | **128889086** | Pres | Pres | Pres | Pres | ? | Pres |

Continues from the previous page

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 4 | 128952327 | 128952319 | ? | Pres | ? | Pres | ? | Pres |
| 4 | 130431164 | 130431155 | ? | Pres | ? | Pres | ? | Pres |
| **4** | **137580885** | **137580875** | ? | Pres | Pres | Pres | ? | Pres |
| **4** | **142409112** | **142409109** | ? | Pres | Pres | Pres | ? | Pres |
| 4 | 175932709 | 175932703 | ? | Pres | ? | Pres | ? | Pres |
| 4 | 177601071 | 177601062 | ? | Pres | ? | Pres | ? | Pres |
| **4** | **184491964** | **184491964** | ? | Pres | ? | Pres | Pres | Pres |
| **4** | **186849700** | **186849694** | ? | Pres | Pres | Pres | ? | Pres |
| 4 | 189124477 | 189124468 | ? | Pres | ? | Pres | ? | Pres |
| **4** | **189444846** | **189444837** | ? | Pres | Pres | Pres | ? | Pres |
| 4 | 189627369 | 189627359 | ? | Pres | ? | Pres | ? | Pres |
| 4 | 26600015 | 26600009 | ? | Pres | ? | Pres | ? | Pres |
| 4 | 40602369 | 40602363 | ? | Pres | ? | Pres | ? | Pres |
| 4 | 40924988 | 40924979 | ? | Pres | ? | Pres | ? | Pres |
| 4 | 78236288 | 78236288 | Pres | ./. | ? | ./. | ? | ./. |
| 4 | 85637667 | 85637657 | ? | Pres | ? | Pres | ? | Pres |
| 4 | 85691186 | 85691176 | ? | Pres | ? | Pres | ? | Pres |
| **4** | **92643806** | **92643801** | Pres | Pres | ? | Pres | ? | Pres |
| 4 | 95546606 | 95546597 | ? | Pres | ? | Pres | ? | Pres |
| **5** | **110076370** | **110076370** | Pres | Pres | ? | Pres | ? | Pres |
| 5 | 132918979 | 132918980 | ? | Pres | ? | Pres | ? | ./. |
| **5** | **137022577** | **137022578** | ? | ./. | Pres | ./. | Pres | Pres |
| 5 | 139150953 | 139150948 | ? | Pres | ? | Pres | ? | Pres |
| 5 | 144473272 | 144473274 | ? | Pres | ? | ./. | ? | ./. |
| **5** | **146857126** | **146857117** | Pres | Pres | ? | Pres | ? | Pres |
| 5 | 156568095 | 156568088 | ? | Pres | ? | Pres | ? | Pres |
| 5 | 160152464 | 160152455 | Pres | Abs | ? | Pres | ? | Pres |
| 5 | 21721786 | 21721777 | ? | Pres | ? | Pres | ? | Pres |
| 5 | 86907763 | 86907763 | ? | Pres | ? | Pres | ? | Pres |
| 5 | 94261110 | 94261100 | ? | Pres | ? | Pres | ? | Pres |
| 6 | 101215476 | 101215470 | ? | Pres | ? | Pres | ? | Pres |
| **6** | **113707586** | **113707576** | Pres | Pres | ? | Pres | ? | Pres |
| 6 | 130979656 | 130979650 | ? | Pres | ? | Pres | ? | Pres |
| **6** | **150412110** | **150412110** | Pres | Pres | ? | Pres | ? | Pres |
| **6** | **152851758** | **152851748** | Pres | Pres | ? | Pres | Pres | Pres |
| 6 | 16444065 | 16444058 | ? | Pres | ? | Pres | ? | Pres |
| 6 | 18434845 | 18434836 | ? | Pres | ? | Pres | Pres | Pres |
| 6 | 25688747 | 25688747 | ? | Pres | ? | Pres | ? | Pres |
| **6** | **3298924** | **3298915** | Pres | Pres | ? | Pres | ? | Pres |
| 6 | 36270879 | 36270875 | ? | Pres | ? | Pres | ? | Pres |
| 6 | 48540906 | 48540906 | ? | Pres | ? | Pres | ? | Pres |
| **6** | **55052605** | **55052605** | Pres | Pres | ? | Pres | ? | Pres |
| **6** | **70584361** | **70584352** | ? | Pres | Pres | Pres | Pres | Pres |
| 6 | 7076454 | 7076446 | ? | Pres | ? | Pres | ? | Pres |
| 6 | 7962089 | 7962089 | ? | Pres | ? | Pres | Pres | Pres |
| 7 | 134350004 | 134350007 | ? | Pres | Pres | Abs | ? | ./. |
| **7** | **143692540** | **143692532** | ? | Pres | ? | Pres | Pres | Pres |
| **7** | **1839523** | **1839514** | Pres | Pres | ? | Pres | ? | Pres |
| 7 | 2494553 | 2494548 | ? | Pres | ? | Pres | ? | Pres |
| 7 | 51714462 | 51714459 | ? | Pres | ? | Pres | ? | Pres |

