Opposing selective forces operating on human-specific duplicated TCAF genes in Neanderthals and humans

PingHsun Hsieh
University of Washington School of Medicine
Vy Dang
University of Washington School of Medicine
Mitchell Vollger
University of Washington School of Medicine
Yafei Mao
University of Washington
Tzu-Hsueh Huang
University of Washington School of Medicine
Philip Dishuck
Univ of WA https://orcid.org/0000-0003-2223-9787
Carl Baker
University of Washington School of Medicine
Stuart Cantsilieris
University of Washington School of Medicine
Alexandra Lewis
University of Washington School of Medicine
Katherine Munson
University of Washington https://orcid.org/0000-0001-8413-6498
Melanie Scofield
University of British Columbia
AnneMarie Welch
University of Washington School of Medicine
Jason Underwood
Pacific Biosciences (United States)
Evan Eichler (eee@gs.washington.edu)
University of Washington School of Medicine https://orcid.org/0000-0002-8246-4014

Article

Keywords: Human-specific genes, thermal regulatory gene duplications, natural selection, archaic and modern humans

Posted Date: October 1st, 2020

DOI: https://doi.org/10.21203/rs.3.rs-77798/v1

Version of Record: A version of this preprint was published at Nature Communications on August 25th, 2021. See the published version at https://doi.org/10.1038/s41467-021-25435-4.

Opposing selective forces operating on human-specific duplicated *TCAF* genes in Neanderthals and humans

PingHsun Hsieh,[1*] Vy Dang,[1,†] Mitchell R. Vollger,[1] Yafei Mao,[1] Tzu-Hsueh Huang,[1] Philip C. Dishuck,[1] Carl Baker[1], Stuart Cantsilieris,[1,‡] Alexandra P. Lewis,[1] Katherine M. Munson, [1] Melanie Sorensen,[1] AnneMarie E. Welch,[1,ǂ] Jason G. Underwood,[1,2] Evan E. Eichler[1,3*]

*Co-corresponding authors

Correspondence: hsiehph@uw.edu

eee@gs.washington.edu

Affiliations:

1. Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA

2. Pacific Biosciences (PacBio) of California, Incorporated, Menlo Park, CA, USA

3. Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA

† Current address: Institute for Cell and Molecular Biology, University of Texas, Austin, TX, USA

‡ Current address: Centre for Eye Research Australia, Department of Surgery (Ophthalmology), University of Melbourne, Royal Victorian Eye and Ear Hospital, East Melbourne, VIC, Australia

ǂ Current address: Brain and Mitochondrial Research, Murdoch Children's Research Institute, Royal Children's Hospital, Melbourne, VIC, Australia

Keywords: Human-specific genes, thermal regulatory gene duplications, natural selection, archaic and modern humans

SUMMARY

TRP channel-associated factor 1/2 (TCAF1/TCAF2) proteins antagonistically regulate the cold-sensor protein TRPM8 in multiple human tissues. Understanding their significance has been complicated given the locus spans a gap-ridden region with complex segmental duplications in GRCh38. Using long-read sequencing, we sequence-resolve the locus, annotate full-length *TCAF* models in human and nonhuman primate genomes, and show substantial human-specific *TCAF* copy number variation. We identify two human super haplogroups, H4 and H5, and establish that *TCAF* duplications originated ~1.7 million years ago but diversified only in *Homo sapiens* by recurrent structural mutations that altered *TCAF* copy number and regulation. Conversely, in all archaic-hominin samples the fixation for a specific H4 haplotype without duplication is likely due to positive selection. The significant, positive effect of H4 on *TCAF2* expression in modern-day humans with candidate associations for hypothyroidism, nerve compression, and diabetes suggests *TCAF* diversification among hominins potentially in response to cold or dietary adaptations.

INTRODUCTION

Gene duplication contributes significantly to molecular evolution by providing the raw material for genetic novelty and organismal adaptation [1,2]. In the human lineage, recently duplicated regions corresponding to segmental duplications (SDs) are known to give rise to new genes contributing to synaptogenesis, neuronal migration, and neocortical expansion [3-7] that distinguish human from other ape species. Although relatively few in number, some of the largest genetic changes that differentiate humans from archaic hominins involve gene-rich SDs [8,9]. Among human-specific SD genes, the duplications of *TCAF1*/*TCAF2* are particularly intriguing. These genes encode TRP channel-associated factors that bind to the protein TRPM8 (transient receptor potential cation channel subfamily M member 8)—the ion channel acting as thermal sensor in peripheral somatosensory neurons [10] and are thought to be under positive selection in Eurasian populations [11]. Both TCAF1 and TCAF2 proteins interact directly with the TRPM8 channel but have antagonistic effects in TRPM8 gating and trafficking into the plasma membrane: while TCAF1 facilitates the TRPM8 channel opening and migration, the activity of TRPM8 is completely suppressed by TCAF2 [12]. These results suggest that the relative abundance of TCAF proteins and their competition in binding TRPM8 is likely critical in physiological regulation of the channel activity [12].

The *TCAF* family originated from an ancient gene duplication event at the basal of mammalian phylogeny and stayed as single-copy genes throughout much of their evolution [8,12]. Within the human lineage, subsequent duplications over the last few million years have changed the copy number of both *TCAF1* and *TCAF2*. In the human reference genome GRCh38, *TCAF1* and *TCAF2* are embedded and span within a complex region of large, highly identical SDs (>99.5%) consisting of >250 thousand base pairs (kbp) sequence and an annotated gap at chromosome 7q35. Studying these SDs is particularly problematic because of the high sequence identity and missing sequences in reference genomes [13-15]. Using large-insert bacterial artificial chromosome (BAC) clones from a haploid hydatidiform mole cell line, we recently assembled a sequence-resolved haplotype spanning over this region and characterized three sets of SDs present in both direct and indirect orientations. Copy number estimations from read-

depth analyses suggest that *TCAF* duplications are completely missing in archaic hominins, such as Neanderthal and Denisovan, and nonhuman great apes [8]. In addition, while their sequence analysis suggested that a single full-length *TCAF1* and *TCAF2* exist at the locus, respectively, additional *TCAF1*/*TCAF2* copies appear to be truncated or incomplete [8].

In the present study, we systematically explore the haplotype structure of the *TCAF* locus in order to study its diversity, annotate the genes, and infer its evolutionary history in the context of selection. We generate 15 sequence-resolved haplotypes from both human and nonhuman primate BAC libraries as well as full-length non-chimeric (FLNC) transcript data from seven tissues using the Pacific Biosciences (PacBio) single-molecule, real-time (SMRT) long-read sequencing technology. We leverage a large collection of over 1,100 publicly available high-coverage Illumina genomes from modern, and archaic humans as well as nonhuman primate samples and document the global diversity of *TCAF* SDs. Finally, we integrate both phylogenetic and population genetic inference approaches to reconstruct the evolution of the *TCAF* SDs in primates and identify putative signals of balancing selection in Native American, Melanesian, and Siberian populations as well as positive selection in both Neanderthal and Denisovan. We provide evidence for differential expressed quantitative trait loci (eQTLs) between two *TCAF* super haplogroups in thyroid, whole blood, tibial nerve, spleen, and lung tissues. In addition, our association analysis using the UK Biobank data suggest that the *TCAF* super haplogroups are likely associated with hypothyroidism, diabetes, and nerve-related traits in some human populations. The study provides one of the most complete genetic investigations of human-specific SDs shedding potential new insights into structural adaptations important in thermal regulation.

RESULTS

Diversity of TCAF copy number in human and nonhuman primates. We applied a read-depth-based genotyper [16] to reevaluate copy number diversity of *TCAF* SDs (duplication segments A, B, and C; **Figure 1A**) in a collection of recently published diverse human and nonhuman samples [17-23]. The set includes high-coverage Illumina genomes from a global panel of 1,102 human samples, four archaic

hominins, and 71 nonhuman great apes. Consistent with previous studies [8], all archaic hominin and nonhuman primate samples are fixed for diploid copy number (CN) 2 (**Figure 1B**). Among contemporary modern human samples, the overall diploid copy number estimates of the three *TCAF* SDs range from two to eight copies and are highly correlated among each other (**Figure S1**), indicating that these three SDs likely appear as a cassette in the evolution of this locus. In contrast to archaic humans where there is a prediction of a single copy of this locus, we estimate that ~98% of modern humans carry more than two copies (**Figures 1B, S1**), suggesting that most of the diversity emerged during the divergence of hominins.

Among humans, we observed a wide range of variation across different geographic locations in contemporary human populations (**Figure 1B-C**). While the median copy numbers are consistently CN4 (with the exception of Native Americans [median = 3]), African samples on average carry more copies (mean CN = 4.6) than other populations (**Figure 1B**). We tested for copy number stratification by applying both the V_{ST} statistic [24] and CN differentiation test [25]. While there is little evidence for population differentiation among pairs of super populations, a reduction in diploid CN among Native American samples distinguishes this group from all others (V_{ST} > 0.21, Bonferroni p value for the CN differentiation test < 0.0014; **Table S1**).

The observation that all diploid CN2 samples (i.e., without duplications) are present solely among non-African populations raises the question as to whether it is due to archaic introgression. To address this, we compared single-nucleotide variation at 20 kbp unique diploid sequences flanking the *TCAF* SD locus (chr7:143,501,000-143,521,000, chr7:143,875,000-143,895,000 [GRCh38]) between archaic and modern human genomes. In general, the archaic flanking haplotypes are similar to those present in modern human samples, including the AFR samples (**Figure S2**). Focusing on sites that are fixed in derived alleles in either the Neanderthal or Denisovan samples, the numbers of derived allele counts in the CN2 samples are comparable to those drawn randomly from African samples (p = 0.276; 1,000 non-parametric bootstrap samples). These results suggest that non-African CN2 haplotypes unlikely arose as a result of introgression from Neanderthal or Denisovan.

Characterization of TCAF haplotype diversity using long-read BAC-based assemblies. These near-perfect duplications have hampered not only genome assembly but also the characterization of underlying genetic diversity as well as our understanding of the *TCAF* evolution. An investigation of four recently published long-read whole-genome assemblies, for example, showed that all four failed to create contiguous sequences for both haplotypes spanning over the *TCAF* SDs (**Figure S3, Table S2**). To resolve this, we selected large-insert BAC clones and used PacBio long-read sequencing to generate 15 high-quality sequence-resolved haplotypes from eight human samples and three nonhuman primates, including chimpanzee, gorilla, as well as an Old World rhesus macaque monkey (**Table 1**, **Methods**).

A sequence comparison analysis among these newly generated haplotypes confirms the complex organization of these *TCAF* SDs [8] and further reveals considerable structural diversity in humans when compared to nonhuman primates (**Figure 2 and S4-S12**). While all the nonhuman primate haplotypes carry a single copy for the *TCAF* SD cassette and single full-length *TCAF1* and *TCAF2* genes (**Figures S4, S11-S12**), among the 12 human haplotypes, we identify five distinct haplogroups that carry one to three copies for the SD cassettes, which range from 145–406 kbp in length (**Figure 2, Table 1**). The two largest human haplogroups (Haplogroups 4 and 5) both carry three copies of the SD cassette (**Figures 3, S5**) but differ by a 100 kbp inversion (Haplogroup5:337,305-437,635; **Figure S5**). The Haplogroup 5 organization is mostly consistent with the current human reference genome (GRCh38), which has a gap in the middle of expanded complex SDs in the reference (**Figure 3**). The newly generated Haplogroup 5 sequence, thus, eliminates the gap by adding 103,616 bp (Haplogroup5:216,541-320,157) in the human reference.

