ALMA MATER STUDIORUM – UNIVERSITÀ DI BOLOGNA

Dottorato di Ricerca in Scienze della Terra, della Vita e dell'Ambiente

Curriculum Biologico – Ciclo XXX

Settore Concorsuale: 05/B1 - Zoologia e Antropologia

Settore Scientifico Disciplinare: BIO/08 - Antropologia

Impact of non-LTR retrotransposons in the differentiation and evolution of Anatomically Modern Humans

Tesi presentata da:

Etienne Guichard

Coordinatore Dottorato:

Prof. Giulio Viola

Supervisore:

Dr. Alessio Boattini

Esame Finale anno 2018

INDEX OF CONTENTS

ABSTRACT

Transposable Elements are biologically important components of eukaryote genomes. In particular, non-LTR retrotransposons (N-LTRrs) extensively shaped the human genome throughout evolution. In this study, we compared retrotransposon insertions differentially present in the genomes of Anatomically Modern Humans, Neanderthals, Denisovans and Chimpanzees, in order to assess the possible impact of retrotransposition in the differentiation of the human lineage. Briefly, we first identified species-specific N-LTRrs and established their distribution in present day human populations. These analyses shortlisted a group of N-LTRr insertions that were found exclusively in Anatomically Modern Humans. Notably, these insertions targeted genes more frequently than randomly expected and are associated with an increase in the number of transcriptional/splicing variants of those genes they inserted in. The analysis of the functionality of genes targeted by human-specific N-LTRr insertions seems to reflect phenotypic changes that occurred during human evolution. Furthermore, the expression of genes containing the most recent N-LTRr insertions is enriched in the brain, especially in undifferentiated neurons, and these genes associate in networks related to neuron maturation and migration. Additionally, we also identified candidate N-LTRr insertions that have likely produced new functional variants exclusive to modern humans, which show traces of positive selection and are now fixed in all present-day human populations. In sum, our results strongly suggest that N-LTRr impacted our differentiation as a species and have been a constant source of genomic variability all throughout the evolution of the human lineage.

INTRODUCTION

Modern and archaic humans

Anatomically Modern Humans (AMH) appeared in Africa at least 200 thousands of years ago (kya), although recent studies antedate them to about 350 kya (Hublin et al. 2017, Schlebusch et al. 2017). AMH went out of Africa between 80-130 kya (Out of Africa II) through the Middle East, expanding towards Europe, Asia and later Southeast Asia, Oceania and the Americas.

Other *Homo* species had already migrated out of Africa before *Homo sapiens* appearance (Out of Africa I): *H. erectus* in Asia, *H. floresiensis* in Southeast Asia, *H. neanderthalensis* and *Denisova* in Europe and Siberia.

Homo neanderthalensis (HN) and *Denisova* (HD) are the closest extinct relatives to AMH (Figure 1). They are sister groups, being more closely related to each other than they are to *Homo sapiens*. Their split from the modern human lineage is estimated to have occurred between 550 thousands of years ago (kya) and 765 kya, after which they colonized Eurasia long before Anatomically Modern Humans (AMH) left Africa. The population split between these archaic populations is estimated at 381-473 kya (Prüfer et al. 2014). However, after AMH colonized Eurasia and encountered HN and HD, gene flow seems to have occurred between the different species: portions of HD- and HN-derived DNA are in fact present in modern human populations.

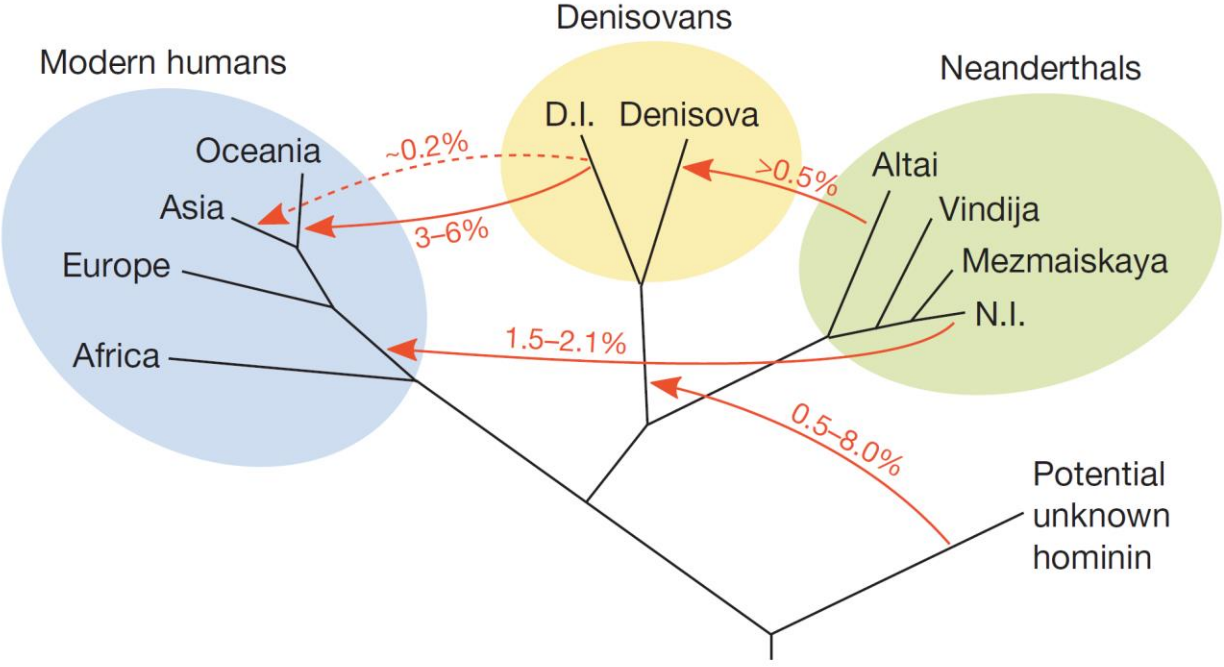


Figure 1. Phylogenesis and possible gene-flow events between AMH, HN and HD (Prüfer et al. 2014).

This gene flow seems to have occurred in both directions: in fact, HN mitochondrial DNA is closer to that of AMH than it is to that of HD. This is explained with an introgression of *Homo sapiens* DNA into the European Neanderthal populations that occurred more than 100 kya (Posth et al. 2017).

Neanderthals seem to have appeared about 500 kya in Europe: their remains have been found in Spain, France, Germany (Neander Valley, from where the species gets its name) (Krings et al. 1997; Schmitz et al. 2002; Sánchez-Quinto and Lalueza-Fox, 2015) and Croatia (Vindija cave) (Kuhlwilm et al. 2016; Ahern et al. 2005); at the same time they were also present in Asia, as testified by remains found in Russia (Mezmaiskaya and Denisova caves) (Pinhasi et al. 2011; Briggs et al. 2009). HN characteristic cranial and post-cranial morphology prompted its classification as a new species. Despite their strong physical features and their remarkable cranial volume (about 1500-1600 cc, slightly larger than *Homo sapiens*) (Stringer 1984; Holloway 1985), Neanderthals probably did not experience a cultural evolution such as the one that occurred for *Homo sapiens* (Pearce et al. 2013).

Denisovans left far less fossil traces than Neanderthals: a distal phalanx of the hand and two molars (dated ~41.000 years ago), all found in the same archaeological site, the Denisova cave in southern Siberia (Sawyer et al. 2015). This is the same cave where the fossil remains of Altai Neanderthals have been found.

Notably, the genomes of some individuals belonging to HN and HD have been previously sequenced, assembled and published (Green et al. 2010; Meyer et al. 2012; Prüfer et al. 2014; Sawyer et al. 2015), two of which are high-coverage: one HD and one HN, both from the Altai mountains in southern Siberia (Meyer et al. 2012; Prüfer et al. 2014). This can allow for molecular comparisons of modern and archaic human DNA.

The human genome and Transposable Elements

The human genome is a complex structure composed of many different types of elements. Only roughly 1.5% of it accounts for protein-coding genes. The vast majority of our genetic material is made of non-genic sequences, such as Satellite DNA and Transposable Elements (Lander et al. 2001).

Transposable Elements (TEs) were discovered in the mid-1940s by Barbara McClintock and are DNA sequences that are able to move or replicate in genomes (Richardson et al. 2015). Although TEs have for long been dismissed as "selfish", "parasites" or simply "junk DNA" (Orgel and Crick, 1980; Doolittle and Sapienza, 1980), the advent of whole genome DNA sequencing, in conjunction with molecular genetic, biochemical, genomic and functional studies, has revealed that TEs are biologically important components of mammalian genomes whose activity has extensively shaped the structure and function of our own genome (Richardson et al. 2015). TEs are known to be involved in several evolutionary and adaptive processes such as the generation of genes and pseudogenes (Ohshima et al. 2003; Moran et al. 1999; Sayah et al. 2004), fine-tuning transcriptional regulation of genes (Speek 2001; Han et al. 2004; Chuong et al. 2017), generation of somatic mosaicism (Bailie et.al. 2011; Muotri et al. 2005; Evrony et al. 2012), the increase in complexity and evolution of gene regulatory networks (Feschotte 2008) and the alteration of epigenetic mechanisms as processes of fine-scale and reversible regulation (Fedoroff 2012). Some of the most notable biological processes associated with the domestication of TE-derived sequences are the insurgence of the V(D)J system of acquired immunity (Kapitonov and Jurka, 2005; Koonin and Krupovic, 2014; Huang et al. 2016) and placental development (Lynch et al. 2011, Lavialle et al. 2013), but they also play key roles in embryogenesis (Gerdes et al. 2016; Friedli and Trono, 2015) and neurogenesis (Notwell et al. 2015; Muotri et al. 2005; Evrony et al. 2012). In sum, in addition to their role in growing the size of eukaryotes' genomes, active TEs are continually impacting the functioning and evolution of genomes.

Notably, the activity of TEs throughout evolution has generated more than two thirds of the human genomic material (Lander et al. 2001; Kapusta et al. 2013).

TEs can be divided in two categories based on their mobilization mechanism and functionality (Figure 2): DNA transposons that move via a cut-and-paste mechanism, being cleaved from their genomic site and inserting in a new target site, and retrotransposons, which are transcribed in a RNA intermediate that is then reverse-transcribed and inserted in the new target site, thus duplicating in a copy-and-paste like mechanism. Retrotransposons are grouped into two classes, depending on the presence or absence of a Long Terminal Repeat (LTR). LTR retrotransposons mimic modern retroviruses and have a similar mobilization mechanism. Non-LTR retrotransposons, instead, mobilize with a unique mechanism and are the most abundant TEs in our genome.

In modern humans, only a limited number of TE subfamilies from the non-LTR-retrotransposon (N-LTRr) class are currently active, i.e. Long and Short INterspersed elements (LINEs and SINEs). Indeed, the ongoing activity of LINEs and SINEs in humans offers a constant source of inter-individual variability in human populations and can sporadically generate new genetic disorders (Kazazian et al. 1988; Richardson et al. 2015).

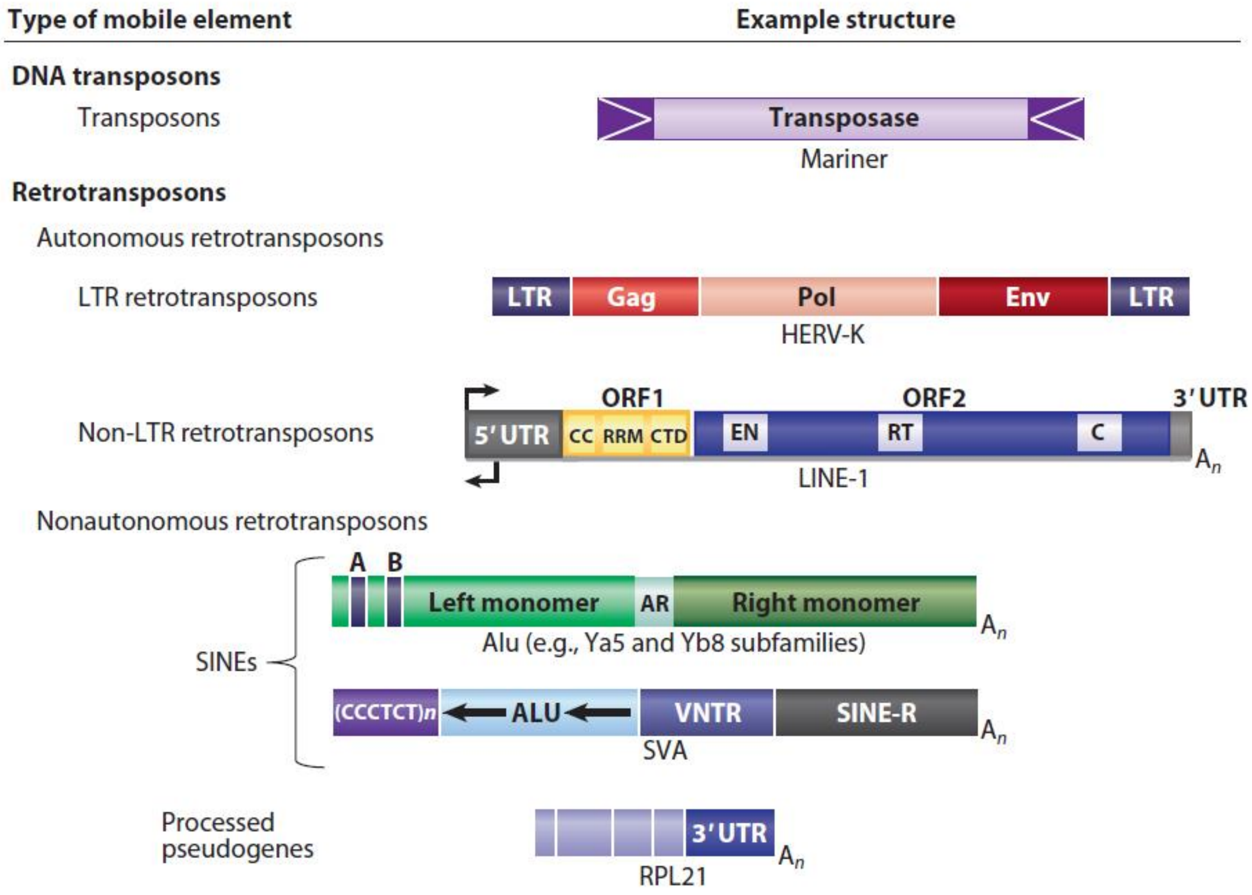


Figure 2. Classification of TEs present in the human genome (Beck et al. 2011).

Active non-LTR retrotransposons in the human genome

LINE-1s (or L1s) are the most successful and abundant (bp-wise) retrotransposons in the human genome: they comprise at least 17% of our nuclear DNA with more than 500.000 copies (Richardson et al. 2015). Despite being the only active autonomous retrotransposons in our genome, the overwhelming majority of L1s are retrotransposition defective (RD-L1) and cannot move because they are 5’ truncated, internally rearranged, or otherwise mutated (Lee et al. 2007). Full-length Retrotransposition-competent L1s (RC-L1s) are abundant, however only 80 to 100 copies of L1s on average are actually active in the human genome (Wei et al. 2001).

A full-length LINE-1 element is around 6kb in length (Dombroski et al. 1991; Scott et al. 1987) and contains a ~900bp long 5' untranslated region (5' UTR) with internal promoter activity (Swergold 1990), two open reading frames (ORFs), a ~150bp long 3' UTR, and a poly(A) tail (Scott et al. 1987). ORF1 encodes for a ~40kDa protein with RNA binding and nucleic acid chaperone activities (Hohjoh and Singer 1997; Kulpa and Moran 2005). ORF2 encodes for a ~150kDa protein with reverse transcriptase (RT) and endonuclease (EN) activities (Feng et al. 1996; Mathias et al. 1991). Both proteins are required for the mobilization of L1 within the human genome (Moran and Gilbert 2002).

SINEs are non-autonomous elements that are L1 retrotransposition-dependent. The two main families of SINEs, in humans, are Alus and SVAs. Alus are, by far, the most common (copies-wise) retrotransposon in our genome. The typical full-length Alu element is ~300 bp long and has a dimeric structure determined by the fusion of two 7SL-RNA-derived monomers (Kriegs et al. 2007); those two monomers are separated by an A-rich linker region. The Alu 5’ region contains an internal RNA polymerase III promoter and the element ends with an oligo dA-rich tail of variable length.

SVAs are hominoid-specific retrotransposons and are named after the three components of their sequence: an antisense Alu-like region, a Variable Number Tandem Repeats (VNTR) region and the 3’LTR of the endogenous retrovirus HERV-K10 (SINE-R) (Wang et al. 2005); they also terminate in a poly(A) tail. The VNTR region is composed of a variable number of copies of a 35–50 bp sequence, 490 bp in total on average, that is derived from the 3’ end of the ENV gene. More than half of the SVA elements in human genomes are full-length (~2kb), but may vary in size as a result of polymorphisms in their VNTR region copy number (48-2306bp) (Wang et al. 2005). It seems that they do not have an internal promoter, but they could have a sequence with promoter activity within the flanking portion of element (Wang et al. 2005).

Retrotransposition

Retrotransposition (Figure 3) starts with the generation of a full-length L1 mRNA from a preexisting L1 in the genome that is subsequently translated to generate a RiboNucleoprotein Particle (RNP) (Moran and Gilbert 2002; Deininger et al. 2003). The L1 RNP is a retrotransposition intermediate that contains both the L1 proteins and their encoding mRNA, associated in *cis*. Indeed, this preferential association, or *cis*-preference, likely represents a mechanism whereby L1 has ensured its evolutionary success as RC-L1 RNPs will most frequently mobilize RC-L1 mRNAs (Moran and Gilbert 2002; Moran et al. 1996; Wei et al. 2001; Goodier and Kazazian 2008). In some cases, however, ORF1 and ORF2 proteins can bind other cellular mRNAs (including Alu and SVA mRNAs), thus mobilizing them in *trans* (Ohshima et al. 2003; Beck et al. 2011); this way, non-autonomous retrotransposons are able to retrotranspose by utilizing L1-encoded proteins.

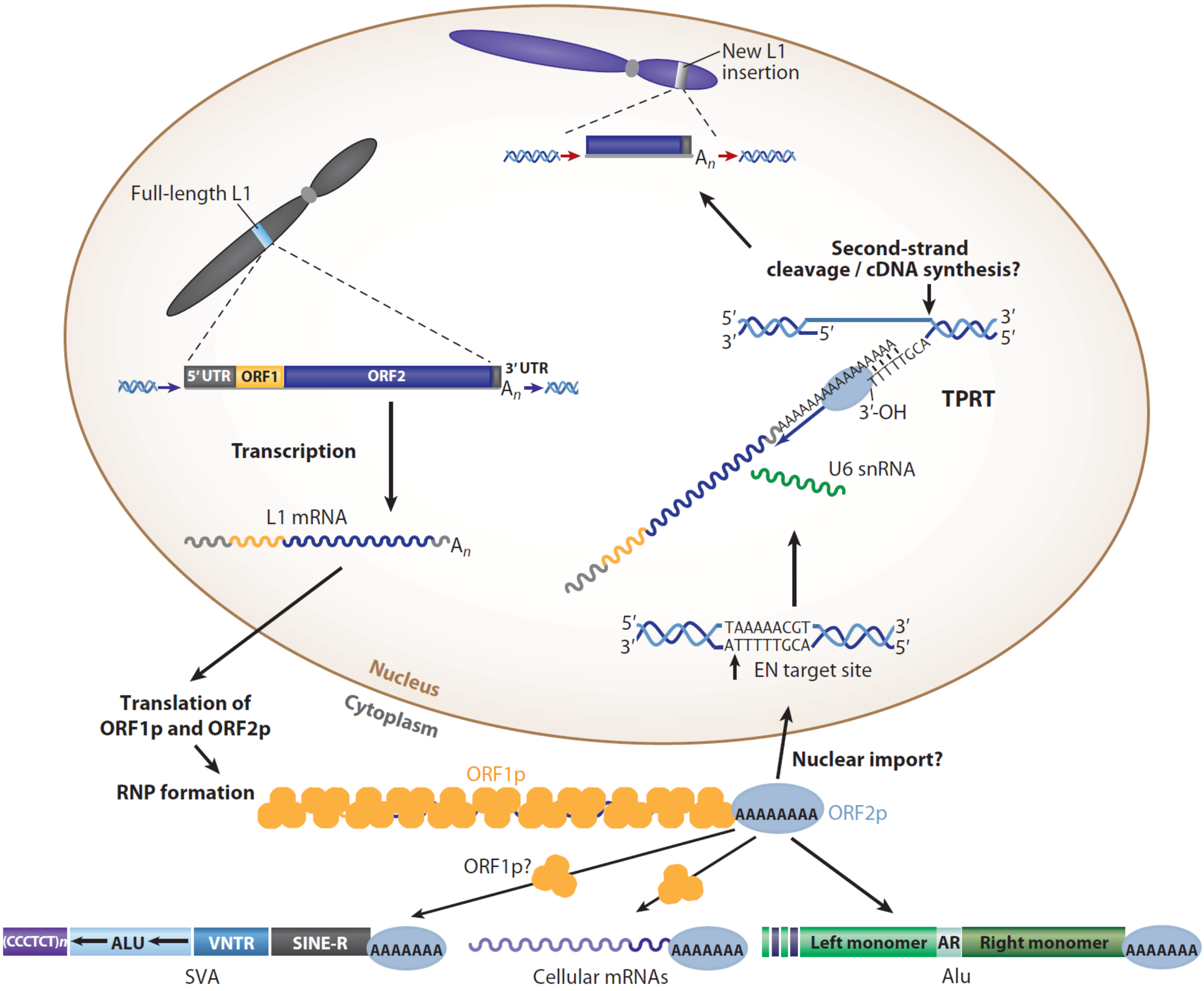


Figure 3. Mechanism of rerotransposition (Beck et al. 2011).

The RNP is then re-introduced in the nucleus where retrotransposition takes place with a process known as Target-site Primed Reverse Transcription (TPRT). In TPRT, the L1-encoded EN recognizes and cleaves a loose consensus sequence, liberating a free 3'OH that is used by the L1-encoded RT to prime first strand cDNA synthesis (Moran and Gilbert 2002; Goodier and Kazazian 2008; Beck et al. 2011). It is thought that the same activities are involved in second strand cDNA synthesis, resulting in a new N-LTRr inserted elsewhere in the genome.

Notably, most de novo N-LTRr insertions are flanked by variable size Target Site Duplications (TSDs) and L1 insertions are often 5'-truncated (Moran and Gilbert 2002; Goodier and Kazazian 2008). Thus, it is possible that 5' truncation could represent a host defense mechanism that reduces the number of potentially active L1s in a genome by aborting reverse transcription before the element can be copied in full.

Host defense mechanisms

Retrotransposition and the cellular DNA repair machinery are known to interact at least in cell culture retrotransposition assays (Morrish et al. 2002; Morrish et al. 2007; Suzuki et al. 2009); various host factors seem to be also involved in discrete stages of retrotransposition and the 5' truncation process (during or immediatly after TPRT), such as the polymerase-delta-associated sliding DNA clamp PCNA, key nonsense-mediated decay factor UPF1 and interactors PABPC1 and MOV10 (Taylor et al. 2013).