Continues from the previous page

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **7** | **56524749** | **56524739** | Pres | Pres | ? | Pres | ? | Pres |
| **7** | **56652489** | **56652490** | Pres | Pres | ? | Pres | Pres | ./. |
| **7** | **67904284** | **67904277** | Pres | Pres | ? | Pres | ? | Pres |
| **7** | **990728** | **990720** | Pres | Pres | ? | Pres | ? | Pres |
| 8 | 109203879 | 109203879 | ? | Pres | ? | Pres | ? | Pres |
| 8 | 129710670 | 129710670 | ? | Pres | ? | Pres | ? | Pres |
| 8 | 130948948 | 130948944 | ? | Pres | ? | Pres | ? | Pres |
| 8 | 33026083 | 33026079 | ? | Pres | ? | Pres | ? | Pres |
| 8 | 66729987 | 66729989 | ? | Pres | ? | Pres | ? | Pres |
| **8** | **68356226** | **68356226** | Pres | Pres | ? | Pres | ? | Pres |
| 8 | 73210032 | 73210026 | ? | Pres | ? | Pres | ? | Pres |
| 8 | 92562537 | 92562537 | ? | Pres | ? | Pres | ? | Pres |
| **9** | **74877805** | **74877805** | Pres | Pres | ? | Pres | ? | Pres |
| **9** | **95093449** | **95093448** | Pres | Pres | ? | Pres | ? | Pres |
| **10** | **10493415** | **10493410** | ? | Abs | ? | Pres | Pres | Pres |
| **10** | **129946574** | **129946565** | Pres | Pres | ? | Pres | ? | Pres |
| 10 | 13258017 | 13258009 | ? | Pres | ? | Abs | ? | Pres |
| 10 | 33492650 | 33492641 | ? | Pres | ? | Pres | ? | Pres |
| 10 | 55958252 | 55958252 | ? | Pres | ? | Pres | ? | Pres |
| 10 | 58863548 | 58863550 | Pres | ./. | ? | ./. | ? | ./. |
| 10 | 74798370 | 74798372 | ? | Abs | ? | Abs | ? | Pres |
| 10 | 92815394 | 92815384 | ? | Pres | ? | Pres | ? | Pres |
| 11 | 104010483 | 104010474 | Pres | Abs | ? | Abs | ? | Pres |
| 11 | 109976712 | 109976712 | ? | Pres | ? | Pres | ? | Pres |
| 11 | 116712474 | 116712466 | ? | Pres | ? | Pres | ? | Pres |
| 11 | 12943734 | 12943733 | ? | Pres | ? | Abs | ? | Pres |
| **11** | **16223168** | **16223160** | Pres | Pres | ? | Pres | ? | Pres |
| 11 | 85211303 | 85211309 | ? | ./. | ? | Pres | ? | ./. |
| 12 | 127499179 | 127499179 | ? | Pres | ? | Pres | ? | Pres |
| **12** | **127656789** | **127656781** | Pres | Pres | ? | Pres | ? | Pres |
| 12 | 129806100 | 129806093 | ? | Pres | ? | Pres | ? | Pres |
| 12 | 27404920 | 27404920 | Pres | Pres | ? | Pres | ? | Pres |
| 12 | 41655279 | 41655269 | ? | Pres | ? | Pres | ? | Pres |
| 12 | 47043012 | 47043006 | ? | Pres | ? | Pres | ? | Pres |
| **12** | **8145512** | **8145502** | Pres | Pres | ? | Pres | ? | Pres |
| 13 | 107922171 | 107922171 | ? | Pres | ? | Pres | ? | Pres |
| 13 | 24067987 | 24067980 | ? | Pres | ? | Pres | ? | Pres |
| **13** | **29864458** | **29864448** | Pres | Pres | ? | Pres | ? | Pres |
| 13 | 29942076 | 29942068 | ? | Pres | ? | Pres | ? | Pres |
| 13 | 31156494 | 31156486 | ? | Pres | ? | Pres | ? | Pres |
| 13 | 42861425 | 42861420 | ? | Pres | ? | Pres | ? | Pres |
| **13** | **57967126** | **57967126** | Pres | Pres | Pres | Pres | Pres | Pres |
| 13 | 63643187 | 63643179 | ? | Pres | ? | Pres | ? | Pres |
| **13** | **79646444** | **79646444** | Pres | Pres | ? | Pres | ? | Pres |
| 13 | 96610409 | 96610401 | ? | Pres | ? | Pres | ? | Pres |
| 14 | 33426572 | 33426563 | ? | Pres | ? | Pres | ? | Pres |
| **14** | **35392013** | **35392003** | ? | Pres | ? | Pres | Pres | Pres |
| 14 | 39012372 | 39012364 | ? | Pres | ? | Pres | ? | Pres |
| 14 | 42748500 | 42748490 | ? | Pres | ? | Pres | ? | Pres |
| **14** | **52962602** | **52962605** | Pres | Pres | ? | Pres | ? | Pres |