Most of the sequence-resolved haplotypes (8 out of 15) consist of two copies of the *TCAF* SD cassette (Haplogroups 2 and 3; **Figure 2, S6-S9**). These two haplogroups can be structurally distinguished by a similar 100 kbp inversion as identified between Haplogroups 4 and 5. We, therefore, use this inversion polymorphism to classify two super haplogroups (H4 and H5, **Figure 2**). The shortest human BAC haplotypes (Haplogroup 1; **Figures 2 and S10**) that we assembled consist of a single copy of the

SD cassette and are similar in structure to those found in nonhuman primates. To further refine the relationship among these assembled haplotypes, we performed a series of sequence alignment analyses among different pairs of haplotypes (**Figure 2**). Based on the organization of the *TCAF* SDs, we conservatively determined that Haplogroups 2 and 5 are closely related and so are Haplogroups 3 and 4. Among Haplogroup 2, two subgroups, Haplogroup 2-1 (n = 1) and Haplogroup 2-2 (n = 5), were further separated and likely emerged due to two independent events, where each involves a different 130 kbp deletion, with respect to Haplogroup 5 (**Figure 2, Table S3**). Similarly, Haplogroup 3 can also be further classified into two subgroups, Haplogroup 3-1 and Haplogroup 3-2, because we predict that they likely derived from two different large deletions with respect to Haplogroup 4 (**Figure 2, Table S3**). Finally, while Haplogroup1 has a similar structure to the nonhuman primate haplotypes, we showed that this group is more closely related to Haplogroup 3 and likely a result of a deletion event over Haplogroup 3 (**Figure S10, Table S3**).

The breakpoints of these six large structural differences map within nearly identical (>99.5%) SDs, consistent with the action of non-allelic homologous recombination (NAHR) (**Table S3**). By identifying the longest, high-identity sequences around the putative breakpoints (**Table S3**), we show that four associate with DupA SDs, while two map around DupC SDs. For example, the inferred inversion breakpoints of Haplogroup 4 are immediately flanking two inversely oriented, nearly identical sequences within DupC SDs (~39.6 kbp; sequence identity > 99.8%). In addition, our analysis of sequence composition reveals the presence of nearly identical LINE/SINE elements in sequences flanking the putative breakpoints (**Table S3**). The inferred breakpoints for the deletion event between Haplogroups 3-1 and 4 overlap with two perfectly identical (100%) 2.5 kbp *CTAGE* (Cutaneous T Cell Lymphoma-Associated Antigen 1) family sequences (**Figure S8, Table S3**). Remarkably, most of the deletion events are similar in size ranging in length from 129.7 to 134.4 kbp. Thus, our data highlight a rapid evolutionary process that involves large-scale duplication, deletion, and inversion events via NAHR but where deletion events are constrained.

Discovery of new TCAF genes/isoforms. The newly generated long-read BAC haplotypes reported allow us to revisit *TCAF* annotation. Functional studies have shown that *TCAF1*/*TCAF2* are highly expressed in the prostate, brain, esophagus, prostate, and skin tissues [12] (GTEx Portal, https://gtexportal.org). To this end, we targeted capture of FLNC *TCAF* transcripts and generated 480,700 FLNC transcripts from six human tissues, including dorsal root ganglion, esophagus, fibroblast, skin, fetal brain, and testis, as well as 50,885 FLNC transcripts from a chimpanzee lymphoblast cell line (**Table S4**). *TCAF* transcript on-target rates range from the lowest in the testis sample (2.2%) to the highest for the chimpanzee lymphoblast sample (15.5%). To identify *TCAF* models, we aligned high-quality FLNC transcripts to the assembled haplogroup sequences and only retained transcripts that have complete open reading frames with >100 amino acids (aa) in length (**Methods**). Of note, we followed the previously described nomenclature[8] to delineate *TCAF* models and isoforms. Overall, our analysis indicates the expression of *TCAF* genes in a wide variety of tissues and reveals new *TCAF* paralogs as well as a variety of new isoforms (**Figures 3, S5-S13; Table S5**). In particular, we recovered only single isoforms for individual genes in the *TCAF1* subfamily. In contrast, the *TCAF2* subfamily shows much more diversity with 10, 2, and 4 isoforms being identified for *TCAF2A*, *TCAF2C1*, and *TCAF2C2*, respectively. With respect to the current annotation in the human reference genome (GRCh38), all haplotypes confirm that *TCAF2A* (or *TCAF2* in RefSeq Release 200) is incorrectly annotated due to the presence of a gap at this locus (**Figure 3**). The gene models and isoforms reported here represent the most correct and comprehensive annotations for the *TCAF* families (**Table S5**).

Compared to the chimpanzee, every additional copy of an SD cassette identified among human haplotypes is associated with an additional copy of *TCAF1A* and *TCAF2C* paralogs, although they are incomplete copies to *TCAF1B* and *TCAF2A*, respectively (**Figures 3, S4-S12**). Specifically, all *TCAF1A1/TCAF1A2* paralogs match to the last seven exons of *TCAF1B*, while the *TCAF2C1/TCAF2C2* paralogs consist of the first three exons of *TCAF2A* (**Figure 3**). In addition, the predicted breakpoints between Haplogroup 2-2 and Haplogroup 5 coincide with the *TCAF1A1* and *TCAF1A2* paralogs. The sequence alignment analysis shows that the breakpoint at Haplogroup 2-2 spans from the third to the fifth

introns, and thus, the *TCAF1A1* copy of Haplogroup 2-2 is a fusion version of the *TCAF1A1* and *TCAF1A2* copies of Haplogroup 5 (**Figure S8**). Because the amino acid sequences encoded from these three paralogs are 100% identical, this suggests the actual breakpoints are within one of the three introns. Finally, the 100 kbp inversion between Haplogroup 4 and Haplogroup 5 raises the question whether the large-scale structural variation event has any effect on *TCAF* coding sequences. Notably, the inferred inversion breakpoints map to the second introns of *TCAF2A* and *TCAF2C2*. There are only two differences in the *TCAF2A* exon alignment between Haplogroups 4 and 5: one is a synonymous change in the second exon, and the other causes a nonsynonymous change (R479Q) in the third exon, which is beyond the inferred breakpoints (**Figure S13**). Thus, our results predict that this inversion shuffles the last six exons of *TCAF2A*, which were ligated with the first two exons of different *TCAF2A*/*TCAF2C2* paralogs on the two haplotypes without disrupting or altering the coding sequence potential of *TCAF2A*/*TCAF2C2*.

It should be noted that although the structural changes largely increase and decrease dosage of individual *TCAF* family members with relatively few predicted amino-acid differences, there is evidence that structural mutations and conversion events are likely affecting the regulatory landscape. We searched for candidate sites of interlocus gene conversion (IGC) among paralogs (**Figure S14**) and observed an IGC site that overlaps with a strong H3K27Ac signal (left white box, **Figure 1A**) an epigenetic signature for enhancers. Using paralog-specific copy number genotypes (**Methods**), we found a reduction in copy number at this locus (i.e., the acceptor site; chr7:143,615,002−143,624,482, GRCh38; the left white box, **Figures 1A and S15**), which, interestingly, is significantly negatively correlated with an increase in copy number at its paralogous locus (the donor site; chr7:143,833,163−143,842,658; the right white box, **Figures 1A and S15**) (Pearson's correlation = -0.3, p value = 5×10^{-16}) in multiple continental populations (**Figure S16**). This result is consistent with the hypothesis of an IGC event copying the donor sequence over the acceptor locus, resulting in reciprocal copy number changes at these two loci. Interestingly, we find that this paralog-specific copy number at the donor locus is negatively correlated (R = -0.18, p value = 3.6×10^{-7}) with the (absolute) latitudinal location of human populations while the relationship for

the acceptor site is positive (R = 0.23, p value = 2.5×10^{-11}; **Figure S17**). While the observed correlations between latitudes of individual populations and paralog-specific copy numbers at the IGC sites may be the result of natural selection, such strong correlations are relatively common among similar loci across the genome (p value = 0.22; 1,000 random SD loci of genomic shuffling), suggesting demographic history among populations also likely contributing to this observation.

Evolution of TCAF segmental duplications. To reconstruct evolutionary origin of the *TCAF* haplotypes in modern humans, we leveraged these high-quality human and nonhuman primate haplotypes for phylogenetic reconstruction (**Tables 1 and S6**). The mean sequence identities for DupA, DupB, and DupC SDs are high among human haplogroups (99.80% [s.d.: 0.04], 99.71% [s.d.: 0.10], and 99.85% [s.d.: 0.05], respectively) (**Figure S18**), while the sequence divergence between human and nonhuman primate *TCAF* haplotypes are compatible to published whole-genome estimates (**Figure S18**) [26,27]. Together, these findings are consistent with a recent *TCAF* SD expansion or IGC. We considered each of the duplicated segments independently guarding against confounders such as IGC and non-allelic homologous recombination events. Because several SDs represent hybrids of two SD paralogs due to NAHR fusion events (**Figure 2**), for example, we restricted our analysis to the larger segments of hybrid SDs. To minimize the effects of IGC (**Figure S14**) [8], we applied GENECONV (v1.81a) to exclude sites corresponding to 80.5%, 61.7%, and 70.6% of DupA, DupB, and DupC sequences, respectively (**Table S7**). Phylogenetic reconstruction and dating estimates on this filtered set showed that all human copies form a single clade with a 100% posterior support and coalesce at 1.06 (95% C.I.:0.87-1.26), 1.72 (95% C.I.: 1.26-2.21), and 1.47 (95% C.I.: 0.74-2.54) million years ago (Mya), for DupA, DupB, and DupC SDs, respectively (**Figures 4A and S19-S21**). While a few differences in topology were noted among the inferred phylogenies (**Figures S19-S21**), these estimates, with the exception of DupB, overlapped coalescent estimates obtained from a 12.3 kbp single-copy unique region (**Figure 2, the right panel**), which suggests that the two supergroups: H5 (Haplogroup 5, Haplogroup 2-1, and Haplogroup 2-2) and H4 (Haplogroup 4, Haplogroup 3-1, Haplogroup 3-2, and Haplogroup 1) diverged from each other ~0.73

247 Mya (kya; 95% C.I.: 0.46-1.03) (**Figure S22A**). Importantly, these findings suggest that the duplications
248 began to occur before humans and archaic hominins diverged.
249 To further refine our inferences on the evolution of *TCAF*, we realigned genome sequence data
250 from four archaic hominins along with human and chimpanzee samples to a custom, gapless human
251 reference chromosome 7 using GRCh38 and sequence-resolved BAC haplotypes to generate a joint
252 single-nucleotide variant call set (**Methods**). Due to the limitations of short-read data, we restricted this
253 analysis to three unique regions (52.3 kbp in total) flanking and internal to the *TCAF* locus (**Figures 1**
254 **and 2**). We identified 1,275 single-nucleotide variants (SNVs) (QV>20) and used these to
255 computationally phase and to construct haplotypes from 802 samples (**Figures S2 and S23**). Using these
256 archaic hominin haplotypes with the sequences of the seven BAC haplogroups, we estimate the time to
257 the most recent common ancestor (TMRCA) of all hominins being 0.78 Mya (95% C.I.:0.55-1.03 Mya),
258 which is consistent with that reported above. Interestingly, all archaic haplotypes and the supergroup H4
259 are more closely related than either is to the supergroup H5, and the clade of archaic and supergroup H4
260 haplotypes coalesces ~0.53 Mya (**Figure 4B**). Thus, our results show that as early as 1.72 Mya the *TCAF*
261 locus (DupB SD) began to duplicate and that by 0.73 Mya most of the of the *TCAF* haplotype diversity
262 that we observe in modern-day humans had already emerged. Archaic hominins, in contrast, show a
263 reduction in genetic diversity without the associated copy number duplication diversity found in modern
264 humans (**Figure 4C**).
265
266 ***Hominin diversity and natural selection of TCAF haplotypes*.** We used the combined haplotypes
267 constructed from the 53.2 kbp of unique regions to further explore genetic diversity among archaic and
268 present-day hominins. A haplotype-based principal component analysis (PCA) reveals distinct clusters
269 (**Figures S24-S25**) corresponding in part to the two human haplogroups H4 and H5 identified from the
270 BAC sequence analysis (**Figure S25**). Using a supervised clustering method based on haplotype PCA and
271 machine-learning algorithm t-SNE [28] for dimensionality reduction and visualization (**Methods**), we show
272 that contemporary hominin haplotypes can be classified into 12 different clusters (**Figures 5A and S26**)

with an estimated haplotype heterozygosity of 0.885 (**Methods**), which is much higher than the reported genome-wide estimated mean of 0.534 [29]. Phylogenetic reconstruction using 10 haplotypes drawn at random from each cluster reproducibly reconstructs a maximum likelihood tree closely resembling this cluster-based relationship (**Figure 5B**). Considering the two flanking and unique regions independently also confirms a phylogeny of 12 distinct clusters with different BAC haplogroup representing different clusters (**Figures S26-S27**). Among humans, no single population is restricted to a specific cluster, although some clusters appear enriched. For example, clusters 2 and 7 largely consist of African (60.3%) and South Asian (50.6%) haplotypes, respectively (**Figure S28**). Assuming an equal split of the overall diploid CN estimates between the two haplotypes in an individual, we found significant differences in *TCAF* copy number among haplotype clusters (Bonferroni corrected p values < 0.05; Mann-Whitney *U* tests) (**Figure S29**).