N-LTRrs are regarded as potentially very dangerous for their host genomes because of their high mutagenic potential. Their activity has been previously associated with apoptosis, DNA damage and repair, tumor progression, cellular plasticity, and stress response. Currently, 124 insertions of L1s, Alus, and SVAs in humans are known to be disease-causing in the germline (Goodier 2016). Studies have characterized several defense mechanisms that act against N-LTRrs mobility other than the common 5’ truncation of new integrants, such as APOBEC proteins (Wissing et al. 2011), the exonuclease Trex1 (Stetson et al. 2008), Piwi proteins and Piwi-interacting RNAs (Malone and Hannon 2009), DNA methylation (Bourc’his and Bestor 2004; Yoder et al. 1997), the Aicardi-Goutières syndrome gene product SAMHD1 (Zhao et al. 2013) and small RNAs generated by the antisense L1 promoter (Yang and Kazazian 2006). These host mechanisms act to control currently active N-LTRrs and likely could have contributed to reduce the mobility of previously active subfamilies, but may require co-evolutionary fine-tuning to control the mobilization of future sub-families (Muñoz-Lopez et al. 2011).

However, these researches do not explain the proliferation of N-LTRrs which, together with the aforementioned impact that insertions can have on their target loci, seems to point out at a possible functional role for N-LTRrs in their host genomes.

Aims of the study

In this study, we aimed to explore the role of N-LTRrs in the differentiation and evolution of the genus *Homo*. In order to do that, we compared the repertoire of Retrotransposon Insertions (RIs) present in the genome of modern humans with those present in modern-day chimpanzees, as well as with those of our closest extinct relatives, Neanderthals and Denisovans.

Although often discussed and speculated, the effects and implications of RIs on the evolution of the human lineage are mostly unknown. Here, and for the first time, we evaluated the potential impact of RIs in our species differentiation; additionally, we have also analyzed how RIs that are specific to AMH and absent in HN, HD and chimpanzees could have affected the genomic loci surrounding them. In order to shed light on the molecular dynamics of AMH differentiation and evolution, we aimed at characterizing Rl locations, identifying potential selective pressures and inferring functional/regulatory alterations that might have occurred as a consequence of species-specific RIs. Thus, the reconstruction of the mechanisms through which retrotransposition impacted the evolution of the human lineage can allow for a better understanding of how our genome is evolving in real time.

METHODS

RI identification between AMH and HN/HD

Available RI identification tools such as RetroSeq (Keane et al. 2012), Tangram (Wu et al. 2014), Tea (Lee et al. 2012), MELT (Gardner et al. 2017), etc. are primarily based on mapping paired-end DNA-sequencing reads. However, and given that a large portion of previously sequenced ancient DNAs is composed of single-ended reads, here we devised a methodology for detecting differentially present RIs in AMH, HD and HN genomes based on single-ended reads (Figure 4). In particular, our methodology is intended to identify confirmed species-specific RIs upon which to infer the impact that RIs might have had in our species differentiation. The methodology uses well-known bioinformatic tools such as the BLAST+ package (Camacho et al. 2008), ABySS (Simpson et al. 2009), BEDTools (Quinlan and Hall, 2010) and RepeatMasker (Smit et al. 2013-2015), implementing them with custom R or Perl scripts for filtering, conversion and general data management.

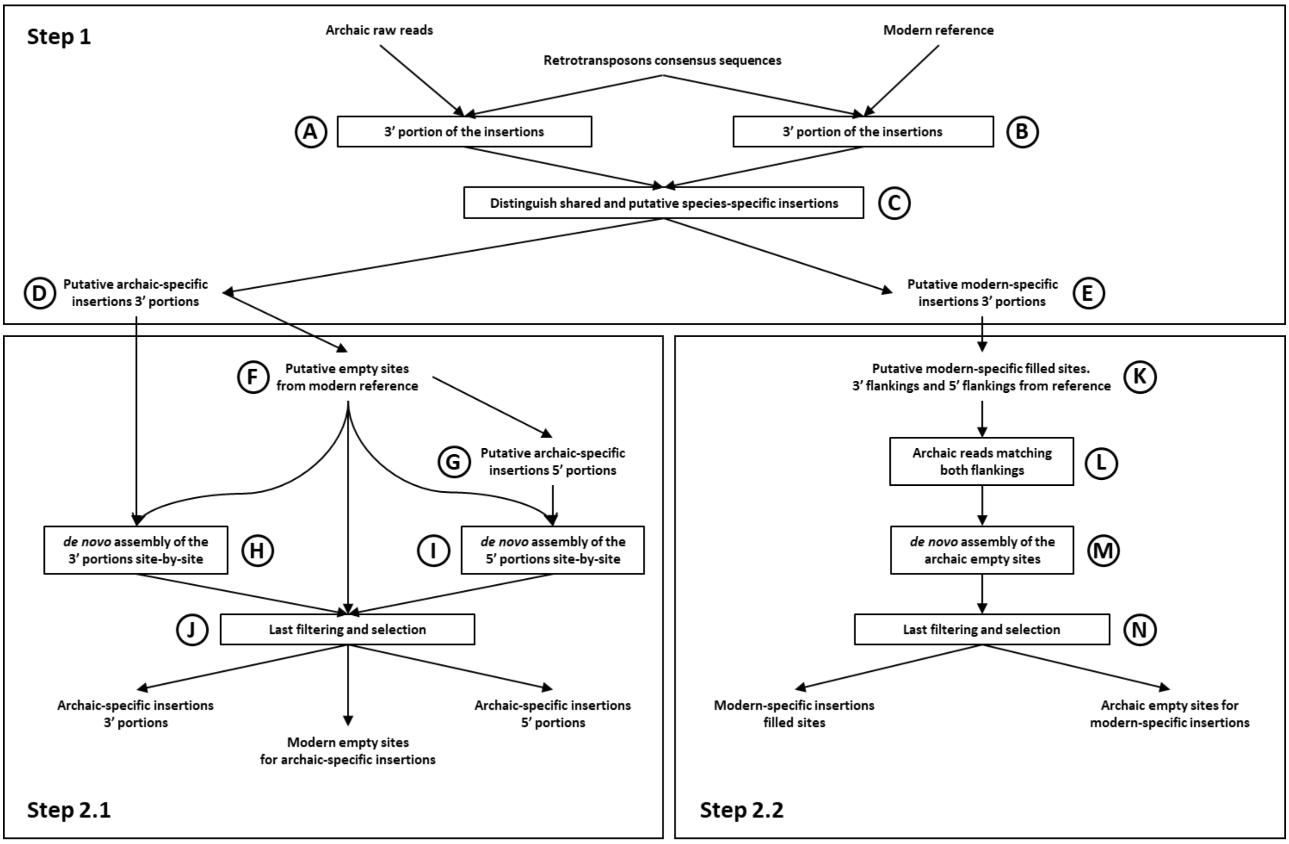

Step 1
Archaic raw reads
Modern reference
Retrotransposons consensus sequences
A
3' portion of the insertions
3' portion of the insertions
B
Distinguish shared and putative species-specific insertions
C
D
Putative archaic-specific insertions 3' portions
Putative modern-specific insertions 3' portions
E
F
Putative empty sites from modern reference
G
Putative archaic-specific insertions 5' portions
de novo assembly of the 3' portions site-by-site
H
I
de novo assembly of the 5' portions site-by-site
J
Last filtering and selection
Archaic-specific insertions 3' portions
Modern empty sites for archaic-specific insertions
Archaic-specific insertions 5' portions
Step 2.1
Putative modern-specific filled sites. 3' flankings and 5' flankings from reference
K
Archaic reads matching both flankings
L
de novo assembly of the archaic empty sites
M
Last filtering and selection
N
Modern-specific insertions filled sites
Archaic empty sites for modern-specific insertions
Step 2.2

Figure 4. Schematic representation of the methodology used for RI identification between pairs of compared species.

The methodology was designed on a simulated dataset composed of 100 random locations in the human reference genome GRCh37-hg19, bot genic and intergenic. In 50 of the 100 random loci, an RI was artificially added simulating a non-reference retrotransposition event, while in the other 50 an existing reference RI was artificially removed reconstructing an empty (pre-insertional) site. RI artificially added or removed accounted for Alu, LINE-1 and SVA elements, both full length and truncated. All the thresholds used during the procedure (most notably blastn ones) were defined in order to retrieve all simulated insertions. After successfully completing the simulated procedure, the same methodology and thresholds were optimized and finally applied to real genomic data.

Step 1) We retrieved consensus sequences of the most recent non-LTR retrotransposons from RepBase (Bao et al. 2015) (AluYa5, AluYb8, AluYb8a1, AluYb9, AluYb10, AluYb11, AluYk13, LINE-1HS, SVA_A, SVA_B, SVA_C, SVA_D, SVA_E, SVA_F), as well as the genomic material of the species to compare (reference sequence GRCh37-hg19 for AMH, the raw reads of the DNA sequencing for both HN and HD). Specifically, the genomes analyzed in this study are those of two individuals, a Neanderthal and a Denisovan, who lived in the Denisova cave at different times (Meyer et al. 2012; Prüfer et al. 2014). These two genomes were selected for their relatively high coverage and ready availability. The selected retrotransposon sequences were identified in both genomes using blastn (A,B), setting the identity parameter to 95%. This was done in order to allow the identification of retrotransposons diverging as much as 5% from their consensus sequence. Because of the repetitive nature of TEs, each insertion has been associated to its unique genomic target. For AMH, this was done by extending TEs matches by 100 bp in the 3’ direction and in the 5’ direction in the reference sequence (3’ and 5’ flanking sequences). The same could not be done for the archaic DNA, having reads averaging 100 bp as a starting point. Many new retrotransposon insertions are 5’ truncated, thus the length and 5’ end of an insertion is not known beforehand. For this reasons, we implemented a custom R script to select the reads that matched at least 30 bp of the 3’ end of the retrotransposon’s sequence and that had at least 30 bp of flanking sequence in the 3’ direction. In order to take account of differential length of the insertions poly-A tails in respect to the consensus sequences we allowed for 25 bp of margin between insertions and flankings. The sequences of the 3’ ends of insertions with their respective flankings were then compared between the two species using blastn, with identity parameter set to 95% (C). Sequences that were present in one species' DNA and not in the other were selected as putatively species-specific insertions, thus producing two lists: putative archaic-specific insertions 3' portions (D) and putative modern-specific insertions 3' portions (E).

Step 2.1) The 3' flanking portions of the putative-archaic specific insertions were used to identify their respective "empty" (pre-insertional) sites in the AMH genome, aligning them with blastn (identity 95%). The selected 3' portions of the empty sites were extended in the 5' direction, thus retrieving the sequence corresponding to the 5' flankings to the putative archaic-specific insertions. The whole empty sites from the AMH genome (200 bp long) and their 3' and 5' portions (both 100 bp) were classified in separate sets, using shared codes for sequences belonging to the same site (F). Then, the 5' portions of the modern-specific empty sites were identified in the archaic DNA-sequencing reads library, using blastn (identity 95%). We then filtered archaic reads containing a match of at least 30 bp to a modern-specific "empty" site's 5' portion and at least 30 bp of non-matching bases 3' of them. These reads should thus contain both the 5' flanking site and the 5' terminal portion of the RI (G). The reads pertaining to the two sets of putative archaic-specific insertions 3' and 5' portions were associated to their corresponding modern-specific "empty" sites. This allowed to perform *de novo* assemblies site-by-site using the software ABySS (parameter k set to 40, H,I). Only sequences that were unambiguously assembled for both the 3' and 5' portions and that had a clear match for a modern-specific empty site were kept to produce the final sets of confirmed archaic-specific insertions 3' and 5' portions, as well as confirmed empty sites from the AMH reference genome (J).

Step 2.2) The putative modern-specific insertions 3' portions were extended to cover the full insertion as well as 100 bp of flankings in both directions (K). Archaic DNA reads were matched against the 3' and 5' flanking sequences using blastn (identity 95%). Reads with at least 30 bp match to both flanking sequences were selected. By doing this, the selected reads spanned the whole empty (pre-insertional) site (L). After associating the archaic reads corresponding to the putative empty sites to their respective modern-specific insertions, *de novo* assembly site-by-site was performed with ABySS (parameter k set to 40, M). Only putative modern-specific insertions whose flankings corresponded to an unambiguously assembled archaic empty site were selected as confirmed AMH-specific insertions (N).

RI identification between AMH and chimpanzee reference sequences and RT-DB

In order to identify RIs differentially present in the genomes of AMH and chimps, we first retrieved all RI from element families which are known to have been recently active (all AluJ, AluS and AluY subfamilies, all LINE-1HS and LINE1-PA subfamilies, all SVAs) in the two species reference sequences (GRCh37-hg19 and panTro5) from RepBase (Bao et al. 2015). The 5' and 3' flanking regions (100 bp) for all retrieved insertions were aligned using blastn (identity 95%) to the genome of the other species in order to find the respective putative empty (pre-insertional) sites. Two matching sequences (at least 85 bp), in close proximity to each other (less than 50 bp), were selected as a putative "empty" site for each "filled" site. These putative empty sites were then aligned back to the first species DNA using blastn (identity 95%) in order to confirm them as pre-insertional loci. After this procedure, we obtained the insertions specific to the first species (i.e. absent in the second) and vice versa.

We also generated a database of insertions, called RT-DB, that contains RIs retrieved from the human reference sequence GRCh37-hg19. RT-DB represent all reference insertions of AluS and AluY subfamilies, LINE-1HS, LINE-1PA2, LINE-1PA3, LINE-1PA4 and all SVAs (~666,000 RIs).

Computational validation procedure

After RI identification, all insertions were validated computationally. Putatively modern-specific RI were selected for having only one matching empty (pre-insertional) site unambiguously assembled from ancient DNA reads. All validated AMH-specific insertions and their absence from the assembled archaic empty sites were verified using RepeatMasker. Putatively archaic-specific insertions were instead selected for having unambiguously assembled both portions of each insertions and matching only one empty (pre-insertional) site in the modern reference genome. All archaic-specific insertions 3' and 5' portions were verified with RepeatMasker, as was their absence from the modern-specific empty sites. All archaic- and AMH-specific insertions were also verified for presence of the poly-A tail of the inserted element and TSDs flanking the RI.

Archaic-specific insertions in 1000 Genomes populations and inter-specific RI estimation

All archaic specific insertions loci were checked in 1000 Genomes (1KG) Phase3 (The 1000 Genomes Project Consortium 2015) .vcf files for the identification of non-reference variants present in modern day individuals. Archaic RI frequency was then averaged in modern populations according to the 1000 Genomes project annotations.

In order to estimate the amount of RI insertions that are polymorphic between populations we checked for the presence of Archaic RI in one AFR individual, then incrementally added other AFR individuals to the comparison. The rate by which polymorphic insertions were identified produced a curve that reaches a plateau after 20 individual confrontations. Applying this model to AMH insertions results in estimated 554 and 376 non-polymorphic-between-species AMH insertions (vs HN and HD respectively).

Assessment of AMH-specific RI in 1KG samples and frequency-based population tree

AMH-specific insertions present in the human reference GRCh37-hg19 may be polymorphic within the broader human population. However, lack of aligned reads spanning the insertion is, in itself, necessary but not sufficient to diagnose the absence of a given insertion within an examined resequenced genome. Even if present, indeed, given the high similarity to other copies of the same transposable element elsewhere in the genome, a given insertion may display no aligned reads due to multiple-mapping filterings. To assess presence/absence of a given insertion we therefore estimated the average coverage of the 1100 bps up and down-stream ("surroundings") of a putative insertion site and compared it with the coverage of the first and last 10 bps within the RI itself ("interfaces"). We therefore avoided any inference based on the coverage of the "core" inserted sequence, since this may have been affected by the multiple-mapping issues described above. We, instead, reckoned that the first and last 10 bps at the interface between the RI and the surrounding loci could be considered unique enough for the mapping algorithm to see them as a single mapping hit. Based on the reads available from the 1000Genomes Phase3 .bam files (The 1000 Genomes Project Consortium 2015) we then considered as:

- “diploid present” an insertion displaying a coverage >0 at both interface regions and where at least one interface region shows a coverage greater than ½ of the average surrounding coverage;

-“haploid present” an insertion displaying a coverage >0 at both interface regions and where both the interface regions show a coverage smaller than or equal to ½ of the average surrounding coverage;

- “absent” if at least one of the interface regions or the surroundings have zero coverage.

Our assessment approach is conservative with respect to the presence of a given insertion, since it is designed to overestimate absence. We then calculated population frequencies of presence of any given insertion, based on all the individuals available from 1000 Genomes Phase 3 (The 1000 Genomes Project Consortium 2015).

For each RI we calculated the absolute delta frequency per each pair of populations and we averaged it for all the insertions. The obtained matrix of average differences in presence/absence of human specific insertions was used to build a neighbour joining tree using the Ape R package (Paradis et al. 2004).

TMRCA estimates of genetic regions surrounding AMH-specific RI

The time to the Most Recent Common Ancestor (TMRCA) of each 10kbp regions encompassing a given insertion was estimated as described elsewhere (Inchley et al. 2016) based on 1000 Genomes sequences of AFR samples to avoid potential backwards biases due to the documented Neanderthal introgression in Eurasians (Green et al. 2010). All AFR individuals, and not only carriers of an insertion, were used for this calculations.

3P-CLR selection estimates for regions surrounding an insertion

For the sites surrounding AMH-specific insertions we aimed at identifying those that underwent positive selection after the split between Africa and Eurasia but prior to population differentiations within Eurasia. To do so, we used the Three Population Composite Likelihood Ratio (3P-CLR) statistic (Racimo 2016), to look for regions in the EGDP dataset (Pagani et al. 2016) that show evidence of selection that likely occurred shortly after the expansion out of Africa. The 3P-CLR statistic assumes a 3-population tree model with no post-split migration. To ensure that the individuals used in the 3P-CLR analyses represent the most basal split within living Eurasian populations , we used for our EAS population only Chinese and Japanese individuals from the Mainland East and Southeast Asia macro-population. The EUR individuals used were a random subset of the South and West Europe and East and North Europe populations. The AFR outgroup population consisted of the Yoruban individuals from the EGDP dataset (Pagani et al. 2016). Following indications (Racimo 2016), 100 SNPs (with at least 20 SNPs between them) were sampled in each window of length 1cM. Upon completion of the scan, sampled SNPs were grouped into 200kb bins that were assigned the maximum 3P-CLR score of the sampled SNPs in the window. Windows containing an AMH-specific insertion site and falling within the top 99th percentile of scores from this 3P-CLR test were considered to be under selection along the shared Eurasian branch.

AMH-specific RI, genes and preferential expression

In order to infer characteristics of the sites where AMH-specific RIs occurred, or impact they might have had on their insertional loci *in cis*, gene- and transcript-annotation tracks for the human reference genome GRCh37-hg19 were retrieved from ENSEMBL (Aken et al. 2016). RI loci for the different databases were identified in those tracks for information on genes containing RI.

Four gene-tracks (All Genes, genes with RT-DB insertions, genes with AMH-specific vs chimp insertions and genes with AMH-specific vs HN/HD insertions) were thus produced and compared for gene proportions and number of annotated transcripts, as well as gene length. Proportion of RI occurred in genes were compared between the tracks and tested with Fisher and binomial tests in R.

The tracks were divided in series containing the number of annotated transcripts for each gene and the gene’s total length, which were then compared between each other and tested with Wilcoxon and Kolmogorov-Smirnov tests in R.

Spearman’s non-parametric correlation tests between gene length, number of transcripts and number of RT-DB insertions (two at a time) were also performed, followed by a partial correlation test between number of transcripts and number of insertions in function of gene length. These tests were executed (in R, functions cor.test and pcor.test respectively; parameter method="spearman" in both cases) in order to exclude biases due to gene length and its impact on the random chance to observe features associated with it.

Functional annotation data on genes belonging to the aforementioned four tracks were also retrieved from DAVID Bionformatics Resources v6.8 (Huang et al., 2009). For general preferential expression information, nomenclature of tissues belonging to cohesive histological complexes was merged under the categories "Brain", "Testis", "Epithelium", "Placenta", "Uterus", "Lung", "Liver", "Lymph", "Kidney", "Eye", "Muscle", "Blood", "Colon" and "Pancreas". Only tissues individually called for preferential expression by at least 5% of all human genes were selected for the comparison.

For genes preferentially expressed in the brain, the categories were unpacked into "Brain (general)", "Undifferentiated Neurons", "Cerebellum", "Amygdala", "Hippocampus", "Peripheral Nervous System", "Thalamus", "Cajal-Retzius Cells", "Cortex", "Pituitary", "Hypothalamus", "Caudate Nucleus", "Dendritic Cells", "Substantia Nigra", "Subthalamic Nucleus", "Corpus Callosum". In this case, only tissues individually called for preferential expression by at least 0.5% of all human genes were selected for the comparison.

Tissue-by-tissue comparisons were tested using Fisher and binomial tests in R.

The lengths of genes displaying preferential expression in the different tissue types were also compared via a pairwise Wilcoxon test (pairwise.wilcox.test function in R with parameter p.adjust.method="bonferroni"). This generated a matrix of p-values (one for each possible pairwise comparison), which was then inverted (1 - p-value) and used as a matrix of distances for a cluster analysis (hclust function in R, parameter method="average"). A boxplot highlighting the length of the genes preferentially expressed in the various tissues and a dendrogram showing how the categories clustered in function of gene length were then plotted in R.

Gene Ontology functional analyses

To identify the gene-ontology category of the genes targeted by the AMH-specific RI, both vs chimp and HN/HD, we used ToppCluster (Kaimal et al. 2010), which allows the identification of biological programs using different gene sets to perform contrast and comparative analysis. ToppCluster was set with a false-discovery-rate (FDR) threshold of 0.05 and using “GO: biological process” annotation. The obtained matrix was used to compute the –log10 p-values to obtain significance scores for each functional term. Next, to reduce the redundancy within the GO terms, we used REVIGO (Supek et al. 2011) with parameters set to C=0.7, similarity measure “SimRel” (Schlicker et al. 2006) and using the *Homo sapiens* database. The scatterplots showing the representation of clusters from multidimensional scaling of the semantic similarities of GO terms were obtained with R. We used these plot to identify GO terms related with similar biological functions and the associated genes were used as input for GENEMania (default parameter) (Warde-Farley et al. 2010). The networks obtained from REVIGO were downloaded and visualized with Cytoscape (Christmas et al. 2005).

Case studies insertional loci annotation and functional inference

The three AMH-specific RI absent in both HN and HD that were identified as recent and displaying peculiar population distribution (see paragraph “Evidences of RIs contribution in the molecular differentiation of AMH” in the RESULTS chapter) were manually characterized using the UCSC Genome Browser (Speir et al. 2016), including genomic insertional locus, conservation of the sequence among primates, RepeatMasker presence/absence of repetitive elements, gene- and transcript-annotation.

To identify splicing motifs at the level of the insertion in the gene SHTN-1 we used Human Splicing Finder (HSF 3.0) (Desmet et al. 2009). HSF 3.0 was interrogated with the sequence of the RI + 100bp of flanking regions and with the reconstructed flanking without the RI itself. This was repeated for the whole intron where the insertion occurred and for the reconstructed intron lacking the specific RI, in order to assess its possible effect in splicing-alteration.