Continues from the previous page

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **14** | **65518515** | **65518505** | Pres | Pres | ? | Pres | ? | Pres |
| **14** | **88885870** | **88885862** | ? | Pres | Pres | Pres | ? | Pres |
| 15 | 34884996 | 34884988 | ? | Pres | ? | Pres | ? | Abs |
| 15 | 55413974 | 55413969 | ? | Pres | ? | Pres | ? | Pres |
| 15 | 71966119 | 71966119 | ? | Pres | ? | Pres | ? | Pres |
| **15** | **75860056** | **75860048** | Pres | Pres | ? | Pres | ? | Pres |
| 16 | 23510964 | 23510964 | ? | Pres | ? | Pres | ? | Pres |
| 16 | 26606850 | 26606853 | ? | Pres | ? | Pres | ? | Pres |
| **16** | **33845731** | **33845721** | Pres | Pres | ? | Pres | ? | Abs |
| 16 | 49975479 | 49975472 | ? | Pres | ? | Pres | ? | Pres |
| 16 | 53854392 | 53854392 | ? | Pres | ? | Pres | Pres | ./. |
| **16** | **86242529** | **86242529** | Pres | Pres | ? | Pres | ? | Pres |
| 17 | 18791154 | 18791153 | ? | ./. | ? | ./. | ? | ./. |
| 17 | 19539953 | 19539944 | ? | Pres | ? | Pres | ? | Pres |
| 17 | 35367471 | 35367462 | ? | Pres | ? | Pres | ? | Pres |
| **17** | **58444186** | **58444178** | Pres | Pres | ? | Pres | ? | Pres |
| 17 | 64927840 | 64927840 | ? | Pres | ? | Pres | ? | Pres |
| 17 | 66948356 | 66948356 | ? | Pres | ? | Pres | ? | Pres |
| 17 | 68139473 | 68139464 | ? | Pres | ? | Pres | ? | Pres |
| 17 | 69849883 | 69849875 | ? | Pres | ? | Pres | ? | Pres |
| **17** | **74970107** | **74970104** | Pres | Pres | ? | Pres | ? | Pres |
| 18 | 21324513 | 21324504 | ? | Pres | ? | Pres | ? | Pres |
| 18 | 28676997 | 28676988 | ? | Pres | ? | Pres | ? | Pres |
| **18** | **40208090** | **40208081** | ? | Pres | Pres | Pres | ? | Pres |
| 18 | 47911019 | 47911019 | Pres | Abs | ? | Pres | Pres | ./. |
| 18 | 51727167 | 51727159 | ? | Pres | ? | Pres | ? | Pres |
| **18** | **68220119** | **68220112** | Pres | Pres | ? | Pres | ? | Pres |
| **18** | **6884091** | **6884081** | ? | Pres | Pres | Pres | ? | Pres |
| **18** | **77092147** | **77092137** | Pres | Pres | ? | Pres | ? | Pres |
| 19 | 11896382 | 11896372 | ? | Pres | ? | Pres | ? | Pres |
| 19 | 20887952 | 20887944 | ? | Pres | ? | Pres | ? | Pres |
| 19 | 21798635 | 21798628 | ? | Pres | ? | Pres | ? | Pres |
| 19 | 34630103 | 34630093 | ? | Pres | ? | Pres | ? | Pres |
| 19 | 44622970 | 44622966 | ? | Pres | ? | Pres | ? | Pres |
| 20 | 34530741 | 34530732 | ? | Pres | ? | Pres | ? | Pres |
| 20 | 46040353 | 46040344 | ? | Pres | ? | Pres | ? | Pres |
| 20 | 53530487 | 53530479 | ? | Pres | ? | Pres | ? | Pres |
| 20 | 54205727 | 54205718 | ? | Pres | ? | Pres | ? | Pres |
| 21 | 20342383 | 20342373 | ? | Pres | ? | Pres | ? | Pres |
| 21 | 22647478 | 22647470 | ? | Pres | ? | Pres | ? | Pres |
| 21 | 26825146 | 26825146 | ? | ./. | ? | ./. | ? | ./. |
| **21** | **36769314** | **36769307** | Pres | Pres | ? | Pres | ? | Pres |
| 21 | 37674076 | 37674071 | ? | Pres | ? | Pres | ? | Pres |
| 21 | 41945009 | 41945009 | ? | Pres | ? | Pres | ? | Pres |
| **21** | **42177624** | **42177614** | Pres | Pres | ? | Pres | Pres | Pres |
| **22** | **33665718** | **33665718** | ? | Pres | ? | Pres | Pres | Pres |
| X | 2133987 | 2133979 | ? | Pres | ? | Pres | ? | Pres |
| X | 44867443 | 44867437 | ? | Pres | ? | Pres | ? | Abs |
| **X** | **8195636** | **8195630** | Pres | Pres | ? | Pres | ? | Pres |