In sharp contrast to the extensive haplotype diversity in modern human samples, the eight archaic hominin haplotypes are virtually identical belonging to a single homogenous group (**Figures 5, S24-S29**). Under a panmictic model and assuming an ancestral hominin origin of ~0.73 Mya for *TCAF* haplotypes, it is highly unlikely that all eight archaic haplotypes would form a single cluster given modern-day human diversity (p value = 2.1×10^{-7}, 100×10^{6} permutations). To formally test whether natural selection plays a role in these unusual patterns in both modern and archaic humans, we performed a selection test using the Tajima's *D* statistic and coalescent simulations under plausible demographic models to assess significance (**Methods**). The results suggest significant distinct signals of selection in humans and archaic hominins separately (**Figure 5C**). Among modern human populations, Native American, Melanesian, and Siberian populations show evidence for excess of heterozygosity and significantly larger Tajima's *D*s compared to those from coalescent simulations under neutral demographic models (**Figure 5C**); though the distributions for Melanesians and Siberians center on zero. Especially among Native Americans (Tajima's $D = 1.24$, p value = 4.3×10^{-9}), these results are consistent with the action of balancing selection at this locus (**Figure 5C**). In contrast, the Tajima's *D* distribution of the archaic hominin samples are strictly negative and significantly lower than the neutral expectation, and this result holds even when the

299 Denisovan sample was removed (Tajima's D = -1.05; p value < 1×10^{-6}) (**Figure 5C**). Our results suggest
300 that the observed archaic hominin haplotype variation pattern is unlikely explained by their demographic
301 history, thus, favoring a hypothesis of positive selection.

302 These dramatic differences in diversity and potentially distinct selection forces pose the question
303 of whether there are any functional differences among the different *TCAF* haplogroups. Because our
304 structural analyses suggested dosage and regulatory differences among common haplogroups (**Figures**
305 **S28-S29**), we investigated *TCAF2* expression using the GTEx multi-tissue data (release v8) and possible
306 genetic associations between traits and *TCAF* haplotypes using the UK Biobank data. Based on the BAC
307 references, we identified four tagging SNVs in the unique diploid region embedded among the *TCAF* SDs
308 that could distinguish super haplogroups H4 and H5 (**Figure 6**). Because the two super haplogroups
309 structurally differ by the 100 kbp inversion, we expect a reduction of homologous recombination events
310 between the two groups and therefore an increase in linkage disequilibrium (LD) in hominins. As
311 expected, we observe substantial LD across all three unique diploid regions in all archaic and modern
312 samples and confirm nearly complete LD among the four tagging SNVs (D' > 0.98, R^2 > 0.89) (**Figure**
313 **S30**). Using these data, we again found that all archaic samples are fixed for the H4 supergroup and
314 inferred the frequencies of H4, H5, and other forms in modern humans being 57.1%, 40.3%, and 2.6%,
315 respectively. While the H4 haplogroup is found at high frequency among non-Africans, especially in
316 Melanesians (72%), some Native Americans (e.g., the Karitiana, >72%), and Northeast Asians (e.g., the
317 Mongolians, >68%), the H5 group segregates in higher frequency among Europeans (e.g., the Basques,
318 >81%), Middle Easterners (e.g., the Druze, >61%), and some Native Americans (e.g., the Surui, >83%)
319 (**Figure 6A**). Consistent with our haplotype-based PCA clustering results, we observed high frequencies
320 of the H4/H5 heterozygous form (orange in **Figure 6B**) across populations, with the highest found in
321 Southeast Asians and Native Americans (>60%) and the lowest in some Melanesian and African
322 populations (<20%).

323 Given the common presence of these two super haplogroups in humans (**Figure 6A-6B**), we
324 collected multi-tissue eQTL data from the GTEx Project (release v8) and found that all four tagging

SNVs are *TCAF2* eQTLs (**Figure 6C**). The H4 haplogroup (reference) alleles are significantly associated with increasing *TCAF2* (p values < 1×10^{-10}) expression in thyroid and tibial nerve tissues in addition to others (**Figure 6C**). Thus, our inferences of the differential *TCAF2* expression between the H4 and H5 haplogroups imply possibly functional differences in the peripheral nervous system and thyroid among hominin groups. Finally, we explored traits that are potentially associated with the *TCAF* haplogroups using the recently released genetic association data from the ancestry-based genetic analysis of the UK Biobank (Pan-UK Biobank as of 2020.08.06; https://pan.ukbb.broadinstitute.org/). We examined SNVs from the three unique diploid regions at the *TCAF* locus and focused on nerve- and thyroid-related traits, including diabetes as thyroid dysfunctions often result in diabetes mellitus[30]. While in general no SNVs reach genome-wide significance for traits that we interrogated, we found several suggestive association signals among specific human populations. For example, we found nominal associations (uncorrected p values < 0.05) for SNVS, including the four eQTLs, for hypothyroidism (phecode-244) in both African and East Asian populations as well as for diabetes medication (Glimepiride) and compressed nerves in samples of European descent (**Figure S31**) consistent with the expression pattern of these genes.

DISCUSSION

Genetic changes in SDs are now well recognized as an important source of functional innovation in human evolution due, in part, to subsequent rounds of duplication, deletion, or inversion [13,14]. Such regions, however, are often frequently overlooked as part of comparative studies, association studies, or scans of selection because of their structural complexity and high-degree of homology. In this study, we first delineate the complex structural diversity of the locus in humans and compare it to nonhuman primates at the sequence level in order to reconstruct its evolutionary history [12]. Of note, complete assembly of 15 haplotypes by tedious BAC-based sequencing was sufficient to capture more than half of the genetic diversity of *TCAF* SDs in humans. Based on these data, we develop a model of rapid diversification with respect to both haplotype and *TCAF* SD copy number diversity (**Figures 1-3**). It is worth noting that despite the recent advances in long-read sequence assembly, the most complex

haplotypes still remain collapsed and unassembled in some of the recently released human genomes [31-33] (**Figure S4, Table S2**). We believe that this is due partly to the read-length limitation of the data as well as the near-perfect sequence identity among *TCAF* SDs that lead to ambiguous read partitions during haplotype assembly. This highlights the importance of alternate approaches, such as large-insert BAC clones, to construct highly accurate haplotype-resolved assemblies. More sophisticated assembly approaches and long-read technology for generating longer and even more accurate sequence reads will likely make such investigations much more routine in the future [34].

The structural complexity and copy number variation of this region underscores its mutational dynamics. The high degree of sequence identity among *TCAF* paralogs (>99.7%), their clustered and tandem orientation, and their large size (10–60 kbp) promote recurrent structural rearrangements making the entire locus genetically unstable as a result of NAHR [13]. Human Haplogroups 4 and 5, in particular, carry more *TCAF* SD paralogs than others, and, thus, both are predicted to be subject to a higher likelihood of structural mutation. Our analysis predicts that H4 would be more predisposed to structural rearrangements than H5 due to a greater number of directly orientated duplications (DupA and DupB) (**Figure 2**). This prediction is consistent with the observation of more subgroups within supergroup H4 in the haplotype-based PCA (**Figures S24-S27**). Surprisingly, this predisposition to recurrent structural diversity is specific to modern-day humans as our sampling of eight archaic hominin haplotypes suggests a static H4 haplogroup without a single duplication event. Given our estimates of origin of the duplications (~1.7 Mya) and haplogroup coalescence (~0.73 Mya), the lack of diversity in the Neanderthal and Denisovan lineages is highly unlikely.

One of the challenges studying this region is the wide-spread IGC we identified among the *TCAF* SDs. IGC essentially erases signatures left by previous mutation events in the acceptor site, thus biasing towards younger TMRCA estimates [35,36]. While we mitigated such a bias by focusing our analysis on non-IGC sequences, this reduced the available dataset limiting our power to infer its evolution at a finer level of resolution. Interestingly, the analysis uncovered a significant correlation between paralog-specific copy number and the latitudinal location of human populations precisely over a potential enhancer region,

suggesting that IGC may have contributed to regulatory differences of *TCAF* genes among human populations. Such co-option of IGC has been noted before among duplicate genes. For example, a human-specific, pseudogene-mediated IGC of *SIGLEC11* replaced the 5' upstream region and coding exons of *SIGLEC11* by its adjacent pseudogene *SIGLEC16P* in the human lineage. This event has been hypothesized as adaptive during human brain development [37]. While we cannot rule out the possibility of confounding effects of demography resulting in the observed correlation, given the importance of *TCAF* in thermal regulation and sensing, it is tempting to hypothesize that IGC contributes to local adaptation by changing dosages through rapid replacement of regulatory sequences in addition to SD.

Our more complete reference genomes coupled with targeted cDNA sequencing also provides the most comprehensive *TCAF* annotation to date for humans and nonhuman primates (**Figures 3 and S4-S12**). We show that each additional *TCAF* SD cassette (DupA, DupB, and DupC) generates additional copies of *TCAF1A* and *TCAF2C* through incomplete gene duplication. Gene duplication has been known to be an important source of genetic novelty and adaptive evolution [1,38], especially for the great ape lineages given the surge of duplications predicted to have occurred in the hominid ancestor [39]. Throughout the evolution of mammals, *TCAF* genes have been highly conserved, and human is the only lineage that is known to carry additional copies. The coding sequences among the individual paralogs are highly conserved as a result of their recent origin arguing for the relative importance of *TCAF* dosage differences in regulating the TRPM8 ion channel as suggested by a previous studies [12]. We also observe, however, considerable isoform diversity as a result of the incomplete nature of the SD particularly among *TCAF2* paralogs suggesting a potentially more complex relationship with the ancestral genes. This pattern of incomplete duplication resembles the evolution of the neural human-specific duplication *SRGAP2* genes [5], where the younger, incomplete duplicated paralog *SRGAP2C* inhibits the function of the ancestral copy *SRGAP2A*, and in mice experiments this in turn prolongs the maturation of synapses and subsequently increases the density of the length of synapses [4]. Taken together, our results bolster the hypothesis of recent incomplete gene duplications and adaptive human evolution.