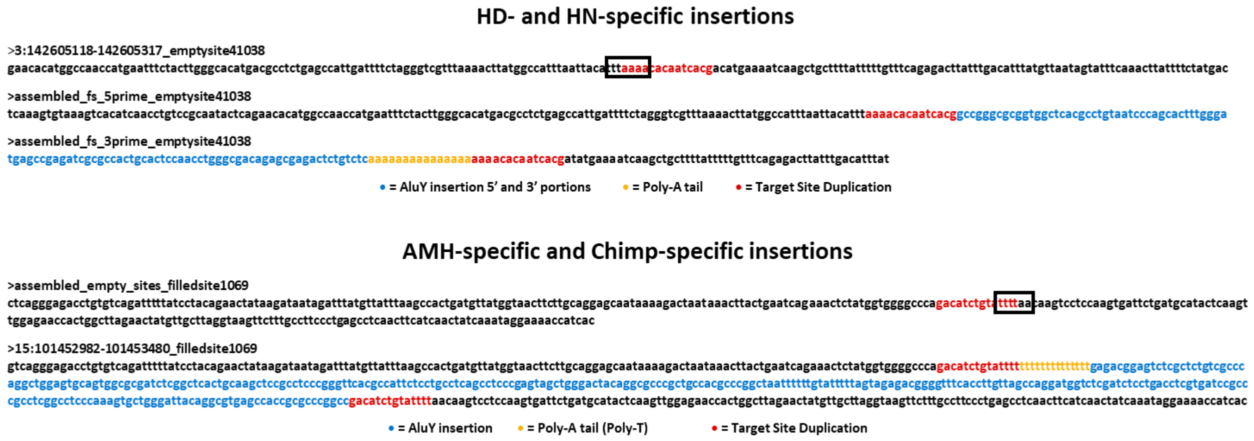
RESULTS

RIs identification

After comparing the genome of AMH to those of HN and HD we identified: i) 507 HN-specific and 331 HD-specific putative RIs, and ii) 3215 and 7185 putative AMH-specific RIs vs HN and HD, respectively.

As for the comparison between AMH (GRCh37-hg19) and chimpanzee (panTro5) genomes, we retrieved all RIs annotated in RepBase (Bao et al. 2015) in these two genomes and analyzed the presence/absence of the insertions in the reference sequences of the two species.

Next, we developed a computational validation procedure, through which we managed to eliminate all those insertions that presented uncertainties in mobile element subfamily attribution or whose location might be ambiguous (see METHODS for details). Thus, the following analyses were only performed on the most reliable canonically-inserted RIs identified (Figure 5).


HD- and HN-specific insertions
• = AluY insertion 5' and 3' portions
• = Poly-A tail
• = Target Site Duplication
AMH-specific and Chimp-specific insertions
• = AluY insertion
• = Poly-A tail (Poly-T)
• = Target Site Duplication

```
>3:142605118-142605317_emptysite41038
gaacacatggccaaccatgaatttctacttgggcacatgacgcctctgagccattgattttctagggtcgtttaaaacttatggccatttaattacatttaaaacacaatcacgacatgaaaatcaagctgcttttatttttgtttcagagacttatttgacatttatgttaatagtatttcaaacttattttctatgac
>assembled_fs_5prime_emptysite41038
tcaaagtgtaaagtcacatcaacctgtccgcaatactcagaacacatggccaaccatgaatttctacttgggcacatgacgcctctgagccattgattttctagggtcgtttaaaacttatggccatttaattacatttaaaacacaatcacggccgggcgcggtggctcacgcctgtaatcccagcactttggga
>assembled_fs_3prime_emptysite41038
tgagccgagatcgcgccactgcactccaacctgggcgacagagcgagactctgtctcaaaaaaaaaaaaaaaaaaacacaatcacgatatgaaaatcaagctgcttttatttttgtttcagagacttatttgacatttat

>assembled_empty_sites_filledsite1069
ctcagggagacctgtgtcagatttttatcctacagaactataagataatagatttatgttatttaagccactgatgttatggtaacttcttgcaggagcaataaaagactaataaacttactgaatcagaaactctatggtggggcccagacatctgtattttaacaagtcctccaagtgattctgatgcatactcaagt
tggagaaccactggcttagaactatgttgcttaggtaagttctttgccttccctgagcctcaacttcatcaactatcaaataggaaaaccatcac
>15:101452982-101453480_filledsite1069
gtcagggagacctgtgtcagatttttatcctacagaactataagataatagatttatgttatttaagccactgatgttatggtaacttcttgcaggagcaataaaagactaataaacttactgaatcagaaactctatggtggggcccagacatctgtattttttttttttttttttgagacggagtctcgctctgtcgccc
aggctggagtgcagtggcgcgatctcggctcactgcaagctccgcctcccgggttcacgccattctcctgcctcagcctcccgagtagctgggactacaggcgcccgctgccacgcccggctaatttttgtatttttagtagagacggggtttcaccttgttagccaggatggtctcgatctcctgacctcgtgatccgcc
cgcctcggcctcccaaagtgctgggattacaggcgtgagccaccgcgcccggccgacatctgtattttaacaagtcctccaagtgattctgatgcatactcaagttggagaaccactggcttagaactatgttgcttaggtaagttctttgccttccctgagcctcaacttcatcaactatcaaataggaaaaccatcac
```

Figure 5. Examples of RI sequences identified by our methodology.

For HD- and HN-specific insertions, the three sequences represent: empty (pre-insertional) site in the modern human reference GRCh37-hg19, 5' and 3' portions of the insertion with flankings assembled from the archaic species DNA. For Chimp- and AMH-specific RI, the sequences are: empty (pre-insertional) site in one species reference genome, insertion with flankings in the other species' reference genome. In all sequences, the inserted retrotransposon is represented in blue, the poly-A tail in yellow, the TSDs flanking the insertion or the single copy of the pre-insertional Target Site in red. The black rectangles on the empty (pre-insertional) sites indicate the exact location where the element inserted.

A number of species-specific RIs were computationally validated: 1906 Chimp-specific, 38 HD-specific, 64 HN-specific, 5402 AMH-specific (against chimps), 548 AMH-specific (against Denisova) and 806 AMH-specific (against Neanderthal) (Table 1). The validation method thus excluded approximately 87% of the identified insertions that could present any sort of bias or uncertainty in attribution. Of the validated AMH-specific RIs, 321 were present in the modern human genome and were absent in both HN and HD genomes (Supplemental Table 1).

	TOTAL	**Alu**	**LINE-1**	**SVA**
Chimp-specific RIs	1906	1370	463	73
HN-specific RIs	64	57	6	1
HD-specific RIs	38	32	6	0
AMH-specific RIs vs chimp	5402	4504	655	253
AMH-specific RIs vs HN	806	728	77	1
AMH-specific RIs vs HD	548	487	58	3
AMH-specific RIs vs both HN and HD	321	295	25	1

Table 1. Identified and validated RIs in chimpanzee, HN, HD and AMH genomes.

A large dataset (defined as RT-DB) of ~666,000 reference retrotransposon insertions from the most recent subfamilies of N-LTRrs (i.e. AluS, AluY, LINE-1HS, LINE-1PA2, LINE-1PA3, LINE-1PA4 and all SVAs) present in the reference GRCh37-hg19 was retrieved in order to assess if characteristics of loci targeted by AMH-specific insertions were random, retrotransposition-dependent or peculiar to the human lineage. The comparison of the identified insertions with RT-DB ones revealed that the activity of N-LTRrs in the human lineage has remained constant. Consistently, Alu RIs are far more common than LINE-1 RIs, while SVAs produced only a handful of insertions.

These results strongly suggest that the retrotransposition rate, or insertion maintenance rate, in the human lineage has remained relatively constant (0.6-0.8 insertions/Ky) and, consistently with previously reported results (Hormozdiari et al. 2013), that the rate of RI accumulation in humans has been approximately 2.5 times higher than in chimps (0.29 insertions/Ky).

Archaic-specific RIs and insertional polymorphisms

HD- and HN-specific RIs (38 and 64, respectively) were compared between th two species and with present-day AMH populations data from 1000 Genomes Project Phase 3 (Figure 6A-B) (The 1000 Genomes Project Consortium 2015). Based on the available 1000 Genomes Project data, three RIs were found in both archaic species, while nearly half of them (49 out of 102) are polymorphic to various degrees in modern populations. Interestingly, 8 of the insertions (1 HD-specific and 7 HN-specific) are absent in African (AFR) individuals and present at a low frequency only in some (or all) non-AFR populations.

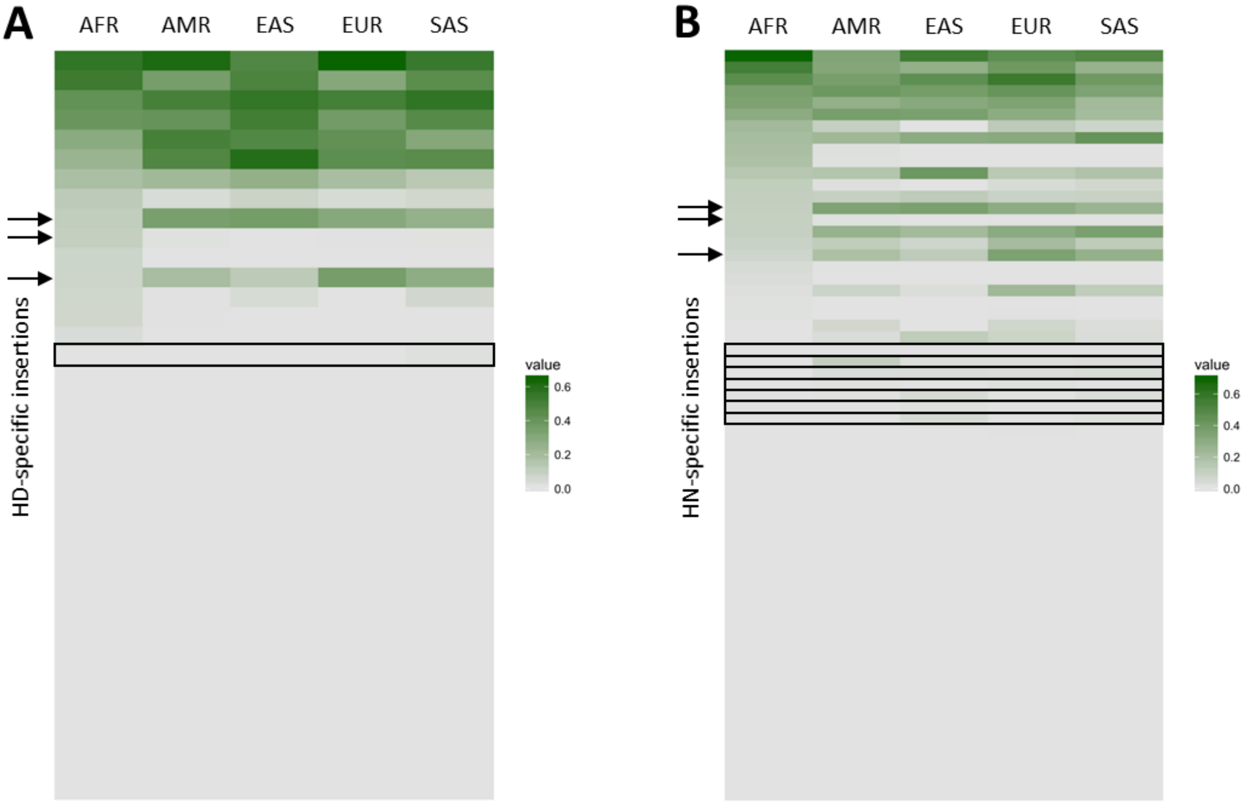


Figure 6. Heatmaps of HD-specific (A) and HN-specific (B) RI distribution in present-day populations.

Each line of the maps represents a single insertion, intensity of the color from grey to green indicates the frequency in modern populations. Arrows indicate three insertions that were identified as shared between HN and HD, rectangles highlight insertions that are putatively introgressed in modern populations post Out-of-Africa.

Thus, we speculate that these RIs might have introgressed in AMH via admixture with the archaic species after *Homo sapiens* migrated out of Africa. Conversely, given the documented presence of Neanderthal introgressed sequences within the genome of Eurasians (Green et al. 2010; Prüfer et al. 2014; Gardner et al. 2017), which in turn forms the majority of the human reference sequence, it may be the case that HN and HD specific insertions present on the human reference due to archaic introgression escaped our detection.

Since some putative archaic-specific RIs are polymorphic in modern humans, it is likely that at least some putative modern-specific insertions might be polymorphic in archaic populations as well; unfortunately, we would need many more available ancient genomes to properly test this. However, in order to estimate the number of potential polymorphic AMH-specific insertions, we took advantage of the large amount of population genetics data provided by the 1000 Genomes project, particularly AFR populations, who are less likely to host Neanderthal-derived genomic traits. Indeed, by randomly sampling AFR individuals we observed that a few samples would be sufficient to identify the vast majority of the archaic-specific polymorphic insertions, reaching a plateau at n=20. Similarly, ~45% of the putative species-specific insertions were shown to be polymorphic (Figure 7).

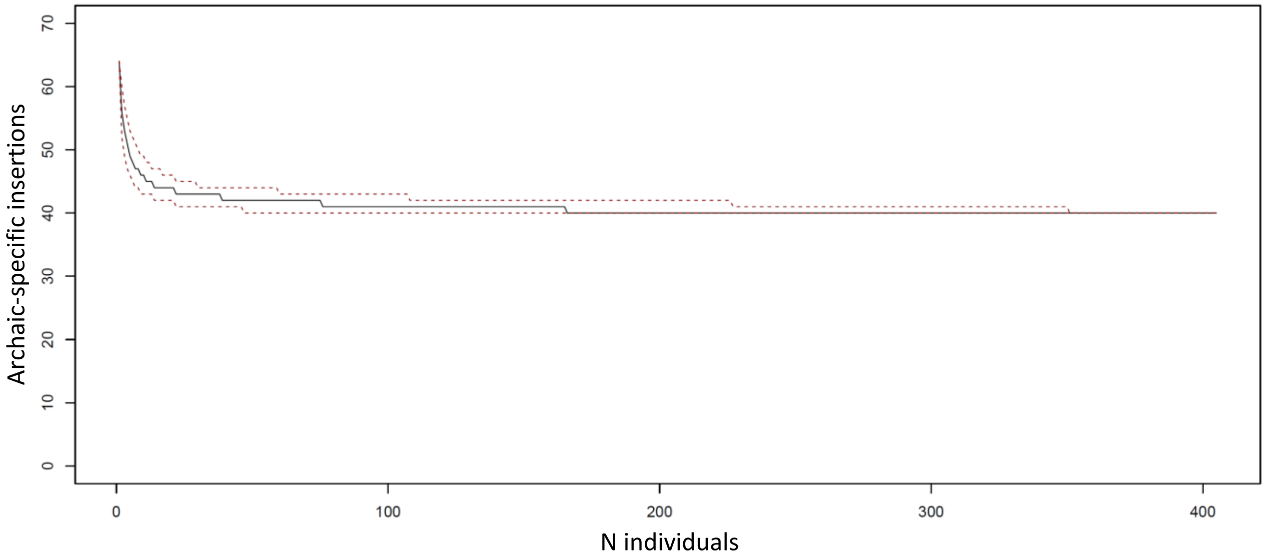


Figure 7. Simulated curve representing random sampling of AFR individuals for the identification and exclusion of polymorphic archaic-specific insertions.

Red dotted lines indicate 95% confidence intervals.

On the other hand, we observed that the 321 detected RIs that were present in AMH and absent in both archaic genomes fall below the ~45% threshold identified with the above procedure (considering polymorphisms against both HD and HN individually). This fact, together with the observation that HD and HN genomes are more divergent than two randomly-chosen AMH genomes (Reich et al. 2010), suggests that the above mentioned 321 insertions may be considered as reliable and truly AMH-specific.

AMH-specific RIs in present-day populations

The fact that the detected 321 AMH-specific RI are present in the human reference sequence (GRCh37-hg19) does not necessary imply that they are fixed in all human populations. We therefore evaluated their distribution in present-day populations by comparing the coverage of the unique 3' and 5' flanking regions with that of the RI/flanking interface in 1000 Genomes Phase 3 data (Figure 8; more details in METHODS).

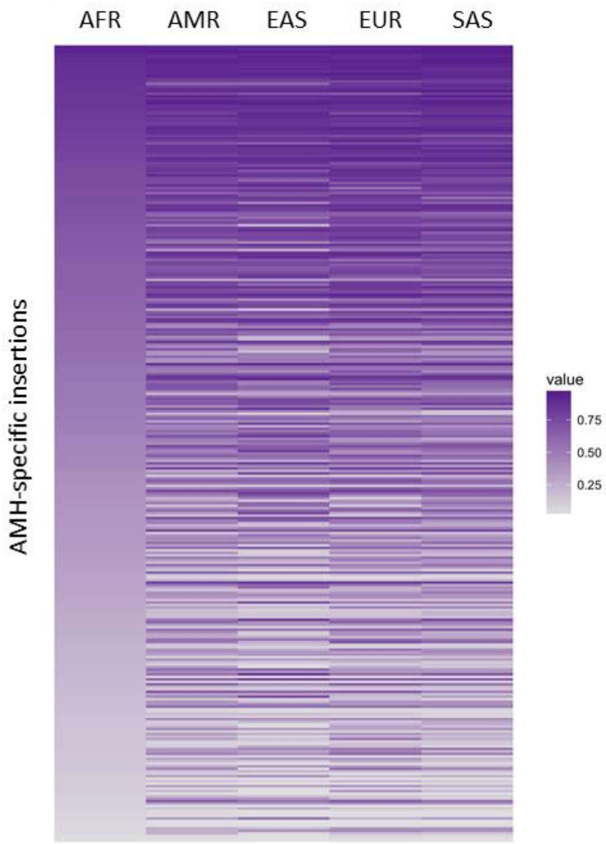


Figure 8. Heatmap of AMH-specific RI distribution in in present-day populations.

Each line of the map represents a single insertion, intensity of the color from grey to purple indicates the frequency in modern populations.

This analysis revealed that, of the 321 AMH-specific insertions, 24 (7,5%) appear to be fixed or almost fixed in all modern populations (allele frequency > 85%), while 8 (2,5%) are polymorphic in AFR individuals but fixed or almost fixed in all non-African populations (allele frequency < 65% in AFR and > 85% in non-Africans), suggesting that their fixation may be related to the Out-of-Africa event.

Interestingly, the patterns of RI distribution seem to closely reflect known pre-historic and historic migrations and population dynamics of AMH (Figure 9). In particular, populations of African descent are the more divergent and the Out-of-Africa groups cluster according to clear phylogenetic/phylogeographic relationships, with the expected exceptions of PUR and CLM who cluster with EUR populations and not with AMR, likely because of admixture during the re-colonization of North and South America (Montinaro et al. 2015).

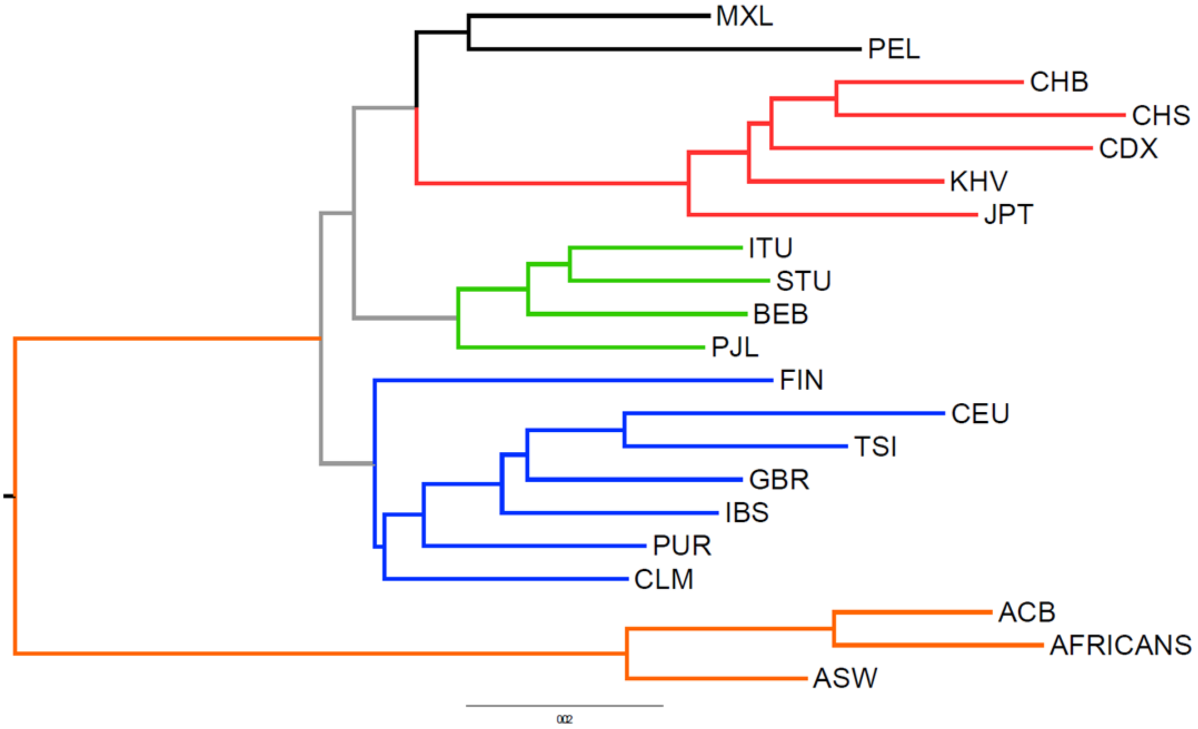


Figure 9. Neighbour joining tree calculated by using AMH-specific RIs in present-day populations as phylogenetic markers.

Branches in orange are from populations with African descent, blue for European descent, green for South-Asian descent, red for East-Asian descent and black for Native-American descent, as clustered by RI distribution.

Times to the Most Recent Common Ancestor (TMRCA) were also calculated for 10kbp sequences (Inchley et al. 2016) surrounding each insertion site (details in METHODS). Of the 24 insertions that are fixed or almost fixed in al modern populations, we selected those showing a TMRCA compatible with the split between AMH and HN/HD (TMRCA < 800 Kya) as potential candidates for selection/spread along the AMH lineage. Accordingly, we identified two RIs (8%), i.e. an AluYg6 insertion in chr1q25.3 that occurred in the gene EDEM3 and an AluYb9 insertion in chr10q25.3 that also occurred within the sequence of the gene SHTN1. Similarly, only one of the 8 AMH-specific insertions that are likely fixated post Out-of-Africa, an AluYa5 insertion in chr16q22.1, also displays a recent TMRCA. However, it is worth noting that TMRCA estimates were obtained from all the AFR individuals and not only from the carriers of an insertion; therefore, they must only be considered as a general indicator of the "age" of a given site surrounding an insertion or, in other words, as an upper-limit for the retrotransposition event itself.

Three Population Composite Likelihood Ratio (3P-CLR) statistic (Racimo 2016) was also performed on 200kbp loci surrounding each insertion. This analysis revealed that 28 (9.2%) out of the 306 AMH-specific RIs autosomal insertional loci are within the top 0.1% of loci subjected to post Out-of-Africa selection (details in METHODS).