From the previous pages. **Supplementary Table S2.** The 198 polymorphic reference TEs in common between *ngs te mapper 2* and MELT-DEL. We compared transposable elements calls (i.e., presence or absence of the TE) and found that 64 elements are present in at least one sample according to both softwares (TEs highlighted in bold. One TE is also colored in yellow: it is the only element present in all three samples according to both softwares).

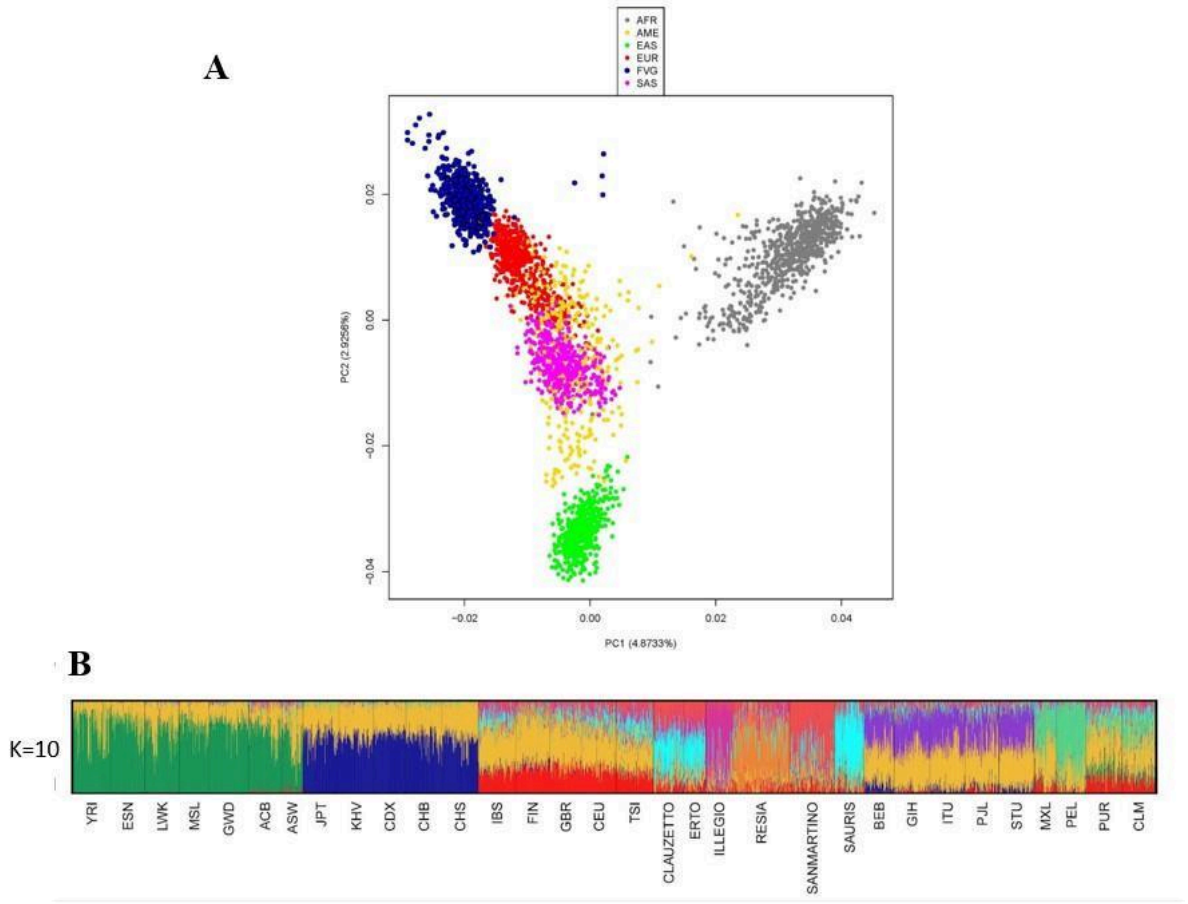
| CHR_POS | Type | Fst | Gene | Location | Fisher_allele | Fisher_geno |
|---|---|---|---|---|---|---|
| chr1_102454988 | AluYa4_5 | 2,581331011 | null | null | 0,003720064571 | 0,006106929124 |
| chr1_227275255 | AluYc1 | 3,296385936 | CDC42BPA | INTRONIC | 0,004689341044 | 0,01377986493 |
| chr1_74682371 | AluYc1 | 3,920654977 | null | null | 0,004583662983 | 0,0125275971 |
| chr1_78141383 | AluYa5 | 3,195664796 | GIPC2 | TERMINATOR | 0,00001175754147 | 0,0003467792578 |
| chr1_79116397 | AluYb9 | 2,731399319 | null | null | 0,0004594876263 | 0,004419344571 |
| chr10_59225161 | AluYa4_2 | 3,657007373 | PHYHIPL | INTRONIC | 0,00002244118112 | 0,0000173397739 |
| chr11_111029132 | AluYd | 2,262552349 | null | null | 0,0006363385996 | 0,000400260859 |
| chr11_26169546 | AluYb9 | 6,585969009 | null | null | 0,00008512214051 | 0,0002660516472 |
| chr11_45940658 | AluY | 4,457465092 | PHF21A | INTRONIC | 0,0003102610168 | 0,000008619566727 |
| chr12_126318397 | L1Ta | 5,353267379 | null | null | 0,000533810953 | 0,00036507081 |
| chr12_16054116 | AluYa | 2,174226654 | null | null | 0,008541329676 | 0,002484588543 |
| chr13_26913989 | L1Ambig | 2,181134736 | null | null | 0,009328209721 | 0,01635198832 |
| chr13_28811284 | AluY | 3,540971145 | null | null | 0,0000518446973 | 0,00003047760619 |
| chr18_47115523 | AluYa5 | 4,075337363 | HDHD2 | INTRONIC | 0,001252997957 | 0,002538619705 |
| chr2_137868315 | AluYg | 2,530582676 | null | null | 0,0006957678499 | 0,001042249434 |
| chr2_15379429 | AluYc1 | 4,222003766 | NBAS | INTRONIC | 0,00001769190101 | 0,0002526856506 |