Remarkably, our data support a model of two distinct forces of natural selection operating on the same locus over the last half million years of hominin evolution. We propose that diversifying or balancing selection is acting in at least some human populations, particularly out-of-African populations such as Native Americans, to maintain and expand haplotype and structural diversity. While the exact mechanism and phenotype that balancing selection are acting upon are not clear, the differential eQTL signal related to *TCAF2* expression between supergroup H4 and H5 alleles in multiple tissues would be consistent with antagonistic pleiotropy, where two haplotypes have opposing effects on two different traits depending on either environmental or life-history conditions [40,41]. In contrast, Neanderthal and Denisovan show a paucity of genetic variation, and while the sample size is still limited, this observation is unlikely to change with the sequencing of additional archaic genomes. Moreover, both the significantly negative Tajima's *D* statistic and the single-haplotype without duplication are highly unlikely observed under neutrality after accounting for differences explained by demographic history. Our inference results, thus, strongly argue that positive selection has reduced genetic diversity at the *TCAF* locus in these archaic hominin lineages.

The opposing effects on the *TCAF2* expression found between H4 and H5 haplogroups in the thyroid is particularly intriguing. Thyroid activity is one of the most important determinants of overall energy expenditure and basal metabolic rate in human body [42]. Not only does the thyroid hormone maintain core body temperature when exposed to cold, but it also regulates metabolism through action in the brain, brown fat, skeletal muscle, etc. [43]. Thyroid dysfunctions, both hyperthyroidism and hypothyroidism, often lead to metabolic disorders, such as diabetes mellitus[30], and also cause neurological disorders during fetal and brain development due to low levels of maternal thyroid hormones from maternal hypothyroidism or lack of iodine in diet [44,45]. Despite no genome-wide significant genetic associations, we did identify suggestive trends between specifically those eQTL variants and hypothyroidism, diabetes medication, as well as pinched nerves. Interestingly, the TRPM8 ion channel, in addition to its role in detecting environmental cold, is also a known regulator of insulin homeostasis through neuronal control of liver insulin clearance and, thus, required for precise thermoregulatory

responses to cold and fasting [46,47]. In addition, a rare variant within the *TCAF* locus (chr7:143735931) was recently reported to be associated with type 2 diabetes in sub-Saharan African populations [48]. Given this variant's restricted geographic distribution and low frequency in the GTEx data (allele frequency: ~0.4%; GTEx portal as of 2020.07.21), it explains neither the observed patterns of diversity nor the selection and the eQTL signals observed here. Therefore, if the *TCAF* haplotype and structural diversity affect an individual's ability to regulate TRPM8 ion-channel gating and, thus, body temperature, it is possible that the antagonistic forces of selection promote adaptations to a cold environment or an enhanced ability to dynamically adapt to changing conditions, such as diets.

While it is still controversial, cold weather is thought to have forced Neanderthals to move from coastal areas to further inland, where they would have had to adapt to iodine-poor diets, which may in turn have affected brain development in this species [44,49]. Iodine deficiency is one of the common causes of hypothyroidism (reduced thyroid hormone levels), which is characterized by lower core body temperature and decreased resting energy expenditure [30]. Despite the limited understanding of the molecular bases underlying the interactions between *TCAF* and thyroid hormones, one possibility may be that selection on H4 increased *TCAF2* expression and subsequently suppressed the activity of TRPM8 in the thyroid of archaic hominins. This may have improved physiologically adaptation to cold weather and low iodine-intake diets.

METHODS

Copy number genotyping in high-coverage Illumina short-read data

To explore the temporal and spatial copy number variation of *TCAF* SDs, we leveraged a large collection of publicly available high-coverage Illumina genomes from 1,102 modern human samples, four archaic hominins, and 71 nonhuman great apes [17-23]. In short, sequencing reads were divided into multiples of 36-mer and then mapped to a repeat-masked human reference genome (GRCh38) using mrsFAST (v3.4.1) [50]. We allowed up to two mismatches per alignment to increase our mapping sensitivity and corrected for possible biases in read depth due to different levels of GC content in sequencing reads. Finally, the copy

number estimate was computed by summarizing over all mappable bases at the locus of interest for each sample.

Library processing and assembly for *TCAF* BAC haplotypes

We obtained the human BAC libraries used in this study from the Virginia Mason Research Center and the nonhuman primate ones from the Children's Hospital Oakland Research Institute. Probes for regions of interest were designed, radioactively labeled and hybridized to the Performa filters, washed, exposed to Phosphor screens, and scanned on a Typhoon scanner. Positives are called and corresponding clones selected from the BAC library. We prepared barcoded libraries from clone DNA using Illumina-compatible Nextera DNA sample prep kits (Epicentre, catalog number GA09115) as described previously [51] and carried out paired-end sequencing (125 bp reads) on an Illumina HiSeq 2500. Reads were then mapped to the reference genome, GRCh38, to identify singly unique nucleotide k-mers (SUNKs) [16]. Non-overlapping BACs were pooled and sheared as described previously [52]. Libraries were processed using the PacBio SMRTbell Template Prep kit following the protocol "Procedure and Checklist—20 kb Template Preparation Using BluePippin Size-Selection System" with the addition of barcoded adaptors during ligation. Up to ten barcoded libraries were then pooled at equimolar amounts and size-selected as a pool on the Sage PippinHT with a start value of 10,000–12,000 and an end value of 50,000. The resulting library was then sequenced on one Sequel SMRT cell 1M by diffusion using Sequel v3.0 chemistry. We performed *de novo* assembly of pooled BAC inserts using Canu (v1.5). Reads were masked for vector sequence (pCC1BAC) and assembled with Canu, then subjected to consensus sequence calling with Arrow (https://github.com/PacificBiosciences/GenomicConsensus). We reviewed PacBio assemblies for misassembly by visualizing the read depth of PacBio reads in Parasight (v7.6, http://baileylab.brown.edu/parasight/download.html), using coverage summaries generated during the resequencing protocol. We performed BLAST (v2.10.1) pairwise alignment between the BAC sequences for the construction of haplotypes. Complete contigs for 15 *TCAF* haplotypes must satisfy the following

two conditions: (1) overlapped DNA segment between BAC sequences must be at least 99.99% identical with no tolerated mismatch and (2) contig must anchor to unique location on both ends when mapping back to the human genome reference GRCh38.

Identification of haplogroups and interlocus gene conversion (IGC) tracts

All sequence alignments in this study were performed using MAFFT (v7.453). To determine the structural differences among the haplogroups, we built single-base level sequence identity profiles using pairwise sequence alignment, followed by smoothing over windows of 500 bp with a step size of 100 bp. Because we are interested in recent, large IGC events, empirical IGC tracts are determined as more than two consecutive windows that have 100% sequence identity. To find additional support, we use a model-based program, GENECONV (v1.81a), to identify pairs of uninterrupted sequences with 100% sequence identity that are longer than expected given the overall pattern of variable sites in an alignment. To be conservative, we considered IGC tracts that have both simulation- and Karlin-Altschul- p values < 0.05 reported by GENECONV.

Full-length non-chimeric (FLNC) transcripts for *TCAF* genes using long-read cDNA sequencing

Total RNA was harvested from six human tissues (dorsal root ganglion: Clontech Laboratories, Inc., Takara [catalog #636150]; esophagus: Clontech Laboratories, Inc., Takara [catalog #636178]; fibroblast (Coriell); skin: Biochain Institute Inc. [catalog #R1234218-P], fetal brain: NCBI BioSample: SAMN09459150; testis: Clontech Laboratories, Inc., Takara [catalog #636533]) and one chimpanzee lymphoblast cell line. We purified polyA RNA using oligo-dT magnetic beads (Dynal: Thermo Fisher Scientific Inc.). Double-stranded cDNA with 96 random barcodes was prepared, amplified and subjected to hybridization capture followed the protocols detailed in (Dougherty et al. 2018). Hybridization probes were designed to target the duplicated exons of *TCAF2*, and sequences are given in Supplemental Table S2b of Dougherty et al., 2018.

Following post-capture PCR, the amplified dsDNA was purified on magnetic beads (AMPure PB; PacBio) and then subjected to library preparation for long-read sequencing (SMRTbell Template Prep Kit 1.0; PacBio with barcoded SMRTbell adapters). SMRT sequencing was performed on the Sequel platform with Sequel v3.0 chemistry (PacBio) with LR SMRT Cells with 2-hour pre-extension and 20-hour movies. Reads corresponding to each sample were extracted by their SMRTbell barcodes and circular consensus sequences were generated from the raw subreads using SMRT Link with minimum number of pass set to 1. The program lima in the Iso-Seq3 pipeline (https://github.com/PacificBiosciences/IsoSeq3) was applied to remove the 5' and 3' dual barcodes and also obtain the unclustered FLNC reads. Parameters used were: lima --isoseq --dump-clips. We did not cluster the FLNC reads further because highly identical paralogous transcripts could undesirably cluster together in this step. Due to the variability of the yields of FLNC transcripts across samples and loci, we used data from various combinations of samples in subsequent analyses when applicable.

To determine the coding potential for each FLNC transcript, we use the program ANGEL (https://github.com/PacificBiosciences/ANGEL) to call open reading frames with a minimum length cutoff of 200 amino acids and proper stop codons. All transcripts with open reading frames were mapped to the seven *TCAF* haplogroups (**Tables S4 and S5**) using minimap2 (v2.17-r941) and individual best placements to specific haplogroups were determined using sequence identity. Given the recent history of the *TCAF* duplications, we required a >99% sequence identity for an alignment to be considered a match between an FLNC read and the haplogroup contig and most have at least an overlap of 200 bases. To improve the sensitivity of FLNC read assignments, we allowed multi-mapping placements for an FLNC read as long as the difference in sequence identity between alignments is < 0.01%. Finally, to determine gene models and isoforms in each haplogroup, we focused our analysis on FLNC transcripts with more than 10 reads (**Table S5**).

Single-nucleotide variant (SNV) calling for unique diploid sequences at *TCAF* locus

To perform population genetic analyses, we generated a joint call set for 840 published high-coverage short-read Illumina genomes, including 828 Human Genome Diversity Project (HGDP) samples, eight chimpanzee samples, and the four high-coverage archaic genomes. Because the current human reference GRCh38 has a gap at the *TCAF* locus, we constructed two custom references for chromosome 7: one with Haplogroup 4 and the other with Haplogroup 5, in additional to a *TCAF*-SDs-hardmasked (chr7:143521769-143874696, GRCh38) chromosome 7. Paired-end data were aligned to the two custom references using BWA-MEM (v0.7.12), while ancient DNA data were mapped using BWA (v0.7.17) with parameters with parameters: –n 0.01 –o 2 –l 16500 [19]. SNV genotypes were jointly generated using haplotype caller FreeBayes (v1.0.2-6-g3ce827d) with the following parameters: "--min-coverage 10 --use-best-n-alleles 4". To ensure genotype quality, we excluded variants that were found within 10 bp of indels and have quality score (QUAL) < 20. We identified a total of 1,275 and 1,295 SNVs from the call sets of Haplogroups 4 and 5, respectively, among the 840 samples. Note that we excluded 38 and 30 samples from the call sets of Haplogroups 4 and 5, respectively, due to >10% missing data to ensure the quality of genotype calls and downstream analyses. Because the results and interpretations of all downstream analyses are highly compatible, we concluded no obvious reference biases between the two sets. Unless otherwise stated, we reported results from the Haplogroup 4 call set because it has slightly better mapping quality.