Genomic features of loci targeted by AMH-specific RIs

The huge amount of genetic and genomic data presently available on modern humans allowed us to perform different exploratory analyses on the AMH-specific RIs and their surrounding genomic loci, aimed at evaluating the possible impact of RIs in our species evolution.

First, we compared selected datasets of RIs (RT-DB, AMH-specific vs chimp and AMH-specific vs both archaic species) with the ENSEMBL gene annotation (Aken et al. 2016) of the reference sequence GRCh37-hg19. We determined that 15367 genes contain insertions of RT-DB (48.7% of the insertions), 1779 genes contain AMH-specific vs chimp RIs (43.9% of the insertions) and AMH-specific RIs targeted 139 genes after the split with HN/HD (43.3% of the insertions) (Figure 10).

These data suggests that RT-DB insertions targeted genes and gene-related sequences, or have been maintained throughout evolution in those sequences, ~30% more frequently than randomly expected in respect to gene-size/genome-size (p-value < 10^{-16}), while AMH-specific RIs, both vs chimp and vs HN/HD, occurred in genes ~17% more frequently than expected (p-values < 10^{-16} and < 0.05 respectively).

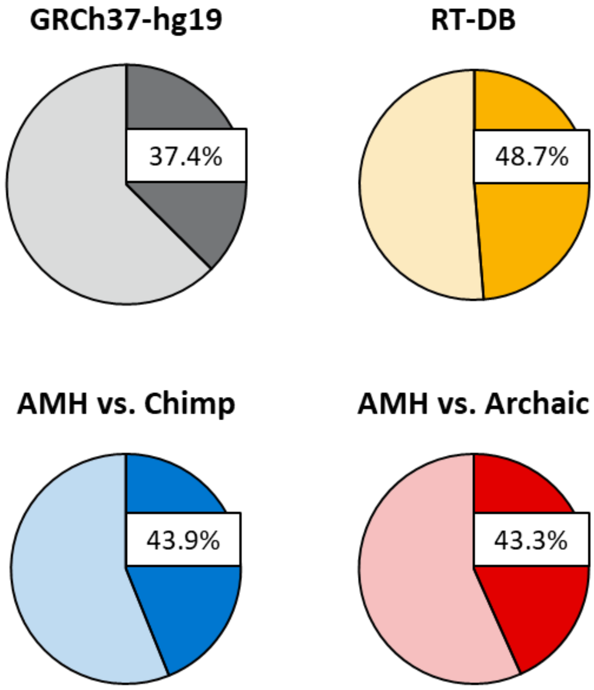


Figure 10. Proportion of ENSEMBL-annotated genes in the whole reference genome GRCh37-hg19 (grey), proportion of insertions that occurred in annotated genes for RT-DB insertions (yellow), AMH-specific RI vs chimp (blue) and AMH-specific RI vs both HN and HD (red).

In each diagram, the darker color denotes the percentage of RIs inserted in genes vs RIs inserted in non-genic regions (lighter color).

In addition, the ENSEMBL gene-annotation data revealed that, in general, the majority of genes in AMH genomes tend to have a low number of annotated transcript/splicing variants, with a decreasing trend between the proportion of genes and the number of transcripts (7.584 on average; mode: 1 transcript/gene). Intriguingly, the comparison of genes targeted by RIs in the human lineage with all others present in AMH genomes (Figure 11) revealed an average increase in the number of transcript and splicing variants for those genes that contain RT-DB insertions (8.428 on average, mode: 3 transcripts/gene; p-value < 10^{-16}). Notably, this trend increases further when analyzing RIs that likely inserted after the split with chimps (9.728 on average, mode: 5 transcripts/gene; p-value < 10^{-16}) and after the split with HN/HD (9.863 on average, mode:

8.5 transcripts/gene; p-value < 10^{-6}). Consistently, genes targeted by AMH-specific RIs, both vs Chimp and vs HN/HD, also have more annotated transcripts than genes containing RT-DB insertions (p-values < 10^{-11} and < 0.005 respectively).

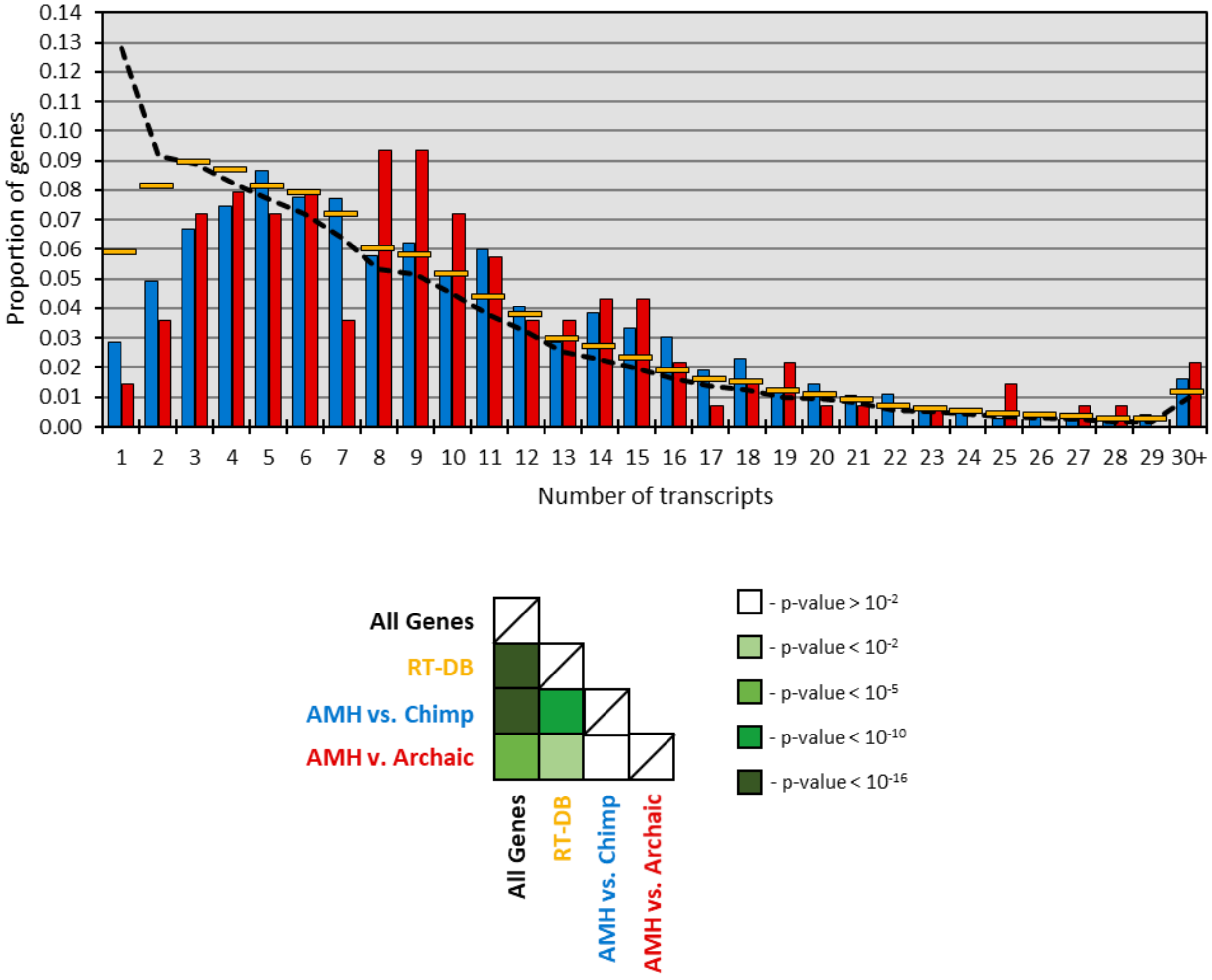


Figure 11. Number of annotated transcripts of genes targeted by retrotransposition.

Proportion of genes per number of annotated transcripts for all ENSEMBL-annotated genes in the reference genome GRCh37-hg19 (black dotted line), for genes targeted by RT-DB insertions (yellow lines), for genes containing AMH-specific RIs vs chimp (blue bars) and for genes with AMH-specific RIs vs both HN and HD (red bars). Below the graph, the table shows statistical significance of the differences between the series calculated with Kolmogorov-Smirnov tests.

We performed further analyses on genes containing RT-DB insertions in order to exclude a possible bias for gene length in the previously reported results. Indeed, a clear correlation is present between number of RIs and number of transcriptional variants (Rho = 0.284, p-value < 10^{-16}), but a strong correlation also exists between gene length and both parameters (Rho = 0.799 with number of RIs, Rho = 0.267 with number of annotated transcripts, both p-values < 10^{-16}). Thus, we performed a partial correlation test between the number of RIs and transcript variants of genes in respect to their length, which resulted in a statistically significant association of the first two parameters even after accounting for the third (Rho = 0.122, p-value < 10^{-50}). This test was repeated considering only genes containing AMH-specific RIs absent in chimps and AMH-specific RIs also absent in both HN and HD: in both cases the correlation between number of RIs and number of transcripts after accounting for gene length was confirmed (Rho = 0.196 with p-value < 10^{-16} and Rho = 0.240 with p-value < 0.005 respectively).

Next, we retrieved functional annotation data from DAVID Bioinformatics Resources v6.8 (Huang et al. 2009) and we obtained tissue-specific preferential gene expression information for 15126 out of 15367 genes with insertions from RT-DB, 1721 out of 1779 genes targeted by retrotransposition in the human lineage after the split with chimps and 124 out of 139 targeted after the split with HN/HD. Comparisons among these data show that genes targeted by retrotransposition tend to be more expressed than others in specific tissues (Figure 12). Genes containing RT-DB insertions tend to follow the general expression profile of all human genes, with a slight under-expression in some tissues (p-values < 0.05); however, genes targeted by AMH-specific RIs after the split with chimpanzees are more expressed than average in the brain and testis (+8.5% and +2.7% in absolute proportions respectively; p-value < 10^{-12} and < 10^{-2}), while being less expressed in the uterus, lungs, liver, blood and pancreas (decreases between -1.8% and -3.1% in absolute proportions; all p-values < 10^{-2}); finally, genes with AMH-specific RIs absent in both HN and HD are significantly more expressed in the brain and, with respect to genes targeted by RT-DB insertions, in testis (+13.8% and +7.5% in absolute proportions, p-values < $5x10^{-3}$ and < 0.05).

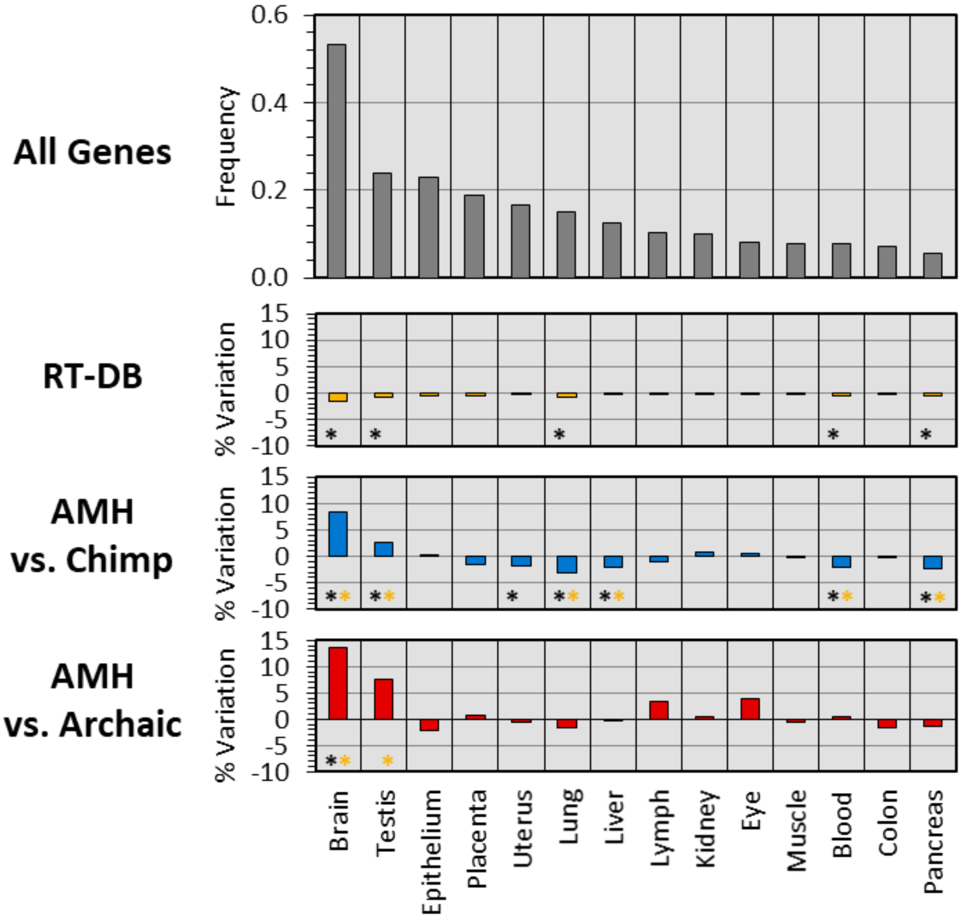


Figure 12. Preferential expression of genes targeted by retrotransposition.

Proportion of all human genes showing preferential expression in different tissues (grey bars); % increase or decrease in absolute proportions for preferential tissue expression of genes targeted by RT-DB insertions (yellow bars), genes containing AMH-specific RIs vs chimp (blue bars) and genes with AMH-specific RIs vs both HN and HD (red bars). Black asterisks mark significant differences between the series and all human genes while yellow asterisks mark significant differences between the series and genes targeted by RT-DB insertions.

We further analyzed genes preferentially expressed in the brain (Figure 13) and observed that genes targeted by RT-DB insertions follow the same expression pattern of all human genes; genes with AMH-specific vs chimps RI are generally highly expressed in the brain and seem to be even more expressed than average in the amygdala and hippocampus, as well as in undifferentiated neurons (+3.4%, +1.5% and +2.0% in absolute proportions respectively, p-values < 10^{-4} for the amygdala and < 0.05 for both hippocampus and undifferentiated neurons), while showing less expression in Cajal-Retzius and dendritic cells (-1.4% and -0.7% in absolute proportions, p-values < 10^{-3} and < 0,05); finally, genes containing AMH-specific RIs absent in both

HN and HD are significantly more expressed than average in undifferentiated neurons (+9.4% in absolute proportions, p-value < 10^{-2}).

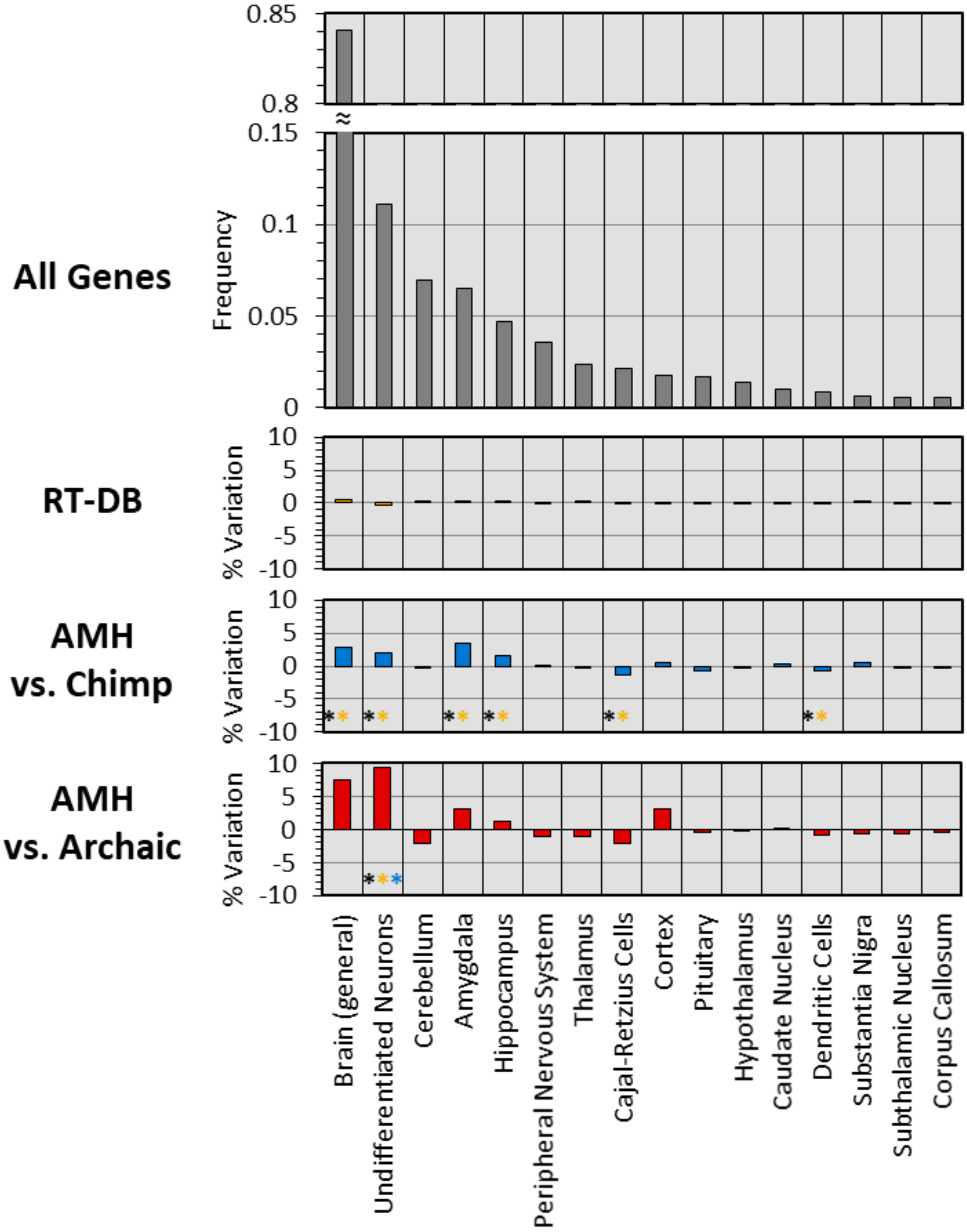


Figure 13. Preferential expression in the brain of genes targeted by retrotransposition.

Proportion of all human genes showing preferential expression in the brain divided by neural regions (grey bars); % increase or decrease in absolute proportions for preferential neural expression of genes targeted by RT-DB insertions (yellow bars), genes containing AMH-specific RIs vs chimp (blue bars) and genes with AMH-specific RIs vs both HN and HD (red bars). Black asterisks mark significant differences between the series and all human genes, yellow asterisks mark significant differences between the series and genes targeted by RT-DB insertions, blue asterisks mark significant differences between the series and genes targeted by AMH-specific RIs vs chimp.

Having previously evidenced a correlation between number of RIs present in genes and their length, we tested the possibility that genes showing preferential expression in specific tissues could exhibit a bias in relation to their length. We thus performed a pairwise Wilcoxon test between all series of lengths of genes preferentially expressed in the different tissues. This test showed a relative homogeneity of length of the various groups of genes, albeit with some pairwise confrontations resulting in statistically significant difference between the pairs of series (Figure 14).

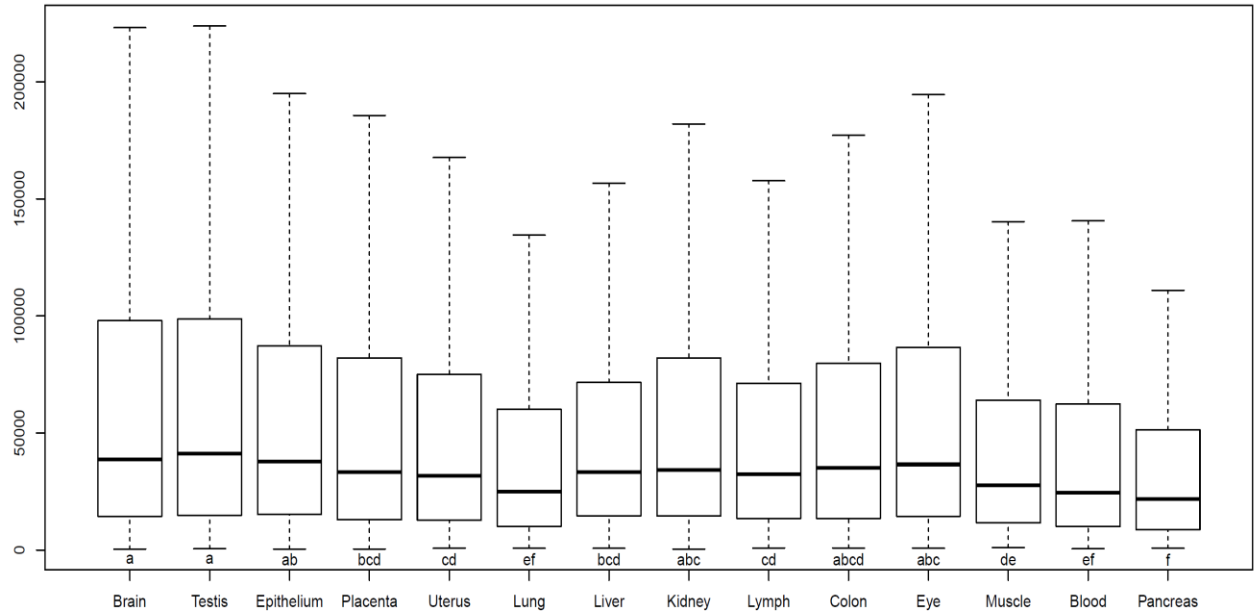


Figure 14. Length of genes grouped by preferential expression.

Boxplot depicting the length of all human genes grouped by preferential expression. The letters under each box represent statistical similarity of the series (p-values < 0.05): two boxes sharing the same letter are not statistically different while two that do not share a letter are.

Furthermore, we used the obtained matrix of p-values of the pairwise comparison as a matrix of distances between the series in order to perform a cluster analysis on the different categories (Figure 15). This resulted in the distinction of 3 major groups of genes based on their length distribution: genes preferentially expressed in the Eye, Kidney, Epithelium, Brain and Testis are generally longer, genes expressed in the

Pancreas, Blood, Lung and Muscle are shorter, while genes expressed in the Colon, Lymph, Liver, Placenta and Uterus fall in between.

Focusing on genes preferentially expressed in the Brain, which have been highlighted as containing more AMH-specific insertions, they do not seem significantly larger than others and instead form a cohesive group with genes expressed in other tissues that do not seem to contain as many AMH-specific RIs. Thus, these results seem to imply that the impact of gene length on previous analyses (if any) was negligible.