| chr2_172657443 | AluYa | 2,600236998 | null | null | 0,008919143519 | 0,002040464798 |
| chr20_1565582 | AluYa4_2 | 2,901074197 | SIRPB1 | INTRONIC | 0,000009500992147 | 0,0001942957943 |
| chr20_28912248 | L1Ambig | 4,934481195 | null | null | 0,000154208187 | 0,00002305296 |
| chr21_24102304 | AluYb8 | 2,378529924 | null | null | 0,0003591000218 | 0,001866679057 |
| chr3_120524105 | AluY | 3,723155517 | null | null | 0,00001594245188 | 0,000006087094009 |
| chr3_122439516 | AluYc1 | 2,361067465 | KPNA1 | INTRONIC | 0,003636429339 | 0,002957506235 |
| chr3_29508893 | AluYb6_2 | 2,58148742 | RBMS3 | INTRONIC | 0,0009837050309 | 0,004513805552 |
| chr4_159798809 | AluYe | 5,353267379 | null | null | 0,001164963601 | 0,0008560572593 |
| chr5_144951450 | L1Ta | 5,519387202 | null | null | 0,000137345054 | 0,00170005426 |
| chr6_16950725 | L1Ambig | 2,926787252 | null | null | 0,002613353161 | 0,00707507442 |
| chr6_81458432 | AluY | 4,823626028 | null | null | 0,0002012910819 | 0,00000013632772 |
| chr7_154268965 | AluYc1 | 4,096354877 | DPP6 | INTRONIC | 0,000250374424 | 0,001188803351 |
| chr8_18085627 | AluYb8 | 4,161395121 | ASAH1-AS1 | INTRONIC | 0,008680090465 | 0,03058534682 |
| chr9_100940618 | AluYb3a1 | 2,620896073 | null | null | 0,001639069642 | 0,005661979604 |
| chr9_122657291 | L1Ambig | 3,736515487 | OR1L1 | PROMOTER | 0,00025421156 | 0,00036256005 |
| chr9_78833763 | AluYb8 | 2,474242704 | null | null | 0,007017872413 | 0,002993334911 |

**Supplementary Table S3.** The 32 transposable elements that emerge as able to discriminate between high-altitude and both middle-altitude and low-altitude populations, according to fixation index (Fst) and differential allele frequencies. “CHR_POS” = chromosome and position (hg38) of the TE; “Type” = TE family/subfamily; “Fst” = fixation index score; “Gene” = gene mapped by the polymorphic TE; “Location” = region in which the TE is located (null = intergenic); “Fisher_allele” and “Fisher_geno” = Fisher p-value for allele and genotype frequencies, respectively.