Phylogenetic analyses

To infer the phylogenetic relationships for *TCAF* SDs and the unique diploid sequences in primates, we performed both maximum likelihood (IQ-TREE, v1.6.11)[53] and Bayesian phylogenetic-based (BEAST v2.5.0) [54] analyses. First, we determined parameters and models using ModelGenerator (v0.85) [55] and compared with the recommendations from the model-selection feature of IQ-TREE (using the parameters: "-bb 1000 -redo -nt 1 -m MFPMERGE"). Based on the best-fit model recommended, to run BEAST, in general we set 1) GAMMA Category Count = 5, shape parameter = 0.1, Proportion Invariant = 0.38, and

using the GTR substitution model (all rates = 1.0) for the Site Model, and 2) Relaxed Clock Log Normal over branches on the tree for Clock Model. For tree priors, we used Calibrated Yule Model, and while kept most of the parameters of the priors as default, for birthRate and clockRate we used Gamma(0.001, 1000) and set calibrations based on human–chimpanzee and human–rhesus macaque divergence following the distributions of a log-normal(M = 6500000, S = 0.09) and log-normal(M = 10000000, S = 0.09), respectively. For the analysis involving the inferred archaic hominin haplotypes, we also set dates for terminal samples to account for the differences between contemporary and fossil samples. For each run, we drew a date (in years) from a uniform [36000, 100000] for the Neanderthals and a uniform [50000, 80000]. For each locus, we performed five independent runs to infer the phylogeny using a chain length of 50,000,000 samples and recorded every 1,000 samples. We used the accompany program Tracer (v.1.7.1) to determine the quality of each run and, in general, we used the first 10% as burn-in and only kept runs with ESS > 500. All phylogenetic trees were plotted using Figtree (v1.4.3).

Population genetic analyses

All population genetic analyses in this study were based on SNVs from the three unique diploid regions at the *TCAF* locus (chr7:143,501,000-143,521,000, chr7:143,729,525-143,741,525, chr7:143,875,000-143,895,000 [GRCh38]) using the HGDP, archaic hominin, and chimpanzee samples. Haplotypes were inferred by applying BEAGLE (v5.1-25Nov19.28d) with default settings to modern human, archaic hominin, and chimpanzee samples separately. We explored the diversity of the *TCAF* haplotype using a haplotype-based PCA to classify and assign haplotypes into individual clusters. Cluster assignments were based on the supervised k-means algorithm using the top N informative PCs, where $\sum_{i=1}^{N} Var(PC_i) > 0.9$. To determine the best number of cluster k, for each possible k between 1 and 20, we computed the within groups sum of squares (WSS) using a distance matrix among the top N PCs and determined the best k when $WSS_{k+1} - WSS_k < delta$, where delta was chosen as 500 according to our analysis. Data visualization of the haplotype clustering was performed using the machine learning technique t-distributed stochastic neighbor embedding (t-SNE) [28] implemented in the R package Rtsne

(https://github.com/jkrijthe/Rtsne) using the following parameters: "perplexity=50, max_iter=5000, early_exag_coeff=12, exaggeration=4, stop_lying_iter=1000, check_duplicates = FALSE, dims=3". We optimized the t-SNE analysis by running five replicates of this computation and identifying the smallest Kullback-Leibler distance as the candidate for final visualization [28].

To test for signals of natural selection, we computed the Tajima's *D* statistic [56] over 2000 bp windows over the 52.3 kbp unique diploid sequences for individual populations as well as the archaic samples. Because the Tajima's D statistic is sensitive to population history, we performed a large coalescent simulation using MaCS [57] and 1,000 demographic models as described previously [25,58] to account for past histories between modern and archaic hominins and among multiple African and non-African populations. Note that our simulations carefully match the local mutation rate and recombination rate variation at these regions to avoid possible biases to our selection inferences. To estimate the TMRCA for a locus of interest, we used the Thomson's estimator [59]. LD measurements among SNVs were computed using Lewontin's D´ [60] and R^2 implemented in PLINK (v1.09, www.cog-genomics.org/plink/1.9/) [61]. Note only SNVs with minor allele frequencies > 10% were included for this analysis because lower frequency variants are less informative about linkage.

ACKNOWLEDGEMENTS

The authors thank T. Brown for assistance in editing this manuscript. We also thank S.C. Murali and D.S. Gordon for help submitting data to the NCBI database. **Funding:** This work was supported, in part, by the US National Institutes of Health (NIH) grant R01HG002385 to E.E.E. P.H. is supported by the NIH Pathway to Independence Award (NHGRI, K99HG011041). S.C. was supported by a National Health and Medical Research Council (NHMRC) C. J. Martin Biomedical Fellowship (1073726). E.E.E. is an investigator of the Howard Hughes Medical Institute. **Author contributions:** P.H. and E.E.E. designed and planned experiments. V.D., C.B., S.C., A.P.L., K.M.M., M.S., A.E.W., and J.G.U. prepared libraries and generated and analyzed sequencing data. P.H., V.D., M.R.V., and T.H. performed variant calling and bioinformatics analyses. P.H., V.D., M.R.V., and P.C.D. analyzed long-read sequencing data and

assembled contigs. P.H. and Y.M. performed population genetic and phylogenetic inferences. A.P.L., K.M.M., P.C.D., and J.G.U. generated Iso-Seq transcript data. P.H. and E.E.E. wrote the manuscript.

Declaration of interests: J.G.U. is an employee of Pacific Biosciences, Inc.

DATA AVAILABILITY: All data used in this study, including assembled BAC contigs and Iso-Seq capture transcript data, are deposited in NCBI under the BioProject: PRJNA657884, BioSample: SAMN09459150; these data are available to anyone for purposes of reproducing or extending the analysis.

REFERENCES

1. Ohno, S. *Evolution by gene duplication*, xv, 160 p. (Springer-Verlag, New York, 1970).
2. Kimura, M. & Ohta, T. On some principles governing molecular evolution. *Proc Natl Acad Sci U S A* **71**, 2848-52 (1974).
3. Boyd, J.L. *et al.* Human-chimpanzee differences in a FZD8 enhancer alter cell-cycle dynamics in the developing neocortex. *Curr Biol* **25**, 772-779 (2015).
4. Charrier, C. *et al.* Inhibition of SRGAP2 function by its human-specific paralogs induces neoteny during spine maturation. *Cell* **149**, 923-35 (2012).
5. Dennis, M.Y. *et al.* Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication. *Cell* **149**, 912-22 (2012).
6. Florio, M. *et al.* Human-specific gene ARHGAP11B promotes basal progenitor amplification and neocortex expansion. *Science* **347**, 1465-70 (2015).
7. Fiddes, I.T. *et al.* Human-Specific NOTCH2NL Genes Affect Notch Signaling and Cortical Neurogenesis. *Cell* **173**, 1356-1369 e22 (2018).
8. Dennis, M.Y. & Eichler, E.E. Human adaptation and evolution by segmental duplication. *Curr Opin Genet Dev* **41**, 44-52 (2016).
9. Nuttle, X. *et al.* Emergence of a Homo sapiens-specific gene family and chromosome 16p11.2 CNV susceptibility. *Nature* **536**, 205-9 (2016).
10. Colburn, R.W. *et al.* Attenuated cold sensitivity in TRPM8 null mice. *Neuron* **54**, 379-86 (2007).
11. Key, F.M. *et al.* Human local adaptation of the TRPM8 cold receptor along a latitudinal cline. *PLoS Genet* **14**, e1007298 (2018).
12. Gkika, D. *et al.* TRP channel-associated factors are a novel protein family that regulates TRPM8 trafficking and activity. *J Cell Biol* **208**, 89-107 (2015).
13. Eichler, E.E. Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends Genet* **17**, 661-9 (2001).
14. Chaisson, M.J., Wilson, R.K. & Eichler, E.E. Genetic variation and the de novo assembly of human genomes. *Nat Rev Genet* **16**, 627-40 (2015).

15. Chaisson, M.J.P. *et al.* Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun* **10**, 1784 (2019).
16. Sudmant, P.H. *et al.* Diversity of human copy number variation and multicopy genes. *Science* **330**, 641-6 (2010).
17. Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222-6 (2012).
18. Prado-Martinez, J. *et al.* Great ape genetic diversity and population history. *Nature* **499**, 471-5 (2013).
19. Prufer, K. *et al.* A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science* **358**, 655-658 (2017).
20. Prufer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43-9 (2014).
21. Mafessoni, F. *et al.* A high-coverage Neandertal genome from Chagyrskaya Cave. *Proc Natl Acad Sci U S A* **117**, 15132-15136 (2020).
22. Bergstrom, A. *et al.* Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367**(2020).
23. Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201-206 (2016).
24. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444-54 (2006).
25. Hsieh, P. *et al.* Adaptive archaic introgression of copy number variants and the discovery of previously unknown human genes. *Science* **366**(2019).
26. Magness, C.L. *et al.* Analysis of the Macaca mulatta transcriptome and the sequence divergence between Macaca and human. *Genome Biol* **6**, R60 (2005).
27. Kronenberg, Z.N. *et al.* High-resolution comparative analysis of great ape genomes. *Science* **360**(2018).
28. Maaten, L.v.d. & Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **9**, 2579-2605 (2008).
29. Stephens, J.C. *et al.* Haplotype variation and linkage disequilibrium in 313 human genes. *Science* **293**, 489-93 (2001).
30. Biondi, B., Kahaly, G.J. & Robertson, R.P. Thyroid Dysfunction and Diabetes Mellitus: Two Closely Associated Disorders. *Endocr Rev* **40**, 789-824 (2019).
31. Nurk, S. *et al.* HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *bioRxiv*, 2020.03.14.992248 (2020).
32. Garg, S. *et al.* Accurate chromosome-scale haplotype-resolved assembly of human genomes. *bioRxiv*, 810341 (2020).
33. Porubsky, D. *et al.* A fully phased accurate assembly of an individual human genome. *bioRxiv*, 855049 (2019).
34. Vollger, M.R. *et al.* Improved assembly and variant detection of a haploid human genome using single-molecule, high-fidelity long reads. *Ann Hum Genet* **84**, 125-140 (2020).
35. Chen, J.M., Cooper, D.N., Chuzhanova, N., Ferec, C. & Patrinos, G.P. Gene conversion: mechanisms, evolution and human disease. *Nat Rev Genet* **8**, 762-75 (2007).
36. Osada, N. & Innan, H. Duplication and gene conversion in the Drosophila melanogaster genome. *PLoS Genet* **4**, e1000305 (2008).
37. Hayakawa, T. *et al.* A human-specific gene in microglia. *Science* **309**, 1693 (2005).

38. Lynch, M. & Katju, V. The altered evolutionary trajectories of gene duplicates. *Trends Genet* **20**, 544-9 (2004).
39. Marques-Bonet, T., Girirajan, S. & Eichler, E.E. The origins and impact of primate segmental duplications. *Trends Genet* **25**, 443-54 (2009).
40. Faria, R., Johannesson, K., Butlin, R.K. & Westram, A.M. Evolving Inversions. *Trends Ecol Evol* **34**, 239-248 (2019).
41. Merot, C., Llaurens, V., Normandeau, E., Bernatchez, L. & Wellenreuther, M. Balancing selection via life-history trade-offs maintains an inversion polymorphism in a seaweed fly. *Nat Commun* **11**, 670 (2020).
42. Kim, B. Thyroid hormone as a determinant of energy expenditure and the basal metabolic rate. *Thyroid* **18**, 141-4 (2008).
43. Mullur, R., Liu, Y.Y. & Brent, G.A. Thyroid hormone regulation of metabolism. *Physiol Rev* **94**, 355-82 (2014).
44. Stenzel, D. & Huttner, W.B. Role of maternal thyroid hormones in the developing neocortex and during human evolution. *Front Neuroanat* **7**, 19 (2013).
45. Forhead, A.J. & Fowden, A.L. Thyroid hormones in fetal growth and prepartum maturation. *J Endocrinol* **221**, R87-R103 (2014).
46. McCoy, D.D. *et al.* Enhanced insulin clearance in mice lacking TRPM8 channels. *Am J Physiol Endocrinol Metab* **305**, E78-88 (2013).
47. Reimundez, A. *et al.* Deletion of the Cold Thermoreceptor TRPM8 Increases Heat Loss and Food Intake Leading to Reduced Body Temperature and Obesity in Mice. *J Neurosci* **38**, 3643-3656 (2018).
48. Chen, J. *et al.* Genome-wide association study of type 2 diabetes in Africa. *Diabetologia* **62**, 1204-1211 (2019).
49. Dobson, J.E. The iodine factor in health and evolution. *Geographical Review* **88**, 3-28 (1998).
50. Hach, F. *et al.* mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nat Methods* **7**, 576-7 (2010).
51. Steinberg, K.M. *et al.* Structural diversity and African origin of the 17q21.31 inversion polymorphism. *Nat Genet* **44**, 872-80 (2012).
52. Vollger, M.R. *et al.* Long-read sequence and assembly of segmental duplications. *Nat Methods* **16**, 88-94 (2019).
53. Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A. & Jermiin, L.S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* **14**, 587-589 (2017).
54. Bouckaert, R. *et al.* BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput Biol* **10**, e1003537 (2014).
55. Keane, T.M., Creevey, C.J., Pentony, M.M., Naughton, T.J. & McLnerney, J.O. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evol Biol* **6**, 29 (2006).
56. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585-95 (1989).
57. Chen, G.K., Marjoram, P. & Wall, J.D. Fast and flexible simulation of DNA sequence data. *Genome Res* **19**, 136-42 (2009).

58. Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H. & Bustamante, C.D. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet* **5**, e1000695 (2009).
59. Thomson, R., Pritchard, J.K., Shen, P., Oefner, P.J. & Feldman, M.W. Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. *Proc Natl Acad Sci U S A* **97**, 7360-5 (2000).
60. Lewontin, R.C. The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. *Genetics* **49**, 49-67 (1964).
61. Chang, C.C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).

Table 1. Summary of 15 assembled *TCAF* haplotypes constructed using large-insert BAC libraries and long-read sequencing. BAC clones were selected and sequenced using the PacBio long-read sequencing technology and assembled into individual haplotypes (**Methods**). Copy number of *TCAF* segmental duplication (SD) cassettes and the classification for individual haplotypes were determined by Miropeats and sequence alignment analysis (**Figures 2 and S5-S13**).

Haplotype ID	BAC library (species or population)	Length (bp)	Length of *TCAF* SD cassettes (bp)	Copy number of *TCAF* SD cassettes	%GC	Haplogroup
CHM1	CHM1	368,013	277,806	2	39.83	Haplogroup 2-1
VMRC53_hapA	NA12878 (European)	433,048	277,806	2	39.56	Haplogroup 2-2
VMRC53_hapB	NA12878 (European)	425,306	273,483	2	39.63	Haplogroup 3-2
VMRC61_hapA	HG00732 (Puerto Rican)	337,690	277,856	2	39.91	Haplogroup 2-2
VMRC61_hapB	HG00732 (Puerto Rican)	435,583	405,366	3	39.71	Haplogroup 4
VMRC62_hapA	HG00733 (Puerto Rican)	395,405	277,854	2	39.60	Haplogroup 2-2
VMRC64_hapA	NA19240 (Yoruba)	323,367	260,853	2	39.94	Haplogroup 2-2
VMRC64_hapB	NA19240 (Yoruba)	348,654	277,808	2	40.07	Haplogroup 2-2
VMRC66_hapA	NA19434 (Luhya)	496,357	406,131	3	40.00	Haplogroup 5
VMRC69_hapA	HG00514 (Han Chinese)	387,079	277,712	2	39.29	Haplogroup 3-1
VMRC73_hapA	GM10539 (Melanesian)	247,628	145,427	1	39.93	Haplogroup 1
VMRC73_hapB	GM10539 (Melanesian)	222,558	145,424	1	39.85	Haplogroup 1
CH251_contig	CH251 (Pan troglodytes)	273,442	127,988	1	39.90	Ancestral-like
CH277_contig	CH277 (Gorilla gorilla)	241,956	140,234	1	39.98	Ancestral-like
CH250_contig	CH250 (Rhesus macaque)	225,909	140,184	1	40.49	Ancestral-like

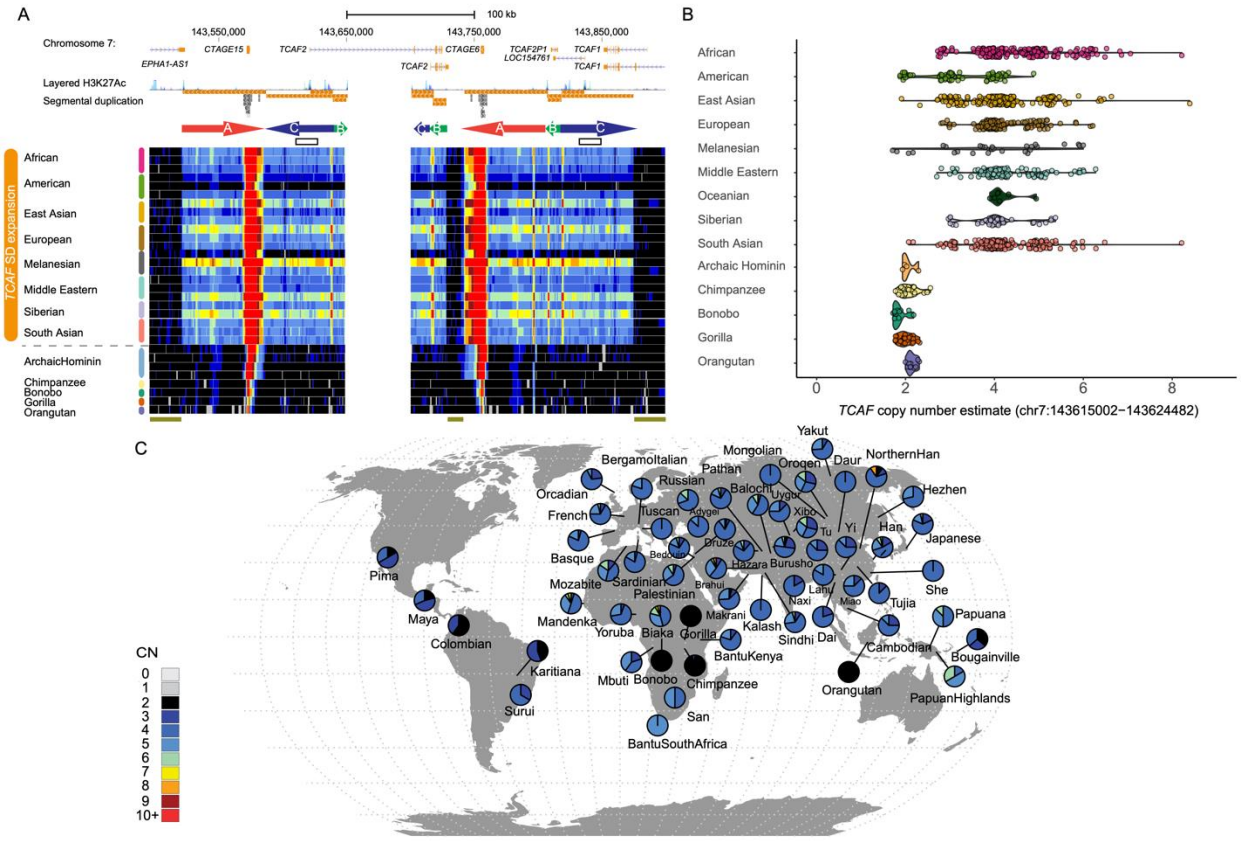

Figure 1. Copy number variation of *TCAF* SDs in a collection of diverse human and nonhuman samples. Copy numbers were estimated using read-depth based genotyping method. (A) Copy number (CN) heat maps, where each row represents the CN of a sample over the *TCAF* locus. The colored arrows (A, B, and C) represent the three major SD blocks in this region. The white area in the middle represents the gap present in the human reference genome (GRCh38). The two white boxes show a putative interlocus gene conversion event that correlates with latitudinal locations of populations (**Figure S16-S18**). Dark green bars at the bottom indicate the unique diploid sequences used for downstream phylogenetic and population genetic analyses (chr7:143,501,000-143,521,000, chr7:143,729,525-143,741,525, chr7:143,875,000-143,895,000 [GRCh38]). (B) Distributions of the overall *TCAF* CN genotypes among samples from nonhuman great apes, archaic hominins, and modern humans using a representative region (chr7:143,615,002−143,624,482). (C) Geographic distribution of the overall *TCAF* CN genotypes in the 54 Human Genome Diversity Project (HGDP) populations and the nonhuman great ape samples. Geo-coordinates for the populations are listed in **Table S8**. Pie charts shows the CN frequency distribution for a given population (colors correspond to those in the CN heat maps).

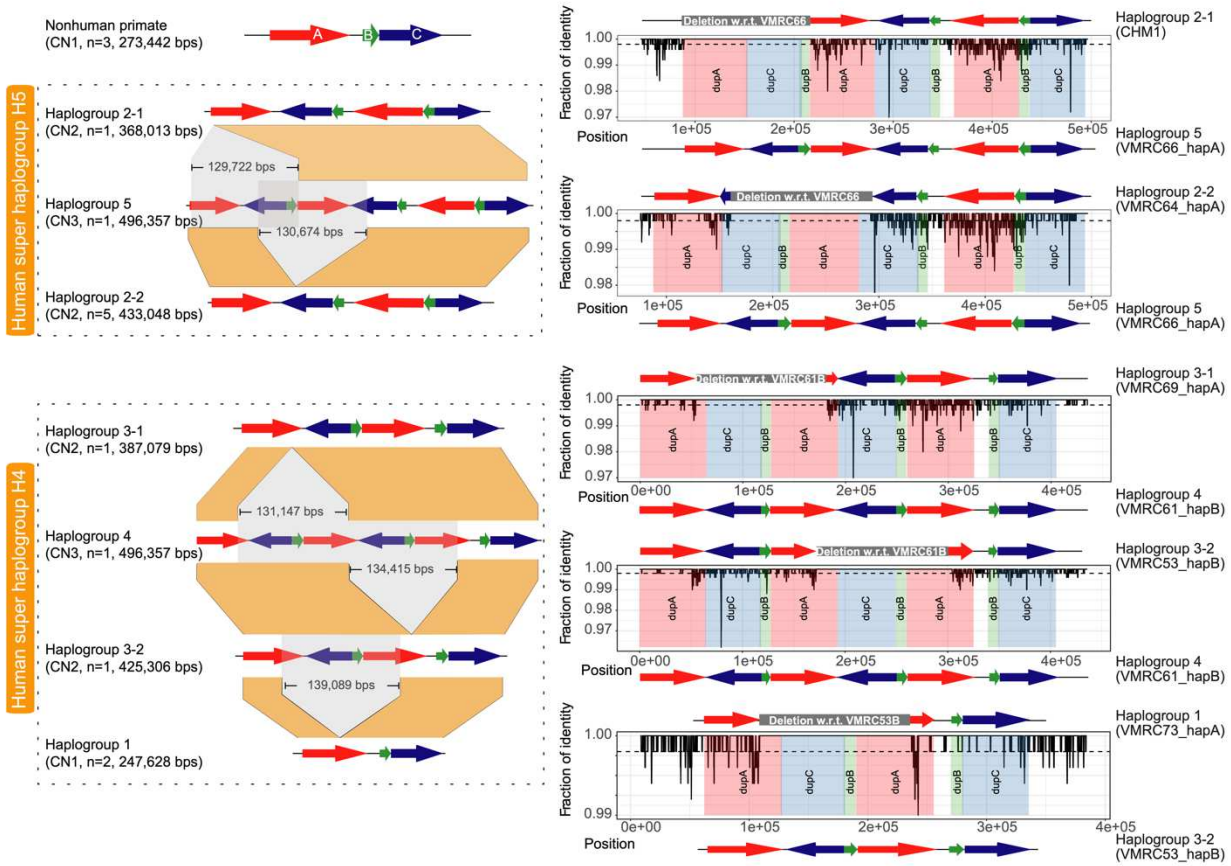

Figure 2. Complex structural haplotype diversity at the 7q35 *TCAF* locus in humans. Left panel: Schematic representations of the major- (super haplogroups H4 and H5) and sub-haplogroups identified from 15 targeted BAC long-read assemblies. Colored arrows are *TCAF* SDs, and the gray and orange areas indicate putative deletion events and orthologous sequences between haplogroups. For example, despite the similarity in structure, Haplogroup 2-1 and Haplogroup 2-2 are likely resulted from different deletion events from Haplogroup 5 based on comparative sequence data. Right panel: The comparative sequence analysis between individual haplogroups shows pairwise sequence identities (black lines) over 500 bp windows (sliding by 100 bp). With each plot, colored rectangles correspond to *TCAF* SDs, while the white region embedded between a DupA (red) and a DupB (green) blocks represents a 12.3 kbp single-copy unique sequence. Gray rectangles above each sequence identity plots indicate the locations of putative structure variants (**Table S3**).