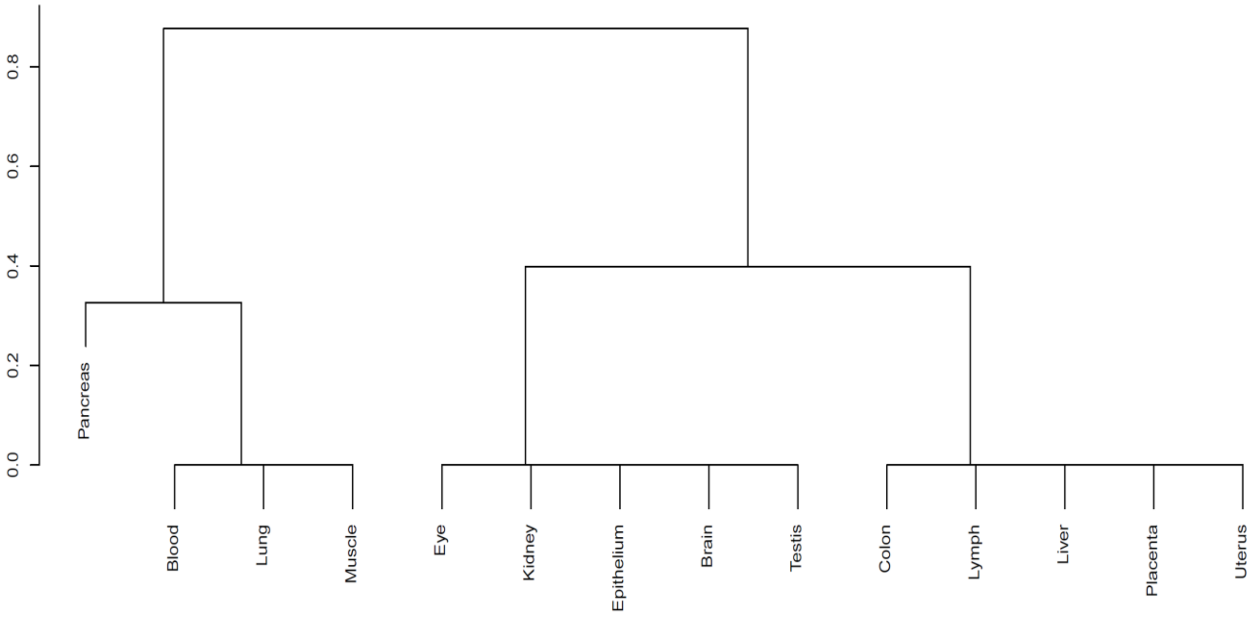


Figure 14. Clustering of genes preferentially expressed in specific tissues based on their length.

Dendrogram representing a cluster analysis performed on all human genes grouped by preferential expression, using p-values resulting from the pairwise Wilcoxon test comparison between all series of length of the genes as a matrix of distances.

Gene Ontology of genes targeted by AMH-specific RIs

In order to examine the functionality of genes targeted by insertions, Gene Ontology (GO) analyses were performed on all genes targeted by AMH-specific RIs, both vs Chimp and vs HN/HD. ToppCluster analyses (Kaimal et al. 2010) revealed that, of the 238 GO terms identified between the two lists, 175 GO terms (73,5%)

were overrepresented in genes targeted by AMH-specific RIs vs chimp, whereas 23 (9,7%) were overrepresented in genes targeted by AMH-specific RIs vs both HN/HD (Figure 15).

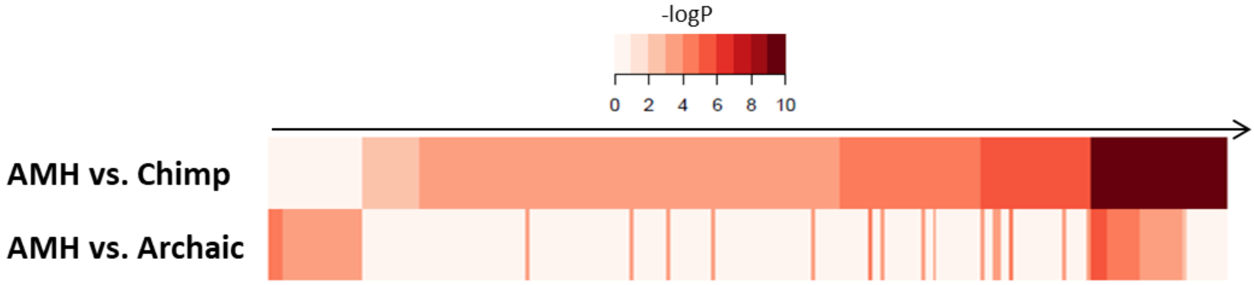


Figure 15. Heat maps representing -log(p-values) of GO terms associated with genes targeted by AMH-specific RIs vs chimp (top) and AMH-specific RIs vs both HN/HD (bottom), order for increased significance in the top row.

Next, we selected the GO terms that were enriched in one group of genes and not in the other as lineage-specific functionalities that might correspond to different moments in the evolution of the human lineage, i.e. hominid-specific GO terms (for terms enriched only in genes targeted by AMH-specific RI vs chimp) and sapiens-specific GO terms (for terms enriched only in genes targeted by AMH-specific RI vs both HN/HD). Interestingly, semantic similarity of hominid-specific GO terms showed that the most enriched functionalities of genes containing AMH-specific RI vs chimp are related to cognition, learning and memory capabilities, vocalization behavior, neuron recognition, dendrite morphogenesis, reflexes and regulation of locomotion (Figure 16). Remarkably, these functionalities associate in networks involving a large number of genes containing AMH-specific RIs vs chimp.

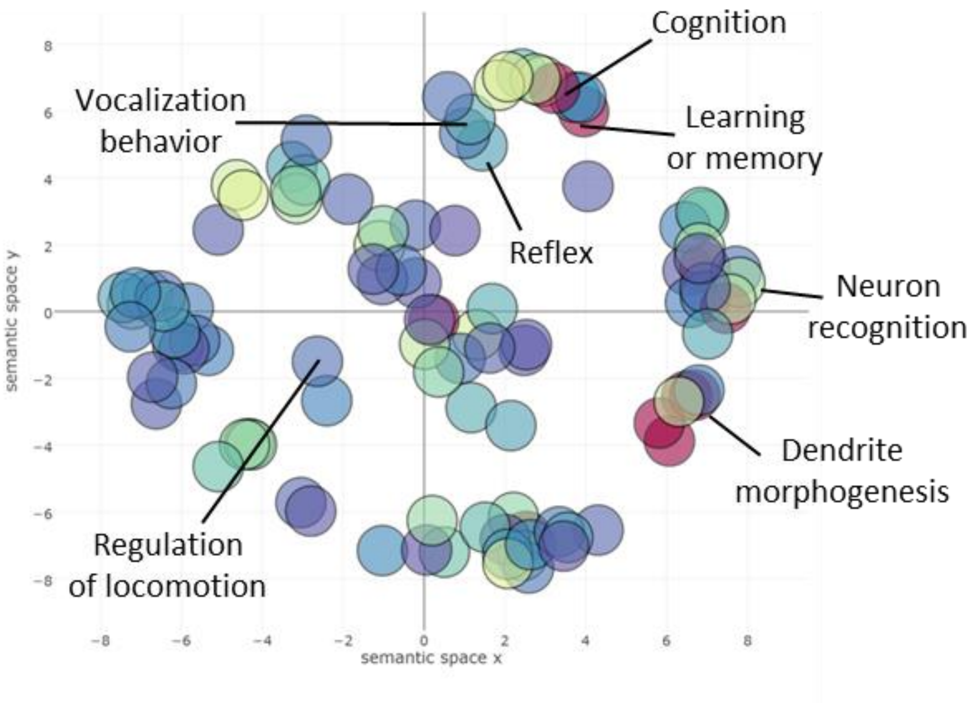


Figure 16. Scatterplot representation of the identified hominid-specific GO terms.

The x and y coordinates of the circles were derived from the Revigo analysis, based on multidimensional scaling of the matrix with the GO semantic similarity values. The functional categories associated with genes that form networks are highlighted and labeled.

Even more noticeably, for enriched sapiens-specific GO terms, all functionalities associated in networks are neural-related: synapse maturation and its regulation, neuron maturation and migration, gliogenesis and glia differentiation (Figure 17A-B).

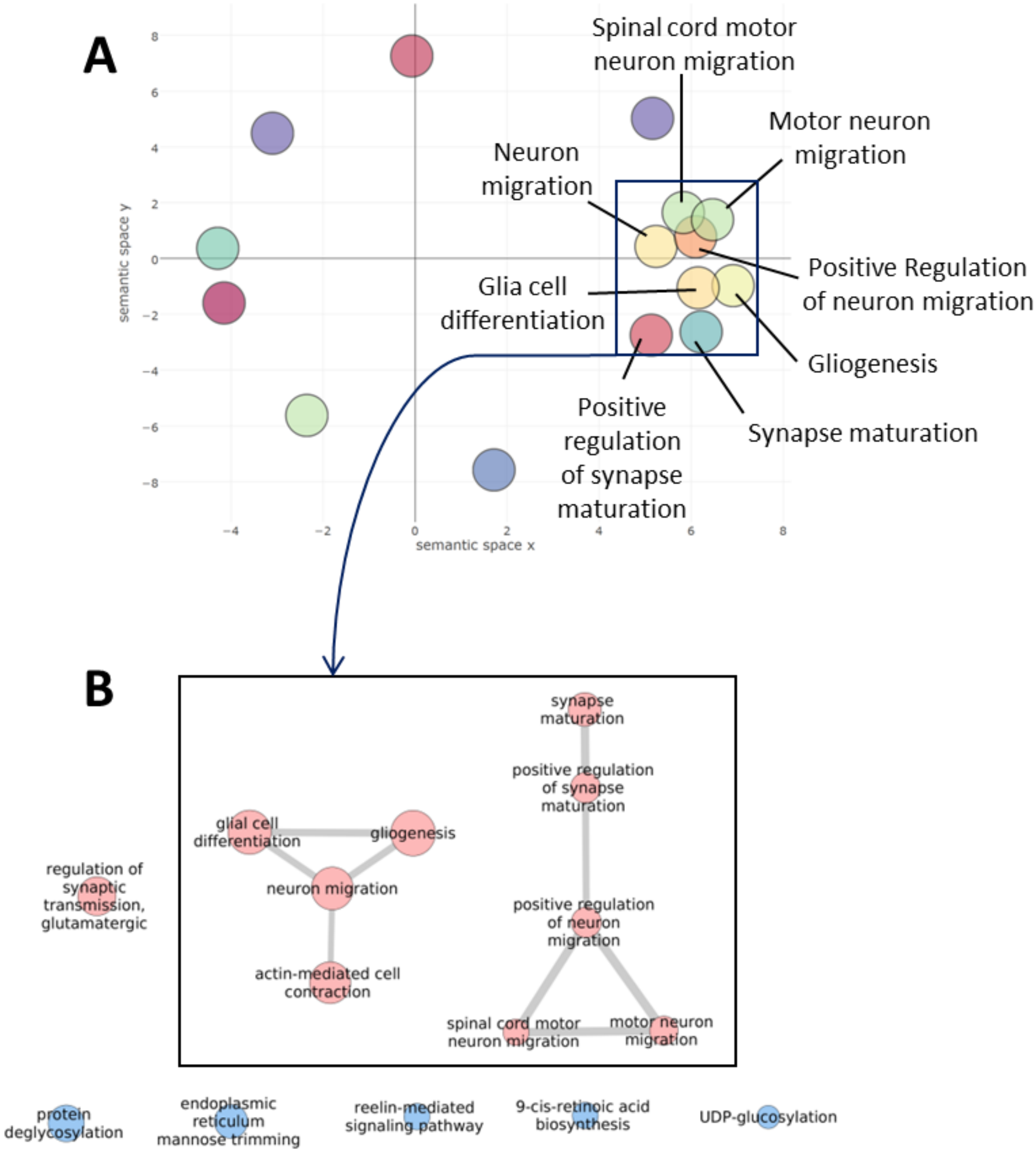


Figure 17. Sapiens-specific GO terms.

A) Scatterplot representation of the identified sapiens-specific GO terms. The x and y coordinates of the circles were derived from the Revigo analysis, based on multidimensional scaling of the matrix with the GO semantic similarity values. The functional categories associated with genes that form networks are highlighted and labeled. B) Functionalities of sapiens-specific GO terms associated in networks (if applicable). Red terms are neural-related while blue terms are not.

Genes associated with these GO terms also form a complex network of interactions (Figure 18). Strikingly, two of the genes with the larger amount of interactions in this network are SHTN1 and EDEM3 (1st and 11th scores in order of significance), which contain the previously identified (see above) AluYg6 RI (chr1q25.3) and the AluYb9 RI (chr10q25.3) respectively.

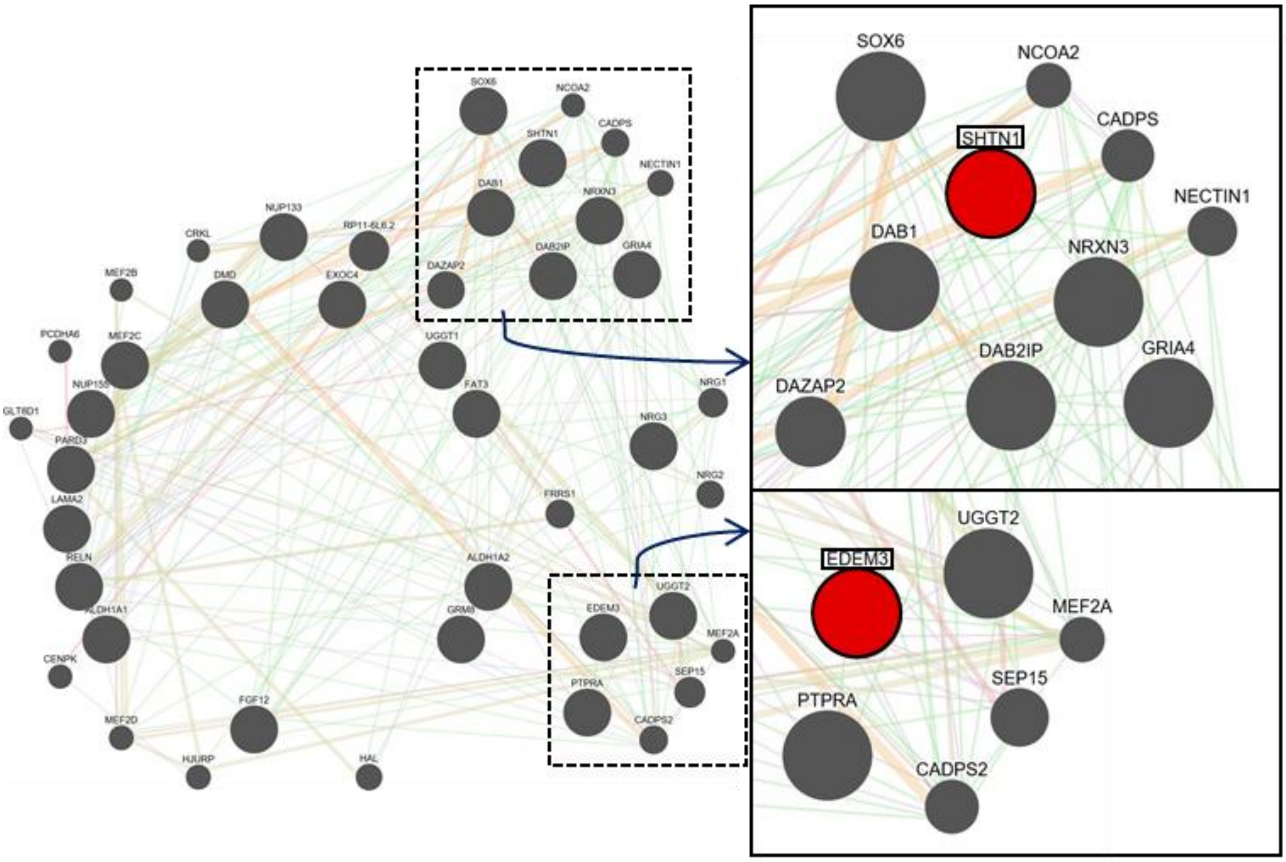


Figure 18. Gene network of genes containing AMH-specific RIs absent in both HN and HD with neural functionalities.

The larger the circle, the more the corresponding gene has interactions with other genes in the network. The sub-network showing strong interactions with the gene SHTN1 is highlighted in the top-right, while the sub-network with more interactions with the gene EDEM3 is highlighted in the bottom-right.

Evidences of RIs contribution in the molecular differentiation of AMH

Since AMH-specific RIs seem to increase the variability of transcripts and tissue-preferential expression of their targeted genes, we next characterized in greater detail the insertional loci of the three "recent" insertions with a peculiar population distribution that were identified above in paragraph "AMH-specific RIs in present-day populations". The first one, an AluYa5 RI in chr16q22.1 (Figure 19), is polymorphic in AFR populations (average frequency of 55%), but fixed or almost fixed in all non-African populations (with the highest difference in frequency between AFR and non-African populations). Although no role in functional alteration was detected for this RI in its insertional locus, the insertion is associated with signs of post Out-of-Africa selection, as revealed by the 3P-CLR selection estimate that places its genomic locus in the top 0.1% loci.

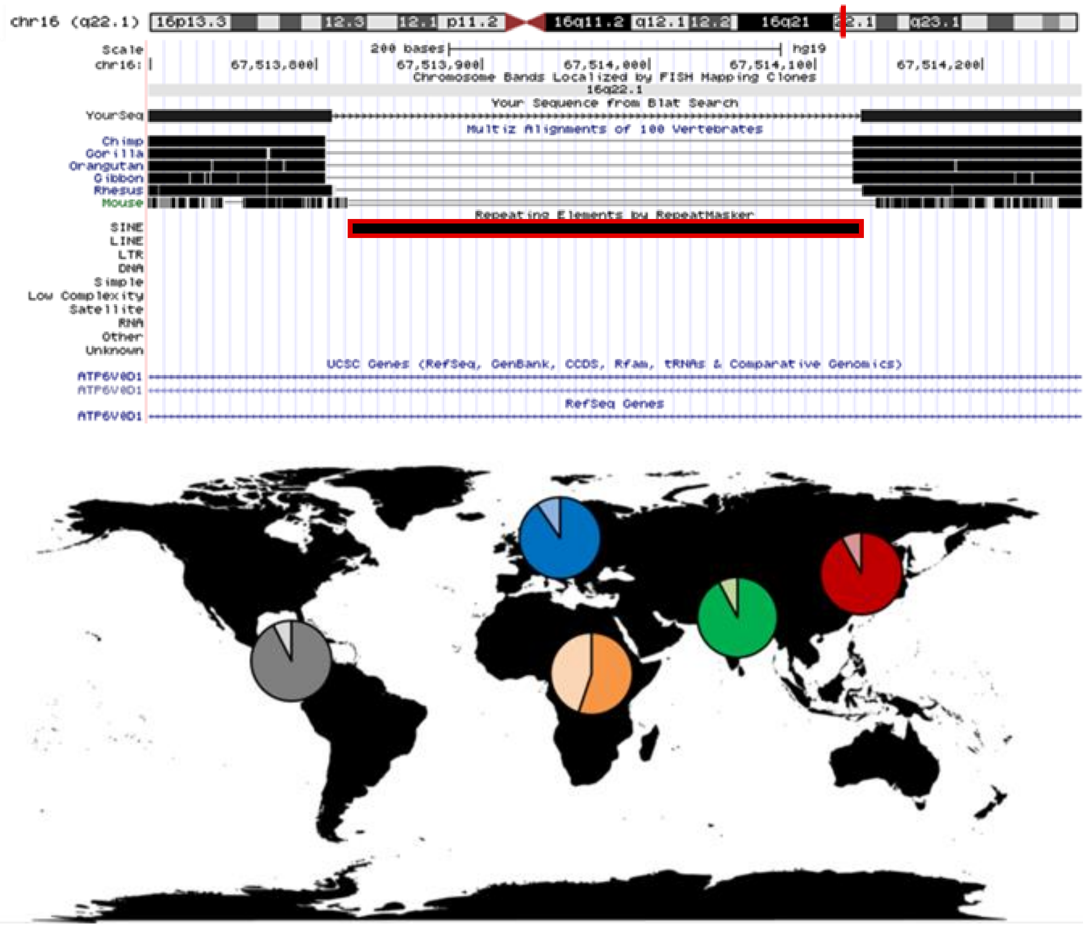


Figure 19. Annotation of the genomic location and distribution in present-day populations of the AluYa5 insertion on chr16q22.1.

The insertion is highlighted in red. In the map, for each diagram, a darker color indicates the presence of the RI and a lighter one its absence.

The second RI analyzed, an AluYg6 insertion in chr1q25.3, is inserted in gene EDEM3 (mentioned above) and is estimated to be fixed or almost fixed in all AMH populations, while completely absent in chimps, HN and HD, as well as in other primates (Figure 20). Remarkably, there is a shorter EDEM3 annotated alternative transcript ending precisely in correspondence with the poly-A tail of the AluYg6 insertion, resulting in exonization of this RI. This alternative EDEM3 transcript is not annotated in chimpanzees, and we suggest that it is extremely likely that this transcript is a direct consequence of the AluYg6 insertion into gene EDEM3.

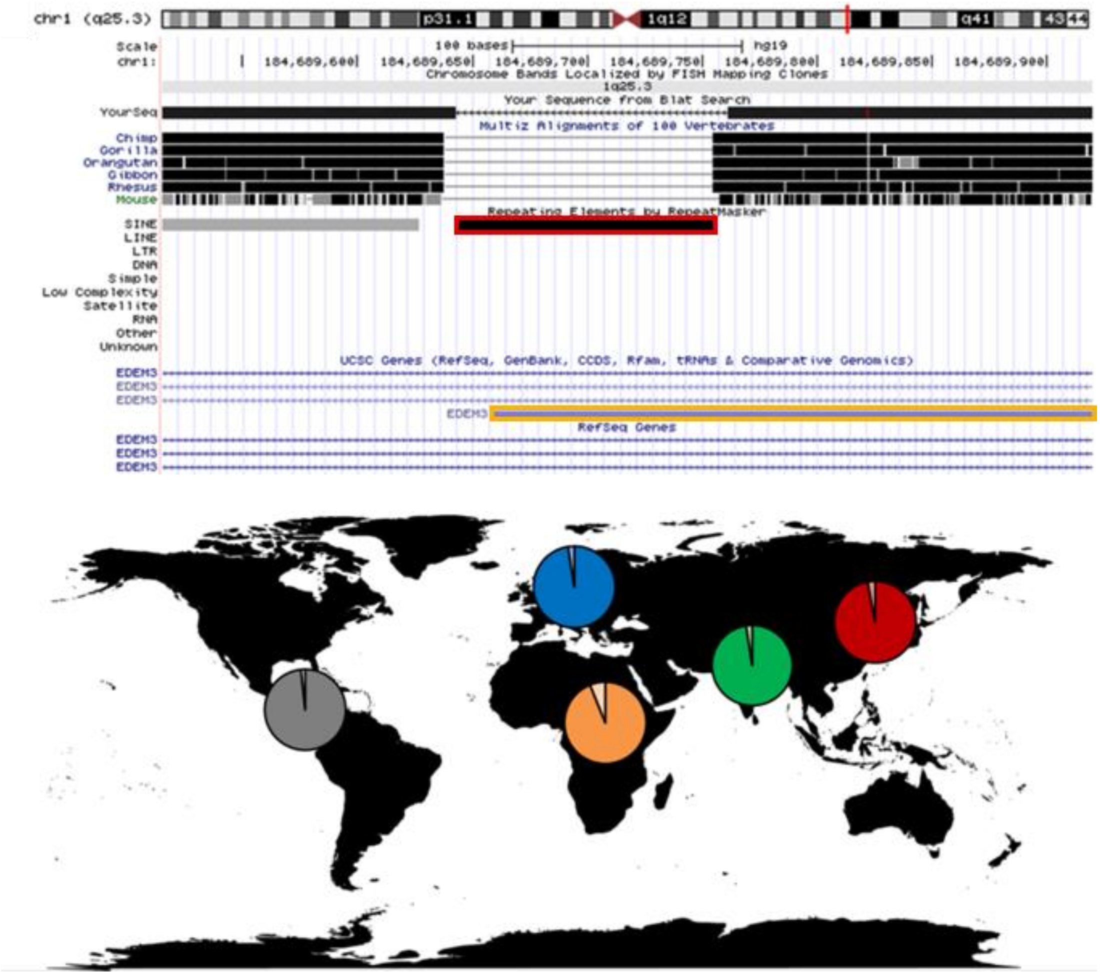


Figure 20. Annotation of the genomic location and distribution in present-day populations of the AluYg6 insertion on chr1q25.3.