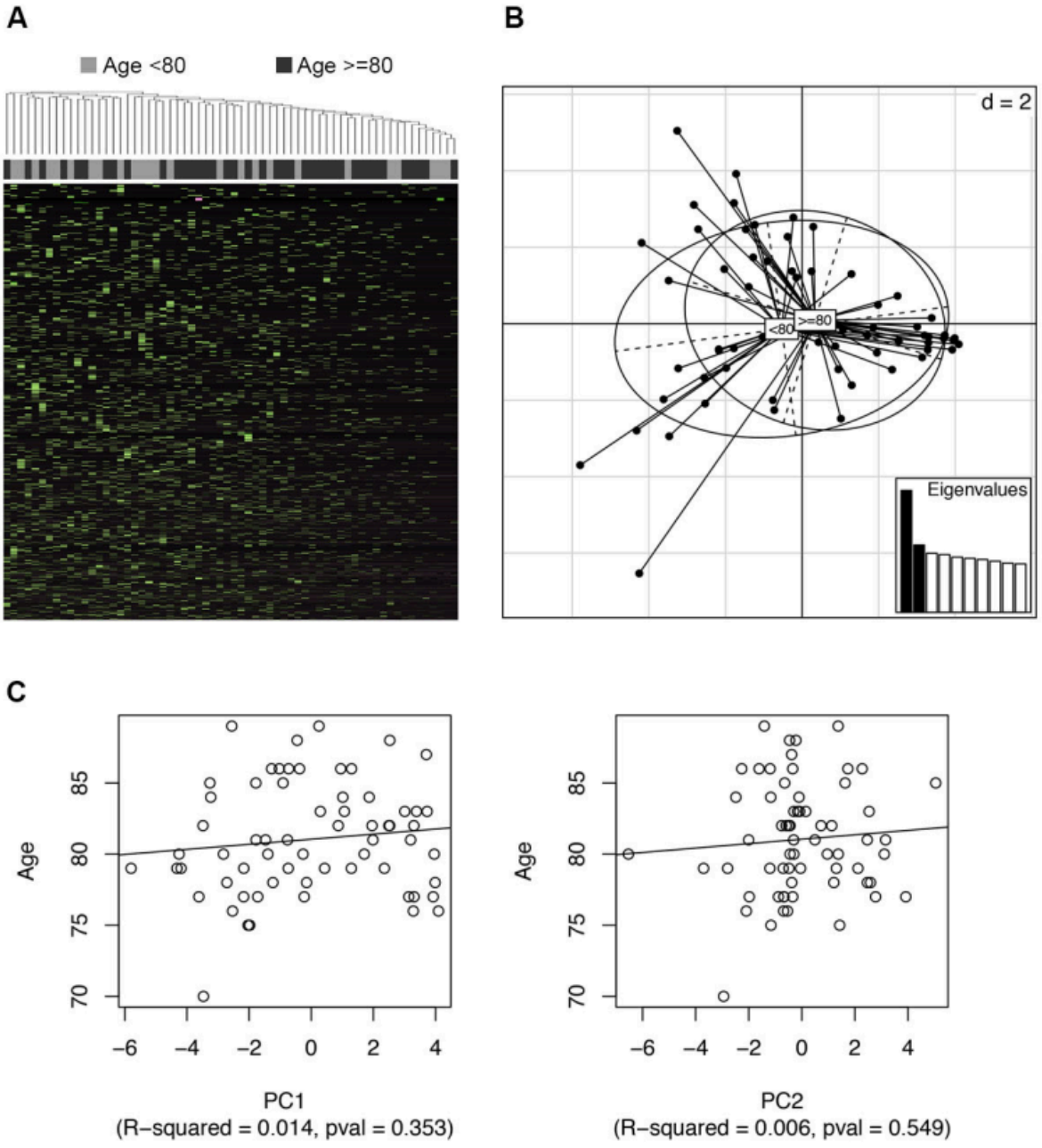
| Chr_pos | PBS | Gene | Location | Type | INS/DEL |
|---|---|---|---|---|---|
| chr1_237405908 | 0,242959 | RYR2 | INTRONIC | AluYa5 | INS |
| chr1_2994973 | 0,309109 | intergenic | | ALU | DEL |
| chr1_74836001 | 0,192827 | null | null | AluYa5a2 | INS |
| chr1_82497609 | 0,308435 | intergenic | | ALU | DEL |
| chr10_113490527 | 0,141907 | null | null | AluYb6_2 | INS |
| chr10_120381091 | 0,172093 | null | null | AluYa | INS |
| **chr11_45940658** | **0,096755** | **PHF21A** | **INTRONIC** | **AluY** | **INS** |
| chr12_105437127 | 0,226608 | null | null | AluY | INS |
| chr12_11105926 | 0,378573 | PRR4 | | ALU | DEL |
| chr12_122525792 | 0,115705 | RSRC2 | | ALU | DEL |
| chr12_126405333 | 0,250548 | null | null | AluYa5 | INS |
| chr12_43919859 | 0,199453 | TMEM117 | INTRONIC | SVA | INS |
| chr12_44670594 | 0,168776 | NELL2 | INTRONIC | AluYe | INS |
| **chr13_28811284** | **0,127321** | **null** | **null** | **AluY** | **INS** |
| chr13_92448897 | 0,102364 | GPC5 | INTRONIC | AluYa | INS |
| chr14_38691565 | 0,147848 | null | null | AluYa4_2 | INS |
| chr14_64710808 | 0,140571 | PLEKHG3 | INTRONIC | AluY | INS |
| chr14_81757125 | 0,113418 | null | null | AluYb7_3 | INS |
| chr14_83339277 | 0,110438 | null | null | AluYa5 | INS |
| chr15_30847459 | 0,54915 | intergenic | | ALU | DEL |
| chr15_51605505 | 0,101694 | DMXL2 | | ALU | DEL |
| chr15_67002522 | 0,140785 | null | null | AluYa4_2 | INS |
| chr16_49945486 | 0,24016 | intergenic | | ALU | DEL |
| chr17_37809362 | 0,321961 | LOC105371757 | | ALU | DEL |
| chr17_83038994 | 0,143938 | B3GNTL1 | INTRONIC | AluY | INS |
| chr18_75591054 | 0,123428 | intergenic | | ALU | DEL |
| chr18_76241984 | 0,249793 | null | null | AluYi | INS |
| chr18_77044929 | 0,140985 | MBP | INTRONIC | AluYb8 | INS |
| chr19_21394832 | 0,123516 | intergenic | | ALU | DEL |
| chr19_21658740 | 0,238356 | null | null | SVA | INS |
| **chr2_137868315** | **0,108998** | **null** | **null** | **AluYg** | **INS** |
| chr2_16110982 | 0,192068 | intergenic | | ALU | DEL |
| chr2_192117971 | 0,253204 | TMEFF2 | INTRONIC | AluYa5 | INS |
| chr2_40751000 | 0,129244 | LINC01794 | | ALU | DEL |
| chr2_46682994 | 0,408802 | intergenic | | ALU | DEL |
| chr2_64936649 | 0,225397 | LINC02245 | PROMOTER | AluYe | INS |
| chr20_17880293 | 0,273156 | null | null | L1Ta1d | INS |
| chr22_34784759 | 0,140118 | null | null | AluYa5 | INS |
| chr3_11658649 | 0,1031 | VGLL4 | INTRONIC | AluYa4_5 | INS |
| **chr3_120524105** | **0,181747** | **null** | **null** | **AluY** | **INS** |
| chr3_153924853 | 0,225229 | null | null | AluYb | INS |
| chr3_76657140 | 0,450974 | ROBO2 | | ALU | DEL |
| chr4_131390572 | 0,165657 | LINC02377 | | ALU | DEL |
| chr4_139016982 | 0,171332 | NOCT | INTRONIC | AluYg | INS |
| chr4_97189494 | 0,389228 | STPG2 | | ALU | DEL |