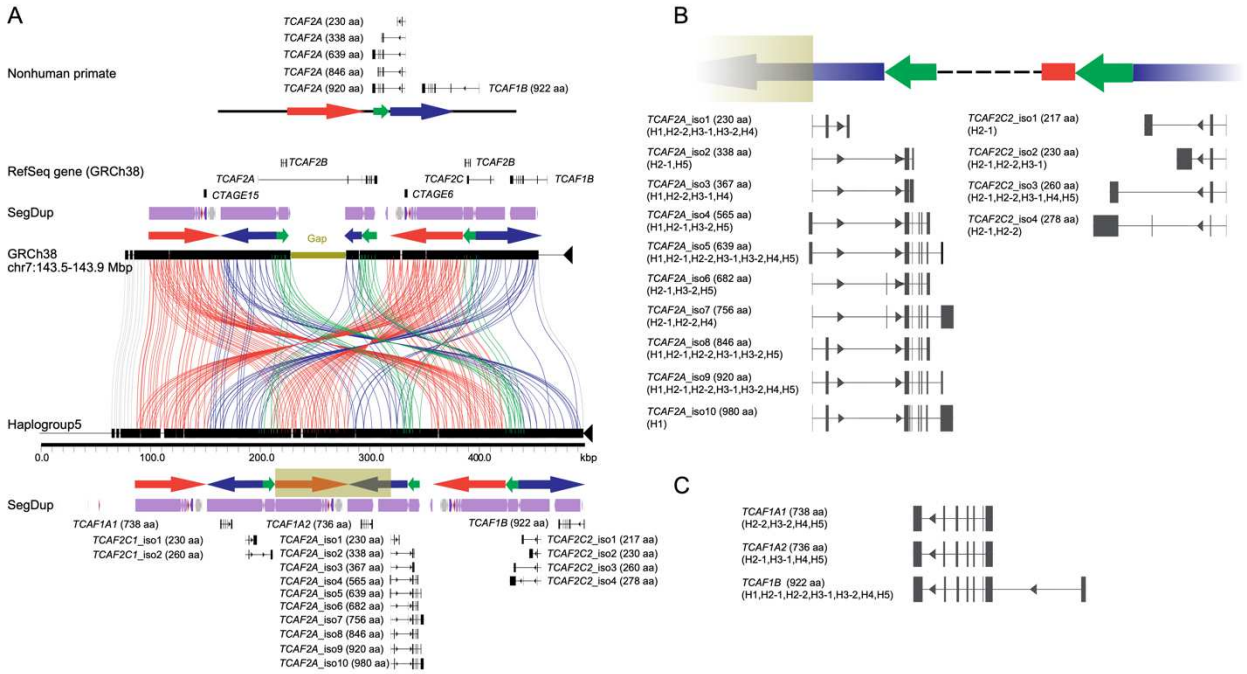

Figure 3. *TCAF* gene and isoform diversity among human haplotypes. (A) Miropeats analysis reveals a great consistency between Haplogroup 5 and the gapped 7q35 *TCAF* locus in the human reference genome (GRCh38). Colored arrows are annotated *TCAF* SDs and lines connecting the sequences show regions of homology. The dark green rectangle in the sequence-resolved, gap-free Haplogroup 5 contig (bottom) represents the region missing in the GRCh38 sequence (dark green bar, top). Additional annotations include a schematic of *TCAF* sequence structure and gene models/isoforms for nonhuman primate, RefSeq gene track (GRCh38), segmental duplication (SegDup) tracks, BAC clones used in the Haplogroup 5 assembly, and predicted *TCAF* models and isoforms using full-length non-chimeric transcripts from six human tissues (**Methods**). Note that the *TCAF2A* gene model in GRCh38 is incorrect due to the presence of the gap in the middle of the *TCAF* SD region. Also note that the numbers above overlap between two BAC clones and indicate the percent sequence identity (#identical bases/total bases). For illustration, all predicted gene models and isoforms are aligned to Haplogroup 5 sequence for (B) *TCAF2A* and (C) *TCAF2C*, along with annotations for amino acid sequence lengths and haplogroups in which they are observed. For simplicity, we skip sequences between *TCAF2A* and *TCAF2C2* (dashed line). Detailed haplotype-specific and/or tissue-specific gene models and isoforms can be found in **Figures S6-S13** and **Table S5**.

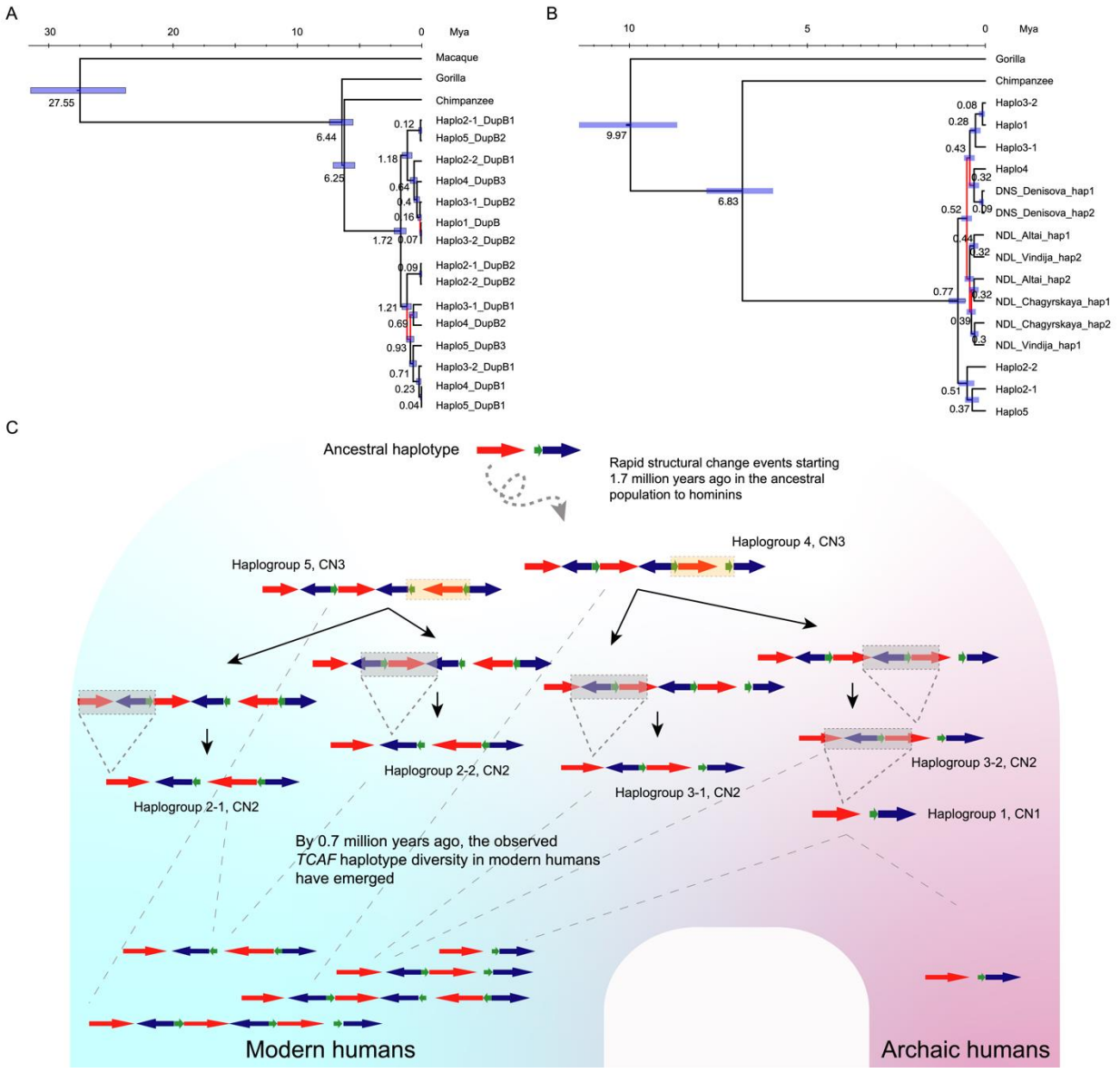

Figure 4. Evolutionary reconstruction of *TCAF* structural diversity. (A) Phylogeny of the haplogroups was inferred using *TCAF* DupB sequences and BEAST (v.2.6.2) with five independent runs of 10 million iterations of Markov Chain Monte Carlo (**Methods**). Numbers and horizontal bars at internal nodes indicate point estimates and 95% highest posterior density intervals for the divergences (in million years ago, Mya), respectively. Branches with posterior probabilities <90% are colored in red. See **Figure S14-S16** for results of other SD sequences. (B) Inferred phylogeny of the modern human haplogroups and archaic hominin haplotypes using the 12 kbp unique sequences embedded within *TCAF* SDs (**Figure 1**). Haplotypes of archaic samples were generated using high-confident single-nucleotide variants (SNVs) called within the unique diploid region. Phylogenetic inference was performed similarly as described above. (C) Schematic model for the evolution of *TCAF* haplotypes in humans based on phylogenetic inferences (**Figures 4A-4B and S14-S17**). Colored arrows are *TCAF* SDs; orange and gray areas indicate relative inversion and deletion events between haplogroups, respectively. Short dashed lines indicate putative breakpoints of structural changes between haplogroups, while the long-dashed lines illustrate lineage sorting.