The insertion is highlighted in red and a yellow rectangle highlights an alternative transcript that terminates precisely at the poly-A tail of the RI. In the map, for each diagram, a darker color indicates the presence of the RI and a lighter one its absence.

Finally, the third RI analyzed is an AluYb9 RI in chr10q25.3, which inserted in the 15th intron of the gene SHTN1, antisense in respect to the gene's transcriptional directionality. This gene has the highest level of interaction in the previously identified network of neural genes. This specific RI is mostly fixed in all AMH populations and absent in HN, HD and chimp genomes, as well as in other primates (Figure 21).

The gene SHTN1 has various annotated transcriptional/splicing variants, two of which lack the first the two exons that flank the intron where the Alu insertion occurred. Intriguingly, the analyses of this intron with Human Splicing Finder 3.0 (Desmet et al. 2009), both as an "empty allele" (with the retrotransposon sequence removed and the original pre-insertional site reconstructed) and as a "filled allele" (with the AMH-specific

AluYb9 insertion), revealed that differences in the predicted Splicing Enhancing/Silencing Matrices are present between the two sequences. Remarkably, these analyses revealed the presence of putative splicing silencing peaks in the filed allele. The strongest peak is located precisely in the inserted AluYb9 sequence (Figure 22), suggesting that the Alu insertion may induce a splicing-silencing effect on the SHTN1 gene.

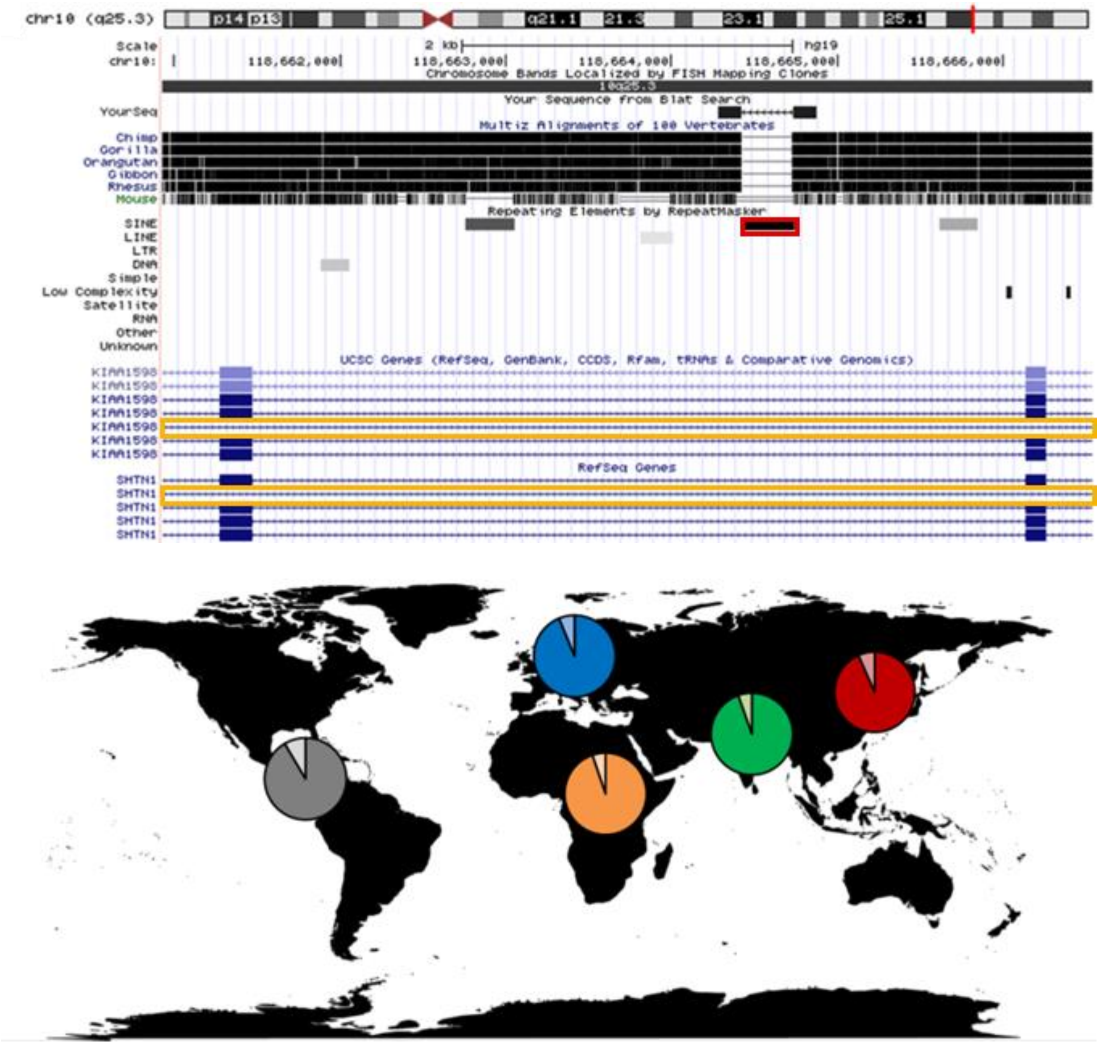


Figure 21. Annotation of the genomic location and distribution in present-day populations of the AluYb9 insertion on chr10q25.3.

The insertion is highlighted in red and yellow rectangles highlight annotated alternatively-spliced products for the gene in which the insertion occurred. In the map, for each diagram, a darker color indicates the presence of the RI and a lighter one its absence.

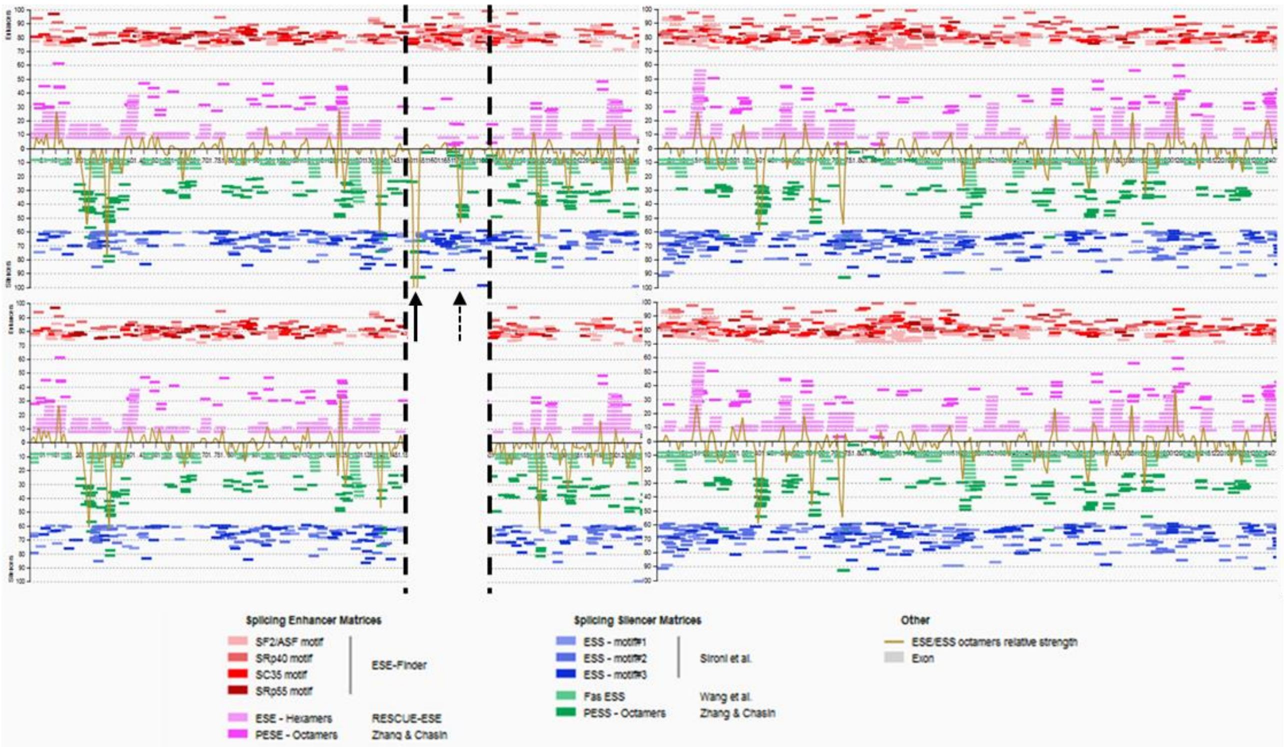


Figure 22. Splicing prediction in the sequence corresponding to the “filled allele” (top, containing the intron and the AluYb9 insertion on chr10q25.3) and in the sequence corresponding to the “empty allele” (bottom, containing just the intron).

The sequence is oriented in the same transcriptional orientation of the gene, black dotted lines highlight the position of the RI. Pink and red lines represent Splicing Enhancer Matrices, green and blue ones Splicing Silencing Matrices; ochre lines represent the combined strength of Enhancer/Silencing Matrices on the sequence. Arrows highlight silencing signal peaks that occur precisely in the RI sequence.

DISCUSSION

In AMH, retrotransposition has been studied mostly for its mutagenic effects and implications in disease insurgence. On the other hand, knowledge about the molecular evolution of our genome relies mostly on simple markers such as SNPs, short InDels and large Copy Number Variants (CNVs), while the role of repetitive/complex regions of the genome is poorly understood. Among others, the complexity of analyses involving repetitive sequences and structural variations in conjunction with the widely used NGS sequencing technology is a challenging task. However, evidences from various Eukaryote organisms suggest that retrotransposition might play an important role in speciation and molecular evolution of genomes (Richardson et al. 2015).

In this study, we evaluated the possible impact of RIs on the differentiation processes that occurred in the human lineage and especially during AMH evolution. In order to do this, we first identified putative species-specific RIs across the genomes of AMH, HN, HD and chimpanzees (Table 1). Notably, the identified RIs reveal a relatively constant retrotransposition/maintenance rate in the human lineage (0.6-0.8 insertions/Ky), which is ~2.5x higher than the rate of N-LTRr mobility/maintenance in chimps (0,29 insertions/Ky). In sum, these data suggest that n-LTRr might have, in a constant manner, impacted our genome throughout the evolution of the human lineage more than they have affected the chimp lineage, despite chimpanzees' shorter generation time. However, the impact of RIs described in this study is just a minor repertoire of the putative effect that RIs can exert on genome structure and regulation, as i) our study is limited to the identification of RIs on very limited sequencing information from Neanderthal, Denisovan and Chimpanzee genomes; and ii) in this study we could only analyze the impact of RIs in *cis*, although RIs are known to impact gene expression and genomic architecture both in *cis* and in *trans* (Garcia-Perez et al. 2016). Thus, we are just starting to uncover the role of n-LTRrs on human evolution, and future genomic studies on Neanderthals and Denisovans will help revealing the full impact of RIs on the evolution of the human lineage.

AMH-specific RIs and functional variability increase at targeted loci

Focusing on the insertions that are specific to AMH, one of our most prominent results is the strong correlation highlighted between RIs that integrated in genes and the increase in the number of annotated transcript for the genes in which the insertion occurred (Figure 11). While this is true for all RT-DB insertions and the corresponding genes, the effect seems even higher for RIs that are exclusive to AMH. This could be interpreted as a sign of target-preferentiality of retrotransposons in general and particularly in the human lineage: RIs might thus preferentially target genes with a high variety of transcripts. Another possibility is that, just by random chance, longer genes might accumulate more RIs than shorter ones and also exhibit a higher variety of transcripts. While there definitely is a strong correlation between gene length and both number of insertions and number of annotated transcripts, this doesn't seem to explain all of the observed variability in human genes annotation: a correlation between number of identified RIs and variety of transcripts remains, in fact, even after accounting for gene length. This observation, together with the characteristics of N-LTRr sequences and their possible effects upon integration (Goodier and Kazazian, 2008; Cordaux and Batzer, 2009), strongly suggest that at least some part of the increase in the variety of transcripts is an effect, and not a cause, of the accumulation of new retrotransposition events. It is tempting to speculate that the new transcript/splicing variants of genes targeted by AMH-specific RIs has led to an increase in their functional complexity.

Another important observation from the analyses included in this study is that genes targeted by AMH-specific RIs tend to be preferentially expressed in specific tissues (Figures 12-13). Notably, in the human lineage these genes are especially likely to be expressed in the brain, with an enrichment of >26% compared to its preferentiality for all human genes. In particular, those genes targeted by AMH-specific insertions vs chimp are more expressed in the amygdala, hippocampus and undifferentiated neurons (up to >52% in respect to each cell-type/tissue expectation), while after the split with HD and HN the enrichment of preferential expression occurs specifically in undifferentiated neurons (>85% in relation to their general baseline). This observation seems to be unaffected by the length of the genes expressed in the different tissues. In fact, genes expressed in the human brain are not significantly longer than ones expressed in other tissues and, instead, when grouped by their dimensions, form a coherent cluster with genes preferentially expressed in the eye, kidney, epithelium and testis.

These results show a strong association between RIs and neural genes in the lineage of AMH. Indeed, GO analyses revealed a consistent pattern of neural-related functionalities for genes containing RIs, which is consistent with the aforementioned tissue-specific preferential expression. Interestingly, hominid-specific GO terms of genes targeted by AMH-specific RI vs chimp are highly related to biological and ethological processes that occurred during the differentiation of hominids after the split from chimpanzees, including: neuronal signaling, cognitive capacity, vocalization behavior, reflexes and locomotion regulation (Figure 16). Strikingly, the most relevant functionalities associated with sapiens-specific GO terms all relate to glia differentiation, synapse maturation and neuron maturation and migration (Figure 17). These functionalities, associated with the preferential expression in undifferentiated neurons of genes that carry AMH-specific RIs absent in both HN and HD, might reflect the importance of these genes in human neural differentiation processes. It is therefore tempting to speculate that the aforementioned increase in transcript variability of specific genes, seemingly induced by RIs, could be tied to the increase in functional complexity of the human brain that occurred all throughout our evolutionary lineage (Hrvoj-Mihic et al. 2013).

Population distribution of RIs and natural selection

These possible variability-increasing effects of RIs in our lineage and their specific relevance in neural genes should, theoretically, have been subjected to natural selection. Among the identified 321 AMH-specific RIs, the most likely candidates for an adaptive effect on their carriers are those insertions that are fixed across all AMH populations or whose distribution reflects a strong phylogenetic/phylogeographic pattern (e.g. fixation post Out-of-Africa). We therefore compared the identified RIs in this study with the genomic variability of present-day human populations provided by the 1000 Genomes Consortium data. AMH-specific RIs, according to calculated TMRCAs (for AFR populations in which the insertion is fixed or almost fixed), seem to have occurred between >6.5 Mya (i.e. before the split with the chimpanzee lineage) and present day. These time estimates must, however, only be intended as upper limits for the actual times of the insertions occurrence. Indeed, a large portion of identified putative HN- and HD-specific insertions was shown to be polymorphic to various degrees in present-day populations. Interestingly, the absence of some archaic-specific insertions in all AFR populations and their presence in some (or all) non-African individuals strongly suggest introgression of genetic material in AMH that occurred post Out-of-Africa, possibly via interbreeding with HN and HD. This observation and the populations in which this phenomenon seems to have occurred is consistent with previously reported examples of interbreeding evidences between AMH and HN/HD after *Homo sapiens* migrated into Eurasia (Green et al. 2010; Prüfer et al. 2014; Gardner et al. 2017).

As expected, RIs seem to have been selectively neutral and polymorphic throughout AMH populations although in some instances they show traces of selection only a long time after their putative occurrence. Amongst all RIs that are present at high frequencies in modern populations, a few of them show a peculiar geographic distribution characterized by polymorphism in Africa and fixation (or almost fixation) in non-African individuals. We hypothesize that these RIs could have reached fixation following the Out-of-Africa event as a consequence of genetic drift or selection. In both cases, these insertions predate the spread of AMH out of Africa. In particular, the AluYa5 insertion in chr16q22.1 identified in this study (Figure 19) has a TMRCA of ~300 Kya and displays rapid fixation in non-African populations, with its surrounding locus being in the top 0.1% 3P-CLR loci. Therefore, we speculate that this insertion was actually subjected to selection post Out-of-Africa, possibly hundreds of Ky after the insertion itself occurred.

While our estimations are only approximations for both the putative age of an insertion and selective pressures acting on an insertional locus, due to the lack of specific methodologies of time estimation and selection for RIs, our results suggest that RIs might occur in a genome and can be maintained randomly within a population under neutral selective pressures. At later times, because of population dynamics or environmental changes, an insertion and the putative novel functional variants it generated might be co-opted and undergo non-neutral selective pressure, in a similar manner as previously reported cases of “soft” selective sweeps detected with SNPs analyses (Pritchard et al. 2010; Hernandez et al. 2011). On an evolutionary timescale, this process seems more likely than insertions having a strong functional-alteration effect immediately upon integration. In fact, most functional regions of a genome are highly conserved and functionality-altering effects would likely be disease-inducing and selected against alleles carrying the RIs. Additionally, genetic drift might also play an important role in the maintenance/diffusion of RIs in human populations, particularly concerning Out-of-Africa bottlenecks.

This interpretation is also consistent with differences in the percentage of RIs targeting genes observed in RT-DB with respect to AMH-specific RIs, both vs chimp and vs HD/HN (Figure 10). Indeed, these three datasets of retrotransposition events are progressively smaller subsets of the same starting pool of insertions and reflect progressively shorter timescales. Importantly, the effects of a retrotransposon insertion can be co-opted even a long time after the insertion itself occurred, creating new functional variants; thus, a dataset of older insertions (on average) is more likely to show annotated functional variants in a modern genome than a dataset of relatively younger insertions.

Impact of RIs in modern humans

Previous studies have revealed how retrotransposons can influence the regulation of the loci in which they inserted in a myriad of ways (Richardson et al. 2015). Besides the activity of the sense and antisense LINE-1 promoters contained within full-length LINE-1s (Swergold 1990; Speek 2001; Macia et al. 2011) and the epigenetic silencing of retrotransposon sequences mediated by DNA methylation (Yoder et al. 1997) or histone modifications (Garcia-Perez et al. 2010; Bulut-Karslioglu et al. 2014) that can directly impact gene expression, other common effects of RIs on targeted genes include premature transcript termination (Perepelitsa-Belancio and Deininger 2003; Han et al. 2004) and alternative post-transcriptional processing of genes (Belancio et al. 2006; Heras SR et al. 2014; Goodier and Kazazian, 2008; Cordaux and Batzer, 2009). Some of these functional impacts are generated by RIs inserted in genes because of the A/T richness of the LINE-1 sequence (Han et al. 2004) and due to the presence of a poly-A tail at the 3´end of the retrotransposon insertion (in both LINEs and SINEs), which can increase the repertoire of transcripts produced from the targeted gene (i.e., generating alternative transcripts). Similarly, Alu elements carry a functional polymarase-III promoter that can directly influence gene expression (Murphy and Baralle 1983); additionally, selected Alu insertions can affect the expression of targeted genes by additional mechanisms (Levanon et al. 2004; Pandey and Mukerji 2011; Elbarbary et al. 2013; Morales-Hernández et al. 2016). Indeed, the AluYg6 insertion on chr1q25.3 identified in this study (Figure 20) seems to directly impact EDEM3 gene expression, as an alternative annotated transcriptional variant in humans terminates precisely in the AluYg6 poly-A tail. Intriguingly, EDEM3 belongs to a group of proteins that accelerate degradation of misfolded or unassembled glycoproteins in the Endoplasmic Reticulum (Hirao et al. 2006). The EDEM3 gene also has a large number of associations in the functional network of neural-related genes containing AMH-specific insertions that are absent in both HN and HD, suggesting a strong relevance for EDEM3 in this network (Figure 18).

Another important effect that RIs can exert on gene expression, especially antisense insertions in respect to the gene's transcriptional orientation, is the alteration of the post-transcriptional processing of mRNAs, which can result in alternatively-spliced RNAs (Keren et al. 2010). Indeed, the identified AluYb9 insertion on chr10q25.3 that occurred in the 15th intron of the SHTN1 gene (Figures 21-22), is likely altering the post-transcriptional processing of SHTN1 mRNAs. Although based on computational predictions, the splicing-silencing peaks associated with the allele containing the AluYb9 sequence strongly suggest that the post-transcriptional processing of this gene is affected by the insertion, and we speculate that the AluYb9 sequence is inducing alternative splicing of SHTN1 transcripts. In sum, these data demonstrate how intronic RIs can contribute to the generation of novel functional variants exclusive to AMHs. Additionally, the gene SHTN1 is highly expressed in the human brain and is involved in the generation of internal asymmetric signals required for neuronal polarization and neurite outgrowth (Sapir et al. 2013); it is, as well, the gene with most interactions and relevance in the detected network of neural-related genes targeted by AMH-specific RIs vs both HN and HD (Figure 18). It is worth mentioning that previous studies demonstrated that the SHTN1 gene has undergone positive selection in AMHs, after the split with HN and HD (Weyer and Paabo, 2015). Accordingly, the AluYb9 insertion's corresponding genomic locus displays a TMRCA of ~560 Kya, which is consistent with possible positive selection and spread after the split between the AMH and HD/HN lineages. Thus, the above-mentioned novel functional variants, likely affected by the RI, might thus have contributed to the establishment of the selective process on this neural gene, which in turn may have affected our species differentiation.

CONCLUSIONS AND FUTURE PERSPECTIVES

The results presented in this study suggest that non-LTR retrotransposons-mediated processes might have played a more than marginal role in recent human evolution. RIs presence/absence polymorphisms in present-day populations can be useful phylogenetic markers and highlight interactions and population dynamics that occurred after the separation from the chimpanzee lineage. RIs display patterns of maintenance and diffusion in modern populations that reflect slow but constant generation of variability. As the new variants can be co-opted at a later moment, selective pressures can arise possibly inducing fixation or purification of those variants. Indeed, non-LTR retrotransposons activity results in an enrichment of pre- and post-transcriptional variants of genes in hominids and can directly generate new functionalities for human genes. This process is particularly evident in the pool of most-recent RIs (AMH-specific ones). In fact, these new variants seem to have been co-opted throughout the evolution of AMH and genes producing those variants are preferentially expressed in specific tissues. Co-optation of putatively RI-induced variants seems to have occurred especially in the brain, where they are related to neuronal maturation and migration, as well as synaptic-recognition; they are also associated to functionalities such as cognitive capability, vocalization behavior and locomotion regulation. Thus, RIs are possibly involved in the differentiation processes of the human brain and its increase in complexity that took place all throughout the evolution of the human lineage. In some instances, as for the AluYg6 insertion on chr1q25.3 and the AluYb9 insertion on chr10q25.3 (which are fixed or almost fixed in all AMH populations), the effects of these RIs on their target in *cis* might have been key contributors to the molecular differentiation of AMH genomes.