| | | | | | |
|---|---|---|---|---|---|
| chr4_99819712 | 0,181747 | DAPP1 | INTRONIC | AluYa4_2 | INS |
| chr5_143555802 | 0,207367 | LOC105378208 | | ALU | DEL |
| **chr5_144951450** | **0,168312** | **null** | **null** | **L1Ta** | **INS** |
| chr5_63408612 | 0,098887 | intergenic | | ALU | DEL |
| chr6_120942545 | 0,156681 | intergenic | | ALU | DEL |
| chr6_45292742 | 0,219776 | SUPT3H | INTRONIC | AluYe | INS |
| chr6_51391489 | 0,098367 | intergenic | | ALU | DEL |
| **chr6_81458432** | **0,202414** | **null** | **null** | **AluY** | **INS** |
| chr7_104215987 | 0,168286 | intergenic | | ALU | DEL |
| chr7_128669158 | 0,297931 | intergenic | | ALU | DEL |
| chr7_141414064 | 0,10577 | TMEM178B | | ALU | DEL |
| chr8_12671195 | 0,093695 | intergenic | | ALU | DEL |
| chr8_15691011 | 0,119551 | TUSC3 | INTRONIC | AluYb8 | INS |
| **chr8_18085627** | **0,177424** | **ASAH1-AS1** | **INTRONIC** | **AluYb8** | **INS** |
| chr8_81350723 | 0,200541 | null | null | AluYa | INS |
| **chr9_100940618** | **0,132969** | **null** | **null** | **AluYb3a1** | **INS** |
| chr9_133719816 | 0,110438 | SARDH | | ALU | DEL |

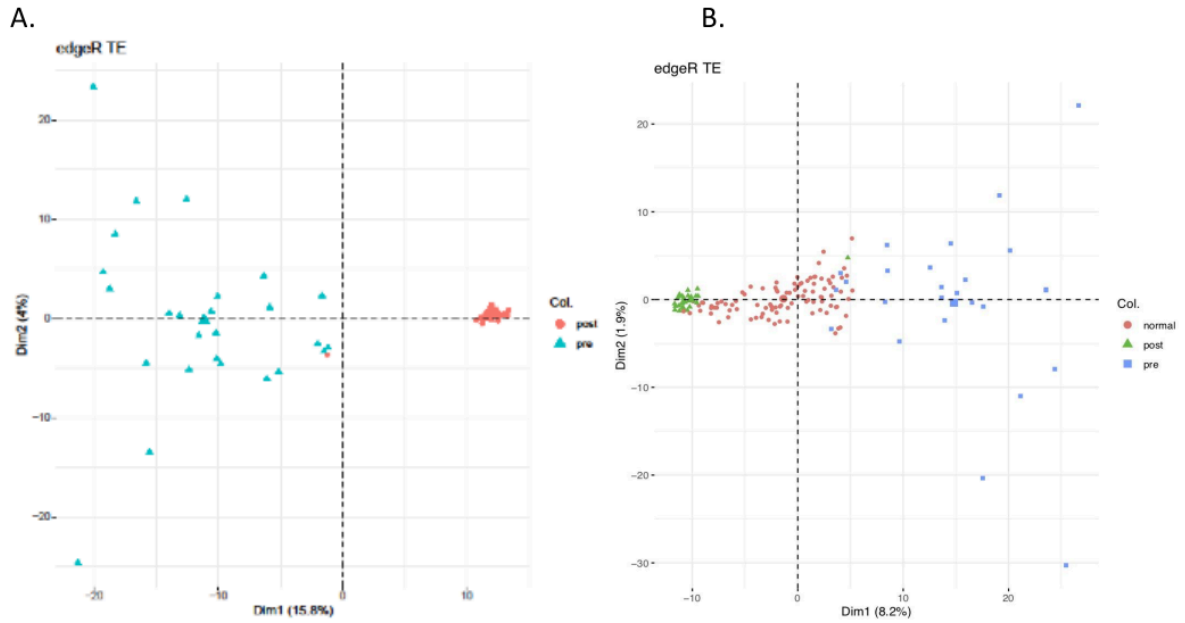
(Continues from the previous page). **Supplementary Table S4.** The 62 polymorphic TEs that are characteristic of differentiation in the direction of the high-altitude group, according to population branch statistics (PBS). 8 TEs are in common with the Fst analysis (bold lines). “PBS” = population branch statistics score; “INS/DEL” = if the TE is a non-reference (INS) or a reference (DEL) element.




**Supplementary Figure S1.** A) Principal Component Analysis (PCA) with Friuli-Venezia Giulia isolates and populations from the 1000 Genomes Project, divided by macro-areas: blue = FVG, red = Europe, pink = South Asia, yellow = America, green = East Asia, gray = Africa. B) Admixture plot at K=10 with all the 26 populations from 1KGP and the six isolates of FVG.




**Supplementary Figure S2.** A) shows the absence of any clustering in the heatmap that compares "young" with "old" normal controls. B) shows the results of a PCA analysis for these 2 groups considering the 1,790 DE TEs presenting an equal dispersion of subjects across PC1 and PC2. C) shows a not significant result for age and both PC1 and PC2 in a linear model.




**Supplementary Figure S3.** A) PCA of edgeR significant TEs using only ConverterPre and ConverterPost individuals. B) PCA of edgeR significant TEs using ConverterPre, ConverterPost and NC individuals.