Since our study is limited to a few Neanderthal, Denisovan and Chimpanzee genomes, and because only RI-associated effects in *cis* could be analyzed, the impact of non-LTR retrotransposonsons on human evolution reported in this work probably reflects only the tip of a much larger iceberg.

A recent study (Gardner et al. 2017), in fact, identified a large number of polymorphic RIs in modern human populations, which have also been found as absent in the genomes of HN and HD. The aforementioned research was conducted as part of the 1000 Genomes project. As soon as the consortium will update the human genetic variability database with the new data, it will be possible to compare the RIs identified in this study with their annotations, in order to better characterize the state of fixation or polymorphism of the insertions themselves in present day populations. Precisely identifying where the annotated insertional variants are present or absent, together with the RNA-Seq analyses conducted by the 1KG Consortium, can allow for further inferences on functional alterations associated with RIs in the loci in which they inserted.

Furthermore, the comparison between our dataset of AMH-specific RIs with other ones can highlight more individual insertions that might display characteristics relevant to the evolution of AMH. The causal relationship between these RIs and the arousal new genomic functionalities/variants could then be tested *in vitro* on human Embryonic Stem Cells (hESCs) or induced Pluripotent Stem Cells (iPSCs). Selected variants can be excised or inserted in the cultured cell lines via CRISPR/Cas recombination methods; hESCs and iPSCs can then be differentiated into Neuronal Precursor Stem Cells (NPSCs) and even into mature neurons, in which alterations of gene expression or regulation can be measured in function of the presence/absence of the specific RI variants.

Finally, the impact of RIs in *trans* should also be studied more in depth: is it possible that long range RNA gene regulation operated by retrotransposon-derived lncRNA or siRNAs/miRNAs is a commonly-occurring process in our genome? Are epigenetic chromatin modifications that occur onto RI targeted loci, which can affect the expression/regulation of large genomic areas at a time, a relevant portion of how our genome is differentially regulated throughout development and tissue-differentiation? How many of the active promoter/enhancers that control gene expression have been originated by retrotransposon activity or are interacting with them?

Much research needs to be done in order to answer these questions and to further develop the knowledge of non-LTR retrotransposition's role and contribution in shaping, regulating and evolving eukaryote genomes, but every additional discovery adds a new layer of understanding in the seemingly unending quest of researchers towards the unraveling of the deep secrets of life and evolution.

REFERENCES

Ahern JCM, Karavanić I, Janković I, Smith FH. 2005. New discoveries and interpretations of hominid fossils and artifacts from Vindija Cave, Croatia. *Journal of Human Evolution* **49**(6): 781-782.

Aken BL, Achuthan P, Akanni W, Amode MR, Bernsdorff F, Bhai J, Billis K, Carvalho-Silva D, Cummins C, Clapham P et al. 2016. Ensembl 2017. *Nucleic Acids Res* **45:** D635-D642.

Baillie JK, Barnett MW, Upton KR, Gerhardt DJ, Richmond TA, De Sapio F, Brennan P, Rizzu P, Smith S, Fell M et al. 2011. Somatic retrotransposition alters the genetic landscape of the human brain. *Nature* **479**: 534-537.

Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* **6**: 11.

Beck CR, Garcia-Perez JL, Badge RM, Moran JV. 2011. LINE-1 Elements in Structural Variation and Disease. *Annu Rev Genomics Hum Genet* **12**:187-215.

Belancio VP, Hedges DJ, Deininger P. 2006. LINE-1 RNA splicing and influences on mammalian gene expression. *Nucleic Acids Res* **34**: 1512-21.

Bourc'his D, Bestor TH. 2004. Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. *Nature* **431**:96-9.

Briggs AW, Good JM, Green RE, Krause J, Maricic T, Stenzel U, Lalueza-Fox C, Rudan P, Brajković D, Kućan Z, et al. 2009) Targeted Retrieval and Analysis of Five Neandertal mtDNA Genomes. *Science* **325**: 318-321.

Bulut-Karslioglu A, De La Rosa-Velázquez IA, Ramirez F, Barenboim M, Onishi-Seebacher M, Arand J, Galán C, Winter GE, Engist B, Gerle B et al. 2014. Suv39h-dependent H3K9me3 marks intact retrotransposons and silences LINE elements in mouse embryonic stem cells. *Mol Cell* **55**: 277-290.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2008. BLAST+: architecture and applications. *BMC Bioinformatics* **10:** 421.

Christmas R, Avila-Campillo I, Bolouri H, Schwikowski B, Anderson M, Kelley R, Landys N, Workman C, Ideker T, Cerami E, Sheridan R et al. 2005. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Am Assoc Cancer Res Educ Book* **2005**: 12-16.

Chuong EB, Elde NC, Feschotte C. 2017. Regulatory activities of transposable elements: from conflicts to benefits. *Nat Rev Genet* **18**: 71-86.

Cordaux R, Batzer MA. 2009. The impact of retrotransposons on human genome evolution. *Nat Rev Genet* **10**: 691–703.

Deininger PL, Moran JV, Batzer MA, Kazazian HH Jr. 2003. Mobile elements and mammalian genome evolution. *Curr Opin Genet Dev* **13**:651-8.

Desmet FO, Hamroun D, Lalande M, Collod-Béroud G, Claustres M, Béroud C. 2009. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res* **37**: 9:67.

Dombroski BA, Mathias SL, Nanthakumar E, Scott AF, Kazazian HH Jr. 1991. Isolation of an active human transposable element. *Science* **254**: 1805-1808.

Doolittle WF, Sapienza C. 1980. Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284**: 601-603.

Elbarbary RA, Li W, Tian B, Maquat LE. 2013. STAU1 binding 3' UTR IRAlus complements nuclear retention to protect cells from PKR-mediated translational shutdown. *Genes Dev* **27**: 1495-510.

Evrony GD, Cai X, Lee E, Hills LB, Elhosary PC, Lehmann HS, Parker JJ, Atabay KD, Gilmore EC, Poduri A et al. 2012. Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell* **151**: 483-496.

Fedoroff NV. 2012. Transposable Elements, Epigenetics, and Genome Evolution. *Science* **338**: 758-767.

Feng Q, Moran JV, Kazazian HH Jr, Boeke JD. 1996. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **87**:905–916.

Feschotte C. 2008. The contribution of transposable elements to the evolution of regulatory networks. *Nat Rev Genet* **9**: 397-405.

Friedli M, Trono D. 2015. The Developmental Control of Transposable Elements and the Evolution of Higher Species*. Ann Rev of Cell and Dev Bio* **31**: 429-451.

Garcia-Perez JL, Widmann TJ, Adams IR. 2016. The impact of transposable elements on mammalian development. *Development* **143**: 4101-4114.

Garcia-Perez JL, Morell M, Scheys JO, Kulpa DA, Morell S, Carter CC, Hammer GD, Collins KL, O'Shea KS, Menendez P et al. 2010. Epigenetic silencing of engineered L1 retrotransposition events in human embryonic carcinoma cells. *Nature* **466**: 769-773.

Gardner EJ, Lam VK, Harris DN, Chuang NT, Scott EC, Pittard WS, Mills RE, 1000 Genomes Project Consortium, Devine SE. 2017. The Mobile Element Locator Tool (MELT): Population-scale mobile element discovery and biology. *Genome Res* doi: 10.1101/gr.218032.116.

Gerdes P, Richardson S, Mager D, Faulkner G. 2016. Transposable elements in the mammalian embryo: pioneers surviving through stealth and service. *Genome Biol* **17**: 100.

Goodier JL, Kazazian HH Jr. 2008. Retrotransposons Revisited: The Restraint and Rehabilitation of Parasites. *Cell* **135(1)**: 23-35.

Goodier JL. 2016. Restricting retrotransposons: a review. *Mobile DNA* **7**: 16.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Hsi-Yang Fritz M et al. 2010. A draft sequence of the Neanderthal genome. *Science* **328**: 710–722.

Han JS, Szak ST, Boeke JD. 2004. Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. *Nature* **429**: 268-274.

Heras SR, Macias S, Cáceres JF, Garcia-Perez JL. 2014. Control of mammalian retrotransposons by cellular RNA processing activities. *Mob Genet Elements* **4**: e28439.

Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, Sella G, Przeworski M, 1000 Genomes Project. 2011. Classic selective sweeps were rare in recent human evolution. *Science* **331**: 920–924.

Hirao K, Natsuka Y, Tamura T, Wada I, Morito D, Natsuka S, Romero P, Sleno B, Tremblay LO, Herscovics A, et al. 2006. EDEM3, a soluble EDEM homolog, enhances glycoprotein endoplasmic reticulum-associated degradation and mannose trimming. *J Biol Chem* **281**: 9650–9658.

Hohjoh H, Singer MF. 1997. Sequence-specific single-strand RNA binding protein encoded by the human LINE-1 retrotransposon. *EMBO J* **16**:6034–6043.

Holloway RL. 1985. The poor brain of Homo sapiens neanderthalensis: see what you please. In Delson, E. Ancestors: The hard evidence. New York: Alan R. Liss.

Hormozdiari F, Konkel MK, Prado-Martinez J, Chiatante G, Herraez IH, Walker JA, Nelson B, Alkan C, Sudmant PH, Huddleston J, et al. 2013. Rates and patterns of great ape retrotransposition. *PNAS* **110**(33): 13457-13462.

Hrvoj-Mihic B, Bienvenu T, Stefanacci L, Muotri AR, Semendeferi K. 2013. Evolution, development, and plasticity of the human brain: from molecules to bones. *Front Hum Neurosci* **7**: 707.

Huang DW, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protocols* **4**: 44-57.

Huang S, Tao X, Yuan S, Zhang Y, Li P, Beilinson HA, Zhang Y, Yu W, Pontarotti P, Escriva H et al. 2016. Discovery of an Active RAG Transposon Illuminates the Origins of V(D)J Recombination. *Cell* **166**: 102-114.

Hublin JJ, Ben-Ncer A, Bailey SE, Freidline SE, Neubauer S, Skinner MM, Bergmann I, Le Cabec A, Benazzi S, Harvati K, Gunz P. 20172. New fossils from Jebel Irhoud, Morocco and the pan-African origin of Homo sapiens. *Nature* **546**: 289-292.

Inchley CE, Larbey CD, Shwan NA, Pagani L, Saag L, Antão T, Jacobs G, Hudjashov G, Metspalu E, Mitt M et al. 2016. Selective sweep on human amylase genes postdates the split with Neanderthals. *Sci Rep* **6**: 37198.

Kaimal V, Bardes EE, Tabar SC, Jegga AG, Aronow BJ. 2010. ToppCluster: a multiple gene list feature analyzer for comparative enrichment clustering and network-based dissection of biological systems. *Nucleic Acids Res* **38**: W96–W102.

Kapitonov V, Jurka J. 2005. RAG1 Core and V(D)J Recombination Signal Sequences Were Derived from Transib Transposons. *PLoS Biol* **3**: e181.

Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, Yandell M, Feschotte C. 2013. Transposable elements are major contributors to the origin, diversification and regulation of vertebrate long noncoding RNAs. *PLoS Genet* **9** doi: 10.1371/journal.pgen.1003470.

Kazazian HH Jr, Wong C, Youssoufian H, Scott AF, Phillips DG, Antonarakis SE. 1988. Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature* **332**: 164-166.

Keane TM, Wong K, Adams DJ. 2012. RetroSeq: transposable element discovery from next-generation sequencing data. *Bioinformatics* **29**: 389-390.

Keren H, Lev-Maor G, Ast G. 2010. Alternative splicing and evolution: diversification, exon definition and function. *Nat Rev Genet* **11**: 345–355.

Koonin E, Krupovic M. 2014. Evolution of adaptive immunity from transposable elements combined with innate immune systems. *Nat Rev Genet* **16**: 184-192.

Kriegs JO, Churakov G, Jurka J, Brosius J, Schmitz J. 2007. Evolutionary history of 7SL RNA-derived SINEs in Supraprimates. *Trends Genet* **23**(4): 158–61.

Krings M, Stone A, Schmitz RW, Krainitzki H, Stoneking M, Pääbo S. 1997. Neandertal DNA sequences and the origin of modern humans. *Cell* **90**: 19-30.

Kuhlwilm M, Gronau I, Hubisz MJ, et al. 2016. Ancient gene flow from early modern humans into Eastern Neanderthals. *Nature* **530**(7591): 429-433.

Kulpa DA, Moran JV. 2005. Ribonucleoprotein particle formation is necessary but not sufficient for LINE-1 retrotransposition. *Hum Mol Genet* **14**:3237–3248.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.

Lavialle C, Cornelis G, Dupressoir A, Esnault C, Heidmann O, Vernochet C, Heidmann T. 2013. Paleovirology of 'syncityns', retroviral env genes exapted for a role in placentation. *Philos Trans R Soc B Biol Sci* **368**: 20120507 doi: 10.1098/rstb.2012.0507.

Lee E, Iskow R, Yang L, Gokcumen O, Haseley P, Luquette LJ, Lohr JG, Harris CC, Ding L, Wilson RK et al. 2012. Landscape of somatic retrotransposition in human cancers. *Science* **337**: 967-971.

Lee J, Cordaux R, Han K, Wang J, Hedges DJ, Liang P, Batzer MA. 2007. Different evolutionary fates of recently integrated human and chimpanzee LINE-1 retrotransposons. *Gene* **390**(1-2): 18–27.

Levanon EY, Eisenberg E, Yelin R, Nemzer S, Hallegger M, Shemesh R, Fligelman ZY, Shoshan A, Pollock SR, Sztybel D et al. 2004. Systematic identification of abundant A-to-I editing sites in the human transcriptome. *Nat Biotechnol* **22**: 1001-5.

Lynch V, Leclerc R, May G, Wagner G. 2011. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nat Genet* **43**: 1154-1159.

Macia A, Muñoz-Lopez M, Cortes JL, Hastings RK, Morell S, Lucena-Aguilar G, Marchal JA, Badge RM, Garcia-Perez JL. 2011. Epigenetic control of retrotransposon expression in human embryonic stem cells. *Mol Cell Biol* **31**: 300-316.

Malone CD, Hannon GJ. 2009. Small RNAs as guardians of the genome. *Cell* **136**:656-68.

Mathias SL, Scott AF, Kazazian HH Jr, Boeke JD, Gabriel A. 1991. Reverse transcriptase encoded by a human transposable element. *Science* **254**:1808–1810.

Meyer M, Kircher M, Gansauge M, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prüfer K, de Filippo C et al. 2012. A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**: 222–226.

Montinaro F, Busby GB, Pascali VL, Myers S, Hellenthal G, Capelli C. 2015. Unravelling the hidden ancestry of American admixed populations. *Nat Commun* **6**:6596 doi: 10.1038/ncomms7596.

Morales-Hernández A, González-Rico FJ, Román AC, Rico-Leo E, Alvarez-Barrientos A, Sánchez L, Macia Á, Heras SR, García-Pérez JL, Merino JM et al. 2016. Alu retrotransposons promote differentiation of human carcinoma cells through the aryl hydrocarbon receptor. *Nucleic Acids Res* **44**: 4665-83.

Moran JV, DeBernardinis RJ, Kazazian HH Jr. 1999. Exon shuffling by L1 retrotransposition. *Science* **283**: 1530-1534.

Moran JV, Gilbert N. 2002. Mammalian LINE-1 retrotransposons and related elements. *Mobile DNA II*: 836–869.

Moran JV, Holmes SE, Naas TP, DeBerardinis RJ, Boeke JD, Kazazian HH Jr. 1996. High frequency retrotransposition in cultured mammalian cells. *Cell* **87**:917-27.

Morrish TA, Garcia-Perez JL, Stamato TD, Taccioli GE, Sekiguchi J, Moran JV. 2007. Endonuclease-independent LINE-1 retrotransposition at mammalian telomeres. *Nature* **446**:208-12.

Morrish TA, Gilbert N, Myers JS, Vincent BJ, Stamato TD, Taccioli GE, et al. 2002. DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. *Nat Genet* **31**:159-65.

Muñoz-Lopez M, Macia A, Garcia-Cañadas M, Badge RM, Garcia-Perez JL. 2011. An epi[c]genetic battle. *Mobile Genetic Elements* **1**:2, 122-127.

Muotri AR, Chu VT, Marchetto MC, Deng V, Moran JV, Gage FH. 2005. Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature* **435**: 903-910.

Murphy MH, Baralle FE. 1983. Directed semisynthetic point mutational analysis of an RNA polymerase III promoter. *Nucleic Acids Res* **11**: 7695-700.

Notwell J, Chung T, Heavner W, Bejerano G. 2015. A family of transposable elements co-opted into developmental enhancers in the mouse neocortex. *Nat Commun* **6**: 6644.

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 Biol* **4**: R74.

Orgel LE, Crick FH. 1980. Selfish DNA: the ultimate parasite. *Nature* **284**: 604-607.

Pagani L, Lawson DJ, Jagoda E, Mörseburg A, Eriksson A, Mitt M, Clemente F, Hudjashov G, DeGiorgio M, Saag L, et al. 2016. Genomic analyses inform on migration events during the peopling of Eurasia. *Nature* **538**: 238-242.

Pandey R, Mukerji M. 2011. From 'JUNK' to just unexplored noncoding knowledge: the case of transcribed Alus. *Brief Funct Genomics* **10**: 294-311.

Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**: 289-290.

Pearce E, Stringer C, Dunbar RIM. 2013. New insights into differences in brain organization between Neanderthals and anatomically modern humans. *Proc R Soc B* **280**(1758): 20130168.

Perepelitsa-Belancio V, Deininger P. 2003. RNA truncation by premature polyadenylation attenuates human mobile element activity. *Nat Genet* **35**: 363-366.

Pinhasi R, Higham TFG, Golovanova LV, Doronichev VB. 2011. Revised age of late Neanderthal occupation and the end of the Middle Paleolithic in the northern Caucasus. *PNAS* **108**(21): 8611-8616.

Posth C, Wissing C, Kitagawa K, Pagani L, Van Holstein L, Racimo F, Wehrberger K, Conard NJ, Kind CJ, Bocherens H, Krause J. 2017. Deeply divergent archaic mitochondrial genome provides lower time boundary for African gene flow into Neanderthals. *Nature Communications* **8**:16046 DOI: 10.1038/ncomms16046.

Pritchard JK, Pickrell JK, Coop G. 2010. The Genetics of Human Adaptation: Hard Sweeps, Soft Sweeps, and Polygenic Adaptation. *Curr Biol* **20**: R208–R215.

Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, de Filippo C, Li H et al. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**: 43–49.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.

Racimo F. 2016. Testing for Ancient Selection Using Cross-population Allele Frequency Differentiation. *Genetics* **202**: 733-50.

Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PLF et al. 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, **468**: 1053-1060.

Richardson SR, Doucet AJ, Kopera HC, Moldovan JB, García-Pérez JL, Moran JV. 2015. The Influence of LINE-1 and SINE Retrotransposons on Mammalian Genomes. *Microbiol Spectr* **3**: MDNA3-0061-2014

Sánchez-Quinto F, Lalueza-Fox C. 2015. Almost 20 years of Neanderthal palaeogenetics: adaptation, admixture, diversity, demography and extinction. *Philosophical Transactions of the Royal Society B: Biological Sciences* **370**(1660): 20130374.

Sapir T, Levy T, Sakakibara A, Rabinkov A, Miyata T, Reiner O. 2013. Shootin1 acts in concert with KIF20B to promote polarization of migrating neurons. *J Neurosci* **33**: 11932-48.

Sayah DM, Sokolskaja E, Berthoux L, Luban J. 2004. Cyclophilin A retrotransposition into TRIM5 explains owl monkey resistance to HIV-1. *Nature* **430**: 569-573.

Sawyer S, Renaud G, Viola B, Hublinc J, Gansaugea M, Shunkov MV, Dereviankod AP, Prüfer K, Kelso J, Pääbo S et al. 2015. Nuclear and mitochondrial DNA sequences form two Denisovan individuals. *Proc Natl Acad Sci* **112**: 15696-15700.

Schlebusch CM, Malmström H, Günther T, Sjödin P, Coutinho A, Edlund H, Munters AR, 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* **358**: 652–655.

Schlicker A, Domingues FS, Rahnenführer J, T Lengauer. 2006. A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics* **7**: 302.

Schmitz RW, Serre D, Bonani G, et al. 2002. The Neandertal type site revisited: Interdisciplinary investigations of skeletal remains from the Neander Valley, Germany. *PNAS* **99**(20): 13342-13347.

Scott AF, Schmeckpeper BJ, Abdelrazik M, Comey CT, O'Hara B, Rossiter JP, Cooley T, Heath P, Smith KD, Margolet L. 1987. Origin of the human L1 elements: proposed progenitor genes deduced from a consensus DNA sequence. *Genomics* **1**: 113–125.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. 2009. ABySS: a parallel assembler for short read sequence data. *Genome Res* **19**: 1117-23.

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

Speek M. 2001. Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. *Molecular Cell Biology* **21**: 1973-1985.

Speir ML, Zweig AS, Rosenbloom KR, Raney BJ, Paten B, Nejad P, Lee BT, Learned K, Karolchik D, Hinrichs AS et al. 2015. The UCSC genome browser database: 2016 update. *Nucleic Acids Res* **44**: D717-D725.

Stetson DB, Ko JS, Heidmann T, Medzhitov R. 2008. Trex1 prevents cell-intrinsic initiation of autoimmunity. *Cell* **134**:587-98.

Stringer C. 1984. Human evolution and biological adaptation in the Pleistocene. In Foley, R. Hominid evolution and community ecology. New York: Academic Press.

Supek F, Bošnjak M, Škunca N, Šmuc T. 2011. REVIGO summarizes and visualizes long lists of Gene Ontology terms. *PLoS ONE* **6**: e21800.

Suzuki J, Yamaguchi K, Kajikawa M, Ichiyanagi K, Adachi N, Koyama H, et al. 2009. Genetic Evidence That the Non-Homologous End-Joining Repair Pathway Is Involved in LINE Retrotransposition. *PLoS Genet* **5**:1000461.

Swergold GD. 1990. Identification, characterization, and cell specificity of a human LINE-1 promoter. *Mol Cell Biol* **10**: 6718-29.

Taylor MS, LaCava J, Mita P, Molloy KR, Huang CRL, Li D, Adney EM, Jiang H, Burns KH, Chait BT, et al. 2013. Affinity Proteomics Reveals Human Host Factors Implicated in Discrete Stages of LINE-1 Retrotransposition. *Cell* **155**:1034-1048.

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation, *Nature* **526**: 68-74.

Wang H, Xing J, Grover D, Hedges DJ, Han K, Walker JA, Batzer MA. 2005. SVA elements: a hominid-specific retroposon family. *J Mol Biol* **354**: 994–1007.

Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, Franz M, Grouios C, Kazi F, Tannus Lopes C, et al. 2010. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res* **38**: W214–W220.

Wei W, Gilbert N, Ooi SL, et al. 2001. Human L1 Retrotransposition: cis Preference versus trans Complementation. *Molecular and Cellular Biology* **21**: 1429-39.

Weyer S, Paabo S. 2015. Functional Analyses of Transcription Factor Binding Sites that Differ between Present-Day and Archaic Humans. *Mol Biol Evol* **33**: 316-322.

Wissing S, Montano M, Garcia-Perez JL, Moran JV, Greene WC. 2011. Endogenous APOBEC3B Restricts LINE-1 Retrotransposition in Transformed Cells and Human Embryonic Stem Cells. *The Journal of Biological Chemistry* **286**: 36427–36437.

Wu J, Lee WP, Ward A, Walker JA, Konkel MK, Batzer MA, Marth GT. 2014. Tangram: a comprehensive toolbox for mobile element insertion detection. *BMC genomics* **15**: 795.

Yang N, Kazazian HH Jr. 2006. L1 retrotransposition is suppressed by endogenously encoded small interfering RNAs in human cultured cells. *Nat Struct Mol Biol* **13**:763-71.

Yoder JA, Walsh CP, Bestor TH. 1997. Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet* **13**: 335-340.

Zhao K, Du J, Han X, Goodier JL, Li P, Zhou X, Wei W, Evans SL, Li L, Zhang W, et al. 2013. Modulation of LINE-1 and Alu/SVA Retrotransposition by Aicardi-Goutières Syndrome-Related SAMHD1. *Cell Rep* **4**(6):1108-1115.

SUPPLEMENTAL MATERIAL

CODE	CHROMOSOME	POSITION	ELEMENT	SENSE
filledsite100se	chr1	234587112-234587612	AluYa5	+
filledsite103se	chr1	240616908-240617419	AluYb8	+
filledsite106se	chr1	247027794-247028274	AluYb8	+
filledsite107se	chr1	247850376-247856637	L1HS	+
filledsite119se	chr1	31040561-31041744	L1HS	+
filledsite126se	chr1	43857371-43857880	AluYb3a1	+
filledsite130an	chr1	210295216-210295721	AluYa5	C
filledsite137se	chr1	56831025-56835069	L1HS	+
filledsite141an	chr1	219923218-219923732	AluYb8	C
filledsite146an	chr1	224392729-224393233	AluYa1	C
filledsite147an	chr1	228572083-228572589	AluYa5	C
filledsite152an	chr1	232423030-232423545	AluYb8	C
filledsite153an	chr1	232775437-232775942	AluYa5	C
filledsite155an	chr1	236054499-236055012	AluYb8	C
filledsite15se	chr1	117078082-117078588	AluY	+
filledsite160an	chr1	241360451-241360952	AluYa5	C
filledsite16se	chr1	118090952-118091360	AluYa5	+
filledsite172an	chr1	26489727-26490233	AluYa5	C
filledsite177se	chr1	81687698-81688211	AluYb8	+
filledsite178se	chr1	83125877-83127603	L1HS	+
filledsite195an	chr1	58343117-58343634	AluYb8	C
filledsite211an	chr1	66024110-66030359	L1HS	C

filledsite21an	chr1	111701427-111701934	AluYa4	C
filledsite224an	chr1	71741251-71741728	AluYb8	C
filledsite234an	chr1	74984430-74985354	L1HS	C
filledsite236an	chr1	77100276-77100756	AluYa5	C
filledsite248an	chr1	83042557-83043063	AluYa5	C
filledsite249an	chr1	85748534-85749049	AluYb8	C
filledsite27an	chr1	119558685-119559191	AluYg6	C
filledsite27se	chr1	150691352-150691857	AluYa5	+
filledsite2se	chr1	103047593-103048099	AluYc1	+
filledsite83se	chr1	217753990-217754482	AluYb9	+
filledsite87an	chr1	182289232-182289735	AluYa5	C
filledsite90an	chr1	184689543-184689854	AluYg6	C
filledsite99an	chr1	188994870-188995383	AluYb8	C
filledsite207se	chr10	10708444-10708950	AluYc1	+
filledsite214se	chr10	111036294-111036800	AluYg6	+
filledsite228se	chr10	118664331-118664840	AluYb9	+
filledsite239se	chr10	133448995-133449508	AluYb8	+
filledsite247se	chr10	20667160-20667667	AluYa5	+
filledsite253se	chr10	28112686-28113193	AluYe5	+
filledsite258se	chr10	32705638-32706145	AluYc1	+
filledsite259se	chr10	34684384-34684895	AluYb8	+
filledsite264se	chr10	43274113-43274619	AluYa4	+
filledsite267se	chr10	4634712-4635218	AluYa5	+
filledsite272se	chr10	49917071-49917581	AluYb8	+

filledsite285an	chr10	10493319-10493832	AluYb8	C
filledsite296an	chr10	111572047-111578314	L1HS	C
filledsite297se	chr10	70272013-70272521	AluYa5	+
filledsite308an	chr10	119634190-119634704	AluYb9	C
filledsite314an	chr10	123489087-123489593	AluYa5	C
filledsite324an	chr10	129565205-129565707	AluYa5	C
filledsite326an	chr10	132858414-132858920	AluYa5	C
filledsite329an	chr10	14936258-14936764	AluYf1	C
filledsite368an	chr10	54860721-54861229	AluYa5	C
filledsite371an	chr10	56516936-56517440	AluYg6	C
filledsite374an	chr10	57701958-57702464	AluYa5	C
filledsite384an	chr10	63743531-63743894	AluYb8	C
filledsite411an	chr10	84400678-84401005	AluYf2	C
filledsite432an	chr10	95853591-95854096	AluYa5	C
filledsite352se	chr11	132558495-132558977	AluYb8	+
filledsite354se	chr11	133302953-133303265	AluYa8	+
filledsite358se	chr11	14348225-14348731	AluYa5	+
filledsite361se	chr11	16597479-16597985	AluYa5	+
filledsite368se	chr11	24349399-24355652	L1HS	+
filledsite403se	chr11	4802436-4802909	AluYb8	+
filledsite408se	chr11	55667320-55667821	AluYc1	+
filledsite432se	chr11	82306796-82307302	AluYa5	+
filledsite440se	chr11	92245907-92246403	AluYc1	+
filledsite443se	chr11	96233033-96233535	AluYc1	+

filledsite448an	chr11	105765417-105765923	AluYb8	C
filledsite451an	chr11	106595790-106596263	AluYa5	C
filledsite514an	chr11	27307785-27308301	AluYb9	C
filledsite527an	chr11	33537589-33538095	AluY	C
filledsite541an	chr11	39805045-39805549	AluYa5	C
filledsite545an	chr11	41006927-41007439	AluYb8	C
filledsite550an	chr11	42464000-42464508	AluYb9	C
filledsite573an	chr11	56407853-56408356	AluYa5	C
filledsite587an	chr11	71861504-71862011	AluYb8	C
filledsite601an	chr11	81556131-81556637	AluYa5	C
filledsite461se	chr12	127502539-127502947	AluYb8	+
filledsite462se	chr12	127660343-127660851	AluYa5	+
filledsite478se	chr12	26510938-26511444	AluYc1	+
filledsite497se	chr12	41659732-41660237	AluYa5	+
filledsite504se	chr12	45011089-45011600	AluYa5	+
filledsite513se	chr12	54245322-54245801	AluYb8	+
filledsite530se	chr12	63232320-63232826	AluY	+
filledsite640an	chr12	102313971-102314263	AluYb9	C
filledsite643an	chr12	104293317-104293823	AluYa5	C
filledsite573se	chr13	101906861-101907367	AluYa5	+
filledsite578se	chr13	103931553-103932059	AluYa5	+
filledsite582se	chr13	105161120-105161626	AluYa5	+
filledsite588se	chr13	111232765-111233269	AluYc1	+
filledsite618se	chr13	48419833-48420181	AluYa5	+

filledsite631se	chr13	63656094-63656600	AluY	+
filledsite650se	chr13	82889112-82889625	AluYb8	+
filledsite660se	chr13	89522726-89523232	AluYa5	+
filledsite663se	chr13	91159954-91160391	AluYb8	+
filledsite673se	chr13	96633337-96633842	AluYh9	+
filledsite677se	chr14	100844914-100845420	AluYa5	+
filledsite702se	chr14	35305204-35305710	AluYc1	+
filledsite726se	chr14	48334154-48334660	AluY	+
filledsite727se	chr14	48826723-48827236	AluYb8	+
filledsite730se	chr14	50118788-50119107	AluYf2	+
filledsite738se	chr14	59545214-59545711	AluYa5	+
filledsite741se	chr14	63887389-63887896	AluYa5	+
filledsite749se	chr14	79653838-79654129	AluSg7	+
filledsite756se	chr14	81538996-81539466	AluYa5	+
filledsite768se	chr14	87555802-87556308	AluYa5	+
filledsite773se	chr15	24206969-24207470	AluYa5	+
filledsite783se	chr15	43660342-43660849	AluYa5	+
filledsite798se	chr15	52525095-52525603	AluYa5	+
filledsite804se	chr15	58282022-58282319	AluYb9	+
filledsite811se	chr15	69532469-69532748	AluYa8	+
filledsite814se	chr15	71885382-71885888	AluYa5	+
filledsite821se	chr15	79891757-79892263	AluYc1	+
filledsite832se	chr15	93515700-93516202	AluYa5	+
filledsite1251an	chr16	63442347-63442853	AluYa4	C

filledsite842se	chr16	19230343-19230856	AluYb9	+
filledsite844se	chr16	20778458-20778964	AluY	+
filledsite856se	chr16	49035989-49036291	AluYa8	+
filledsite871se	chr16	62046777-62047291	AluYb9	+
filledsite878se	chr16	67513721-67514227	AluYa5	+
filledsite883se	chr16	74289332-74289844	AluYb8	+
filledsite895se	chr17	11557291-11557795	AluYi6	+
filledsite905se	chr17	29932375-29932875	AluYa5	+
filledsite938se	chr17	6874635-6875132	AluYe2	+
filledsite942se	chr17	69852055-69852568	AluYb8	+
filledsite945se	chr17	75988198-75988699	AluYg6	+
filledsite1001se	chr18	57258095-57258603	AluYi6	+
filledsite1009se	chr18	67949511-67950022	AluY	+
filledsite1010se	chr18	68846641-68847147	AluYa5	+
filledsite1013se	chr18	73569922-73570434	AluYb8	+
filledsite1018se	chr18	9385305-9385819	AluYb8	+
filledsite947se	chr18	11971830-11972336	AluYa4	+
filledsite948se	chr18	12851097-12851610	AluYb8	+
filledsite956se	chr18	22892394-22892905	AluYb8	+
filledsite968se	chr18	32661873-32662376	AluYa4	+
filledsite985se	chr18	41166600-41167103	AluYb9	+
filledsite991se	chr18	47019769-47020275	AluYa5	+
filledsite994se	chr18	47910920-47911370	L1HS	+
filledsite1025se	chr19	23320704-23321210	AluYa5	+

filledsite1026se	chr19	23986311-23986817	AluY	+
filledsite1038se	chr19	51457377-51457884	AluYa5	+
filledsite1069se	chr2	118906040-118906378	AluYa8	+
filledsite1078se	chr2	128871173-128871671	AluYf2	+
filledsite1080se	chr2	131844255-131844758	AluYe5	+
filledsite1104se	chr2	151902990-151903458	AluYa5	+
filledsite1119se	chr2	168895721-168896232	AluYb8	+
filledsite1134se	chr2	182355883-182356388	AluYg6	+
filledsite1136se	chr2	183315724-183316226	AluYa5	+
filledsite1149se	chr2	194398409-194398915	AluYa5	+
filledsite1153se	chr2	196051732-196052191	L1HS	+
filledsite1161se	chr2	20666579-20667076	AluYa5	+
filledsite1162se	chr2	206731099-206731605	AluYa4	+
filledsite1173se	chr2	213574587-213575081	AluYa5	+
filledsite1177se	chr2	21572694-21573180	AluYc1	+
filledsite1184se	chr2	224473015-224473438	AluYb8	+
filledsite1191se	chr2	231996096-231996601	AluYa5	+
filledsite1196se	chr2	239409282-239409788	AluYa5	+
filledsite1221se	chr2	5209090-5209596	AluYg6	+
filledsite1237se	chr2	60197805-60198304	AluYa5	+
filledsite1269se	chr2	85335271-85335779	AluYb8	+
filledsite1270se	chr2	8616711-8617202	AluYb8	+
filledsite1275se	chr2	89029145-89032380	L1HS	+
filledsite1287se	chr20	14980888-14981401	AluYb9	+

filledsite1302se	chr20	2863917-2864423	AluY	+
filledsite1305se	chr20	33115736-33116242	AluYa5	+
filledsite1309se	chr20	37139172-37139674	AluYa4	+
filledsite1318se	chr20	43295470-43295976	AluYa5	+
filledsite1324se	chr20	57723354-57723860	AluYg6	+
filledsite1325se	chr20	58926628-58927141	AluYb8	+
filledsite1326se	chr20	60526203-60526699	AluYa5	+
filledsite1327se	chr20	60570539-60571051	AluYb9	+
filledsite1328se	chr20	6150860-6151366	AluYa5	+
filledsite1350se	chr21	24742147-24742653	AluYa4	+
filledsite1351se	chr21	26825047-26825501	L1HS	+
filledsite1353se	chr21	30601727-30602240	AluYb8	+
filledsite1360se	chr21	42402797-42403304	AluYa5	+
filledsite1437se	chr3	160237201-160237714	AluYb9	+
filledsite1438se	chr3	161246629-161247135	AluYa5	+
filledsite1458se	chr3	17512965-17513288	AluYf2	+
filledsite1476se	chr3	192063095-192063593	AluYc1	+
filledsite1532se	chr3	68890116-68890597	AluYa5	+
filledsite1533se	chr3	72379627-72380108	AluYa5	+
filledsite1540se	chr3	81345982-81346485	AluYa5	+
filledsite1543se	chr3	81907408-81907915	AluYa5	+
filledsite1573se	chr4	111808666-111809164	AluYb8	+
filledsite1579se	chr4	113469889-113470386	AluYa4	+
filledsite1631se	chr4	150160957-150161474	AluYb8	+

filledsite1647se	chr4	163232315-163232819	AluYa5	+
filledsite1673se	chr4	177869233-177869740	AluYa5	+
filledsite1674se	chr4	178514158-178514667	AluYb8	+
filledsite1683se	chr4	186155455-186155956	AluY	+
filledsite1693se	chr4	21160913-21167153	L1HS	+
filledsite1702se	chr4	28564202-28564708	AluYc1	+
filledsite1703se	chr4	31406587-31407093	AluYa5	+
filledsite1705se	chr4	33473320-33473826	AluYa5	+
filledsite1707se	chr4	36469822-36470328	AluYa4	+
filledsite1711se	chr4	37766193-37766706	AluYb8	+
filledsite1733se	chr4	55808586-55809025	AluYa5	+
filledsite1736se	chr4	58887179-58888432	L1HS	+
filledsite1737se	chr4	59944466-59950701	L1HS	+
filledsite1745se	chr4	65199357-65199873	AluYb8	+
filledsite1750se	chr4	67537849-67538353	AluYa5	+
filledsite1756se	chr4	71926318-71926667	AluYd3	+
filledsite1761se	chr4	78236189-78236645	L1HS	+
filledsite1766se	chr4	80205643-80206158	AluYb8	+
filledsite1770se	chr4	81991431-81991937	AluYc1	+
filledsite1775se	chr4	87066267-87066638	AluYf2	+
filledsite1789se	chr4	97572015-97572529	AluYb9	+
filledsite1790se	chr4	98110557-98111063	AluYa5	+
filledsite1792se	chr4	98590566-98591074	AluYb8	+
filledsite1800se	chr5	103030468-103030974	AluYc2	+

filledsite1802se	chr5	103854190-103860433	L1HS	+
filledsite1816se	chr5	122274254-122274760	AluY	+
filledsite1829se	chr5	130489674-130490178	AluY	+
filledsite1834se	chr5	13894993-13895486	AluYa5	+
filledsite1835se	chr5	143353551-143353878	AluYb9	+
filledsite1840se	chr5	150106957-150107463	AluYa4	+
filledsite1851se	chr5	158592309-158592816	AluYa5	+
filledsite1858se	chr5	162648049-162648556	AluYg6	+
filledsite1863se	chr5	170158722-170159002	AluYa8	+
filledsite1865se	chr5	170857411-170857924	AluYb8	+
filledsite1866se	chr5	172889629-172890135	AluYa5	+
filledsite1877se	chr5	24452753-24453266	AluYb8	+
filledsite1878se	chr5	25275312-25275823	AluYb8	+
filledsite1881se	chr5	26111103-26111609	AluYa5	+
filledsite1899se	chr5	37293720-37294233	AluYb9	+
filledsite1904se	chr5	41599347-41600228	L1HS	+
filledsite1905se	chr5	41603904-41604414	AluYb8	+
filledsite1916se	chr5	53176842-53177337	AluYa5	+
filledsite1927se	chr5	61293541-61294052	AluYb8	+
filledsite1940se	chr5	80039282-80039787	AluY	+
filledsite1942se	chr5	80751057-80751570	AluYb8	+
filledsite1947se	chr5	82156633-82157137	AluYa5	+
filledsite1955se	chr5	88031930-88032443	AluYb8	+
filledsite1963se	chr5	9411674-9412179	AluYa5	+

filledsite1966se	chr5	97843851-97844355	AluYc1	+
filledsite1973se	chr5	99759376-99759882	AluYg6	+
filledsite1974se	chr5	99923448-99923942	AluYc1	+
filledsite1978se	chr6	10070706-10071205	AluYa5	+
filledsite1987se	chr6	112810812-112811319	AluYf2	+
filledsite1993se	chr6	11556008-11556514	AluY	+
filledsite2004se	chr6	119417420-119417932	AluYb8	+
filledsite2008se	chr6	123514969-123515478	AluYe5	+
filledsite2017se	chr6	128989732-128990245	AluYb8	+
filledsite2018se	chr6	129319430-129325678	L1HS	+
filledsite2042se	chr6	158548190-158549147	SVA_F	+
filledsite2062se	chr6	22227378-22227850	AluYa5	+
filledsite2067se	chr6	25529785-25530108	AluYb9	+
filledsite2081se	chr6	44613399-44613905	AluYa5	+
filledsite2087se	chr6	47280054-47280336	AluYb9	+
filledsite2093se	chr6	51256201-51256706	AluYc1	+
filledsite2108se	chr6	65185063-65185569	AluYa5	+
filledsite2138se	chr6	9014733-9015024	AluYa8	+
filledsite2143se	chr6	9460293-9460813	AluYb8	+
filledsite2942an	chr6	23300277-23300783	AluY	C
filledsite2155se	chr7	102343341-102343849	AluYa5	+
filledsite2156se	chr7	102488622-102489128	AluY	+
filledsite2158se	chr7	103462980-103463491	AluYb8	+
filledsite2165se	chr7	108352758-108353271	AluYb8	+

filledsite2168se	chr7	110542599-110543112	AluYa4	+
filledsite2175se	chr7	113953419-113953923	AluYa4	+
filledsite2181se	chr7	11817394-11817908	AluYa5	+
filledsite2189se	chr7	123636845-123637358	AluYb8	+
filledsite2193se	chr7	126775521-126775797	AluSg7	+
filledsite2198se	chr7	133107591-133108105	AluYb9	+
filledsite2224se	chr7	154941833-154942346	AluYb8	+
filledsite2242se	chr7	29476833-29477329	AluY	+
filledsite2244se	chr7	31070490-31071003	AluYb8	+
filledsite2258se	chr7	38512923-38513436	AluYb8	+
filledsite2272se	chr7	51270816-51271282	AluY	+
filledsite2276se	chr7	52094874-52095380	AluYc1	+
filledsite2291se	chr7	79165233-79165729	AluY	+
filledsite2317se	chr7	96894648-96895078	AluYa5	+
filledsite2323se	chr8	106181311-106181825	AluYb8	+
filledsite2324se	chr8	107511009-107511521	AluYb8	+
filledsite2325se	chr8	108120135-108120642	AluYb8	+
filledsite2332se	chr8	113085336-113085852	AluYb8	+
filledsite2348se	chr8	126595030-126601231	L1HS	+
filledsite2361se	chr8	136966696-136967816	L1HS	+
filledsite2372se	chr8	16962123-16962629	AluYc1	+
filledsite2380se	chr8	27662417-27662930	AluYb8	+
filledsite2381se	chr8	27683961-27684433	AluY	+
filledsite2386se	chr8	30392191-30392673	AluYb8	+

filledsite2417se	chr8	5748088-5748601	AluYb9	+
filledsite2431se	chr8	69417585-69418095	AluYb8	+
filledsite2437se	chr8	72551319-72551825	AluYa5	+
filledsite2455se	chr8	91515331-91515781	AluYb8	+
filledsite2461se	chr8	98200058-98200568	AluYb8	+
filledsite2466se	chr8	98783362-98783867	AluYa5	+
filledsite2474se	chr9	10498652-10499154	AluYa5	+
filledsite2478se	chr9	106132449-106132956	AluYa5	+
filledsite2481se	chr9	109024030-109024536	AluYa5	+
filledsite2486se	chr9	115936999-115937468	AluYa5	+
filledsite2489se	chr9	119476977-119477484	AluYa5	+
filledsite2491se	chr9	122531746-122532258	AluYa5	+
filledsite2493se	chr9	124496275-124496780	AluYb8	+
filledsite2500se	chr9	13840687-13841182	AluYc1	+
filledsite2508se	chr9	16389308-16389817	AluYa5	+
filledsite2514se	chr9	18751740-18752214	AluYa5	+
filledsite2528se	chr9	22999402-22999909	AluYe5	+
filledsite2550se	chr9	36423565-36424070	AluY	+
filledsite2568se	chr9	71665240-71665659	AluYc2	+
filledsite2572se	chr9	7371713-7372219	AluYe5	+
filledsite2574se	chr9	74269295-74269768	AluYb8	+
filledsite2576se	chr9	75669431-75669866	AluYa5	+
filledsite2588se	chr9	85033760-85034273	AluYb8	+
filledsite2596se	chr9	9652961-9653474	AluYb8	+

filledsite2628se	chrX	105191063-105191559	AluYa5	+
filledsite2639se	chrX	11096310-11096816	AluYa4	+
filledsite2653se	chrX	11725267-11731524	L1HS	+
filledsite2677se	chrX	135157898-135158407	AluYa5	+
filledsite2708se	chrX	16427596-16428316	L1HS	+
filledsite2710se	chrX	17520531-17521036	AluY	+
filledsite2720se	chrX	24755949-24756462	AluYb9	+
filledsite2725se	chrX	31577378-31577882	AluYg6	+
filledsite2740se	chrX	43669109-43669615	AluY	+
filledsite2756se	chrX	5570251-5570756	AluYa5	+
filledsite2790se	chrX	81096555-81102794	L1HS	+

Supplemental Table S1. List of the identifired AMH-specific RI that are absent in chimps, HN and HD.

For access to other supplemental data, please contact etienne.guichard2@unibo.it