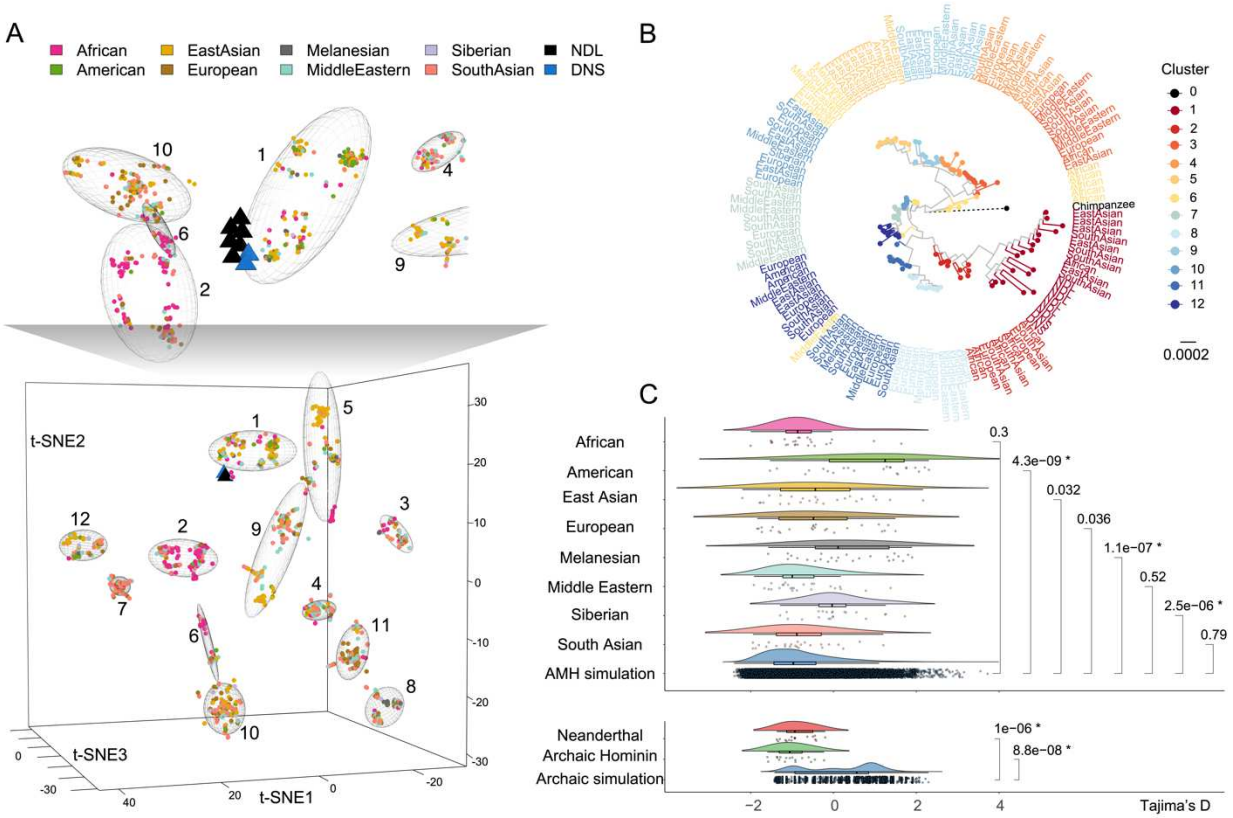

Figure 5. Archaic hominin versus human haplotype diversity. Haplotypes were inferred using 1,275 SNVs in the three unique diploid sequences around the *TCAF* SD region (**Figures 1 and 2**). (A) Haplotype-based principle component analysis was performed, followed by haplotype clustering and cluster visualization using the dimension-reduction technique, t-SNE (**Methods**). On the t-SNE plot, each dot/triangle is a haplotype and colored according to population/species origin. NDL and DNS refer to Neanderthal (black triangle) and Denisovan (blue triangle) haplotypes. Numbers and ellipses in the 3D t-SNE plots indicate individual clusters (see also **Figures S24-S29**). The zoom-in above the 3D t-SNE shows that all archaic haplotypes are in tight proximity to each other. (B) The maximum likelihood phylogeny was constructed using 10 randomly selected haplotypes from the 12 inferred clusters, in addition to eight archaic and one chimpanzee haplotypes. Note that the branch length of chimpanzee (dashed line) is truncated by 90% of its actual length for the purpose of illustration. (C) Distributions of Tajima's D statistic computed using 2 kbp windows sliding over the three unique diploid regions. Significance of natural selection signals was determined using coalescent simulations based on 1,000 different demographic models (**Methods**). The asterisks indicate significant tests after the Bonferroni correction.

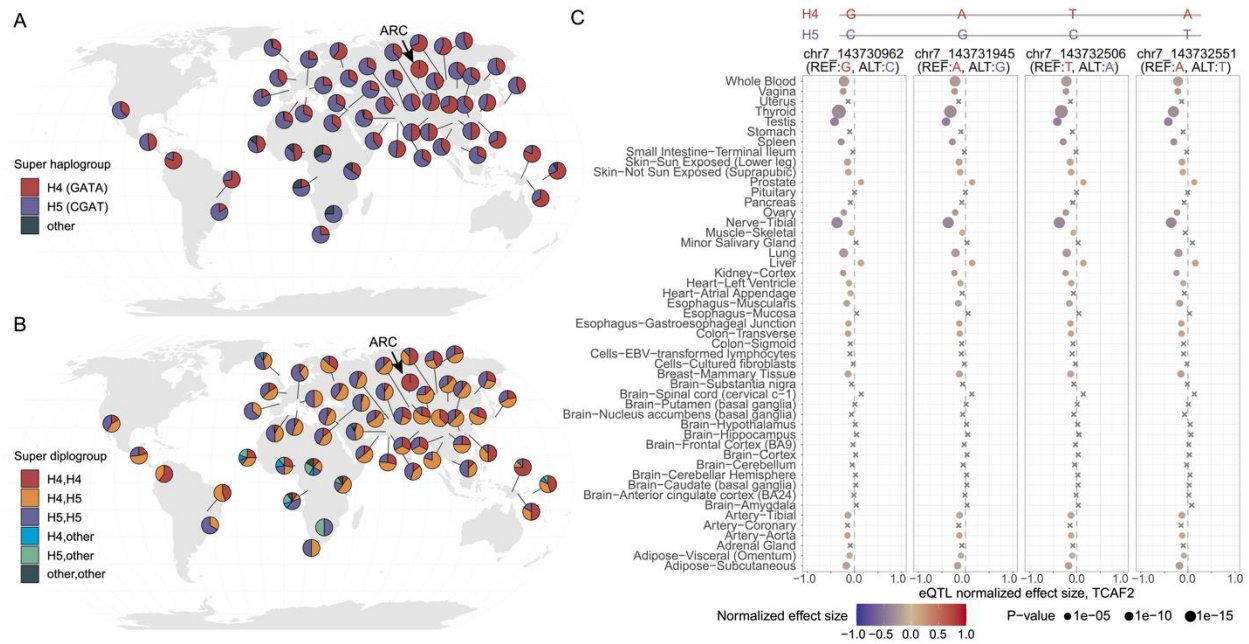

Figure 6. H4 and H5 super haplogroup distribution and eQTL analysis. Four tagging SNVs were first identified by perfectly separating super haplogroups H4 and H5 among the seven BAC haplogroups and confirmed based on patterns of linkage disequilibrium (LD) in samples from the HGDP panel. (A and B) Distributions of super haplogroups H4 (red, 57.1%) and H5 (blue, 40.3%) and their diploid type in the HGDP populations across the world. Other haplotypes were found in ~2.6% of the samples. Note that archaic hominin haplotypes from the Neanderthal (n = 3) and Denisovan (n = 1) samples all carry H4 haplotypes, and as a representation the geographic location of these samples were placed in the Altai Mountains in Siberia (arrows). (C) Multi-tissue eQTL plots show consistent patterns of associations between the four tagging SNVs and expression levels of *TCAF2* across 50 tissues. Effect sizes were calculated as the effect of the alternative allele (blue) relative to the reference allele (read) as defined in GTEx (release v8) and are scaled using color. The (unadjusted) p values of eQTL association are represented by dot sizes. Note that crosses (x) indicate insignificant associations.

Figures

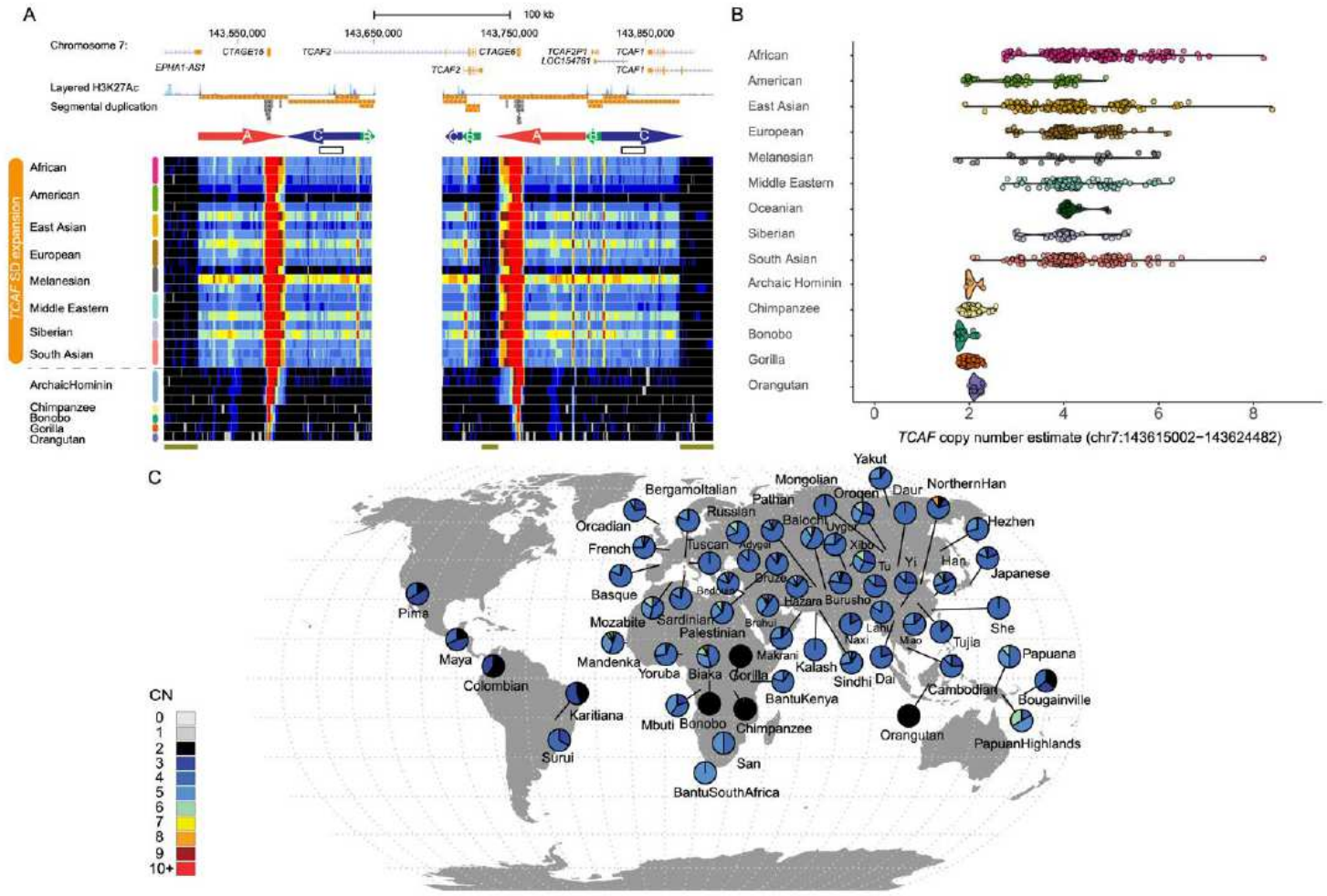

Figure 1

Copy number variation of TCAF SDs in a collection of diverse human and nonhuman samples. Copy numbers were estimated using read-depth based genotyping method. (A) Copy number (CN) heat maps, where each row represents the CN of a sample over the TCAF locus. The colored arrows (A, B, and C) represent the three major SD blocks in this region. The white area in the middle represents the gap present in the human reference genome (GRCh38). The two white boxes show a putative interlocus gene conversion event that correlates with latitudinal locations of populations (Figure S16-S18). Dark green bars at the bottom indicate the unique diploid sequences used for downstream phylogenetic and population genetic analyses (chr7:143,501,000-143,521,000, chr7:143,729,525-143,741,525, chr7:143,875,000-143,895,000 [GRCh38]). (B) Distributions of the overall TCAF CN genotypes among samples from nonhuman great apes, archaic hominins, and modern humans using a representative region (chr7:143,615,002–143,624,482). (C) Geographic distribution of the overall TCAF CN genotypes in the 54 Human Genome Diversity Project (HGDP) populations and the nonhuman great ape samples. Geo-coordinates for the populations are listed in Table S8. Pie charts shows the CN frequency distribution for a given population (colors correspond to those in the CN heat maps). Note: The designations employed

and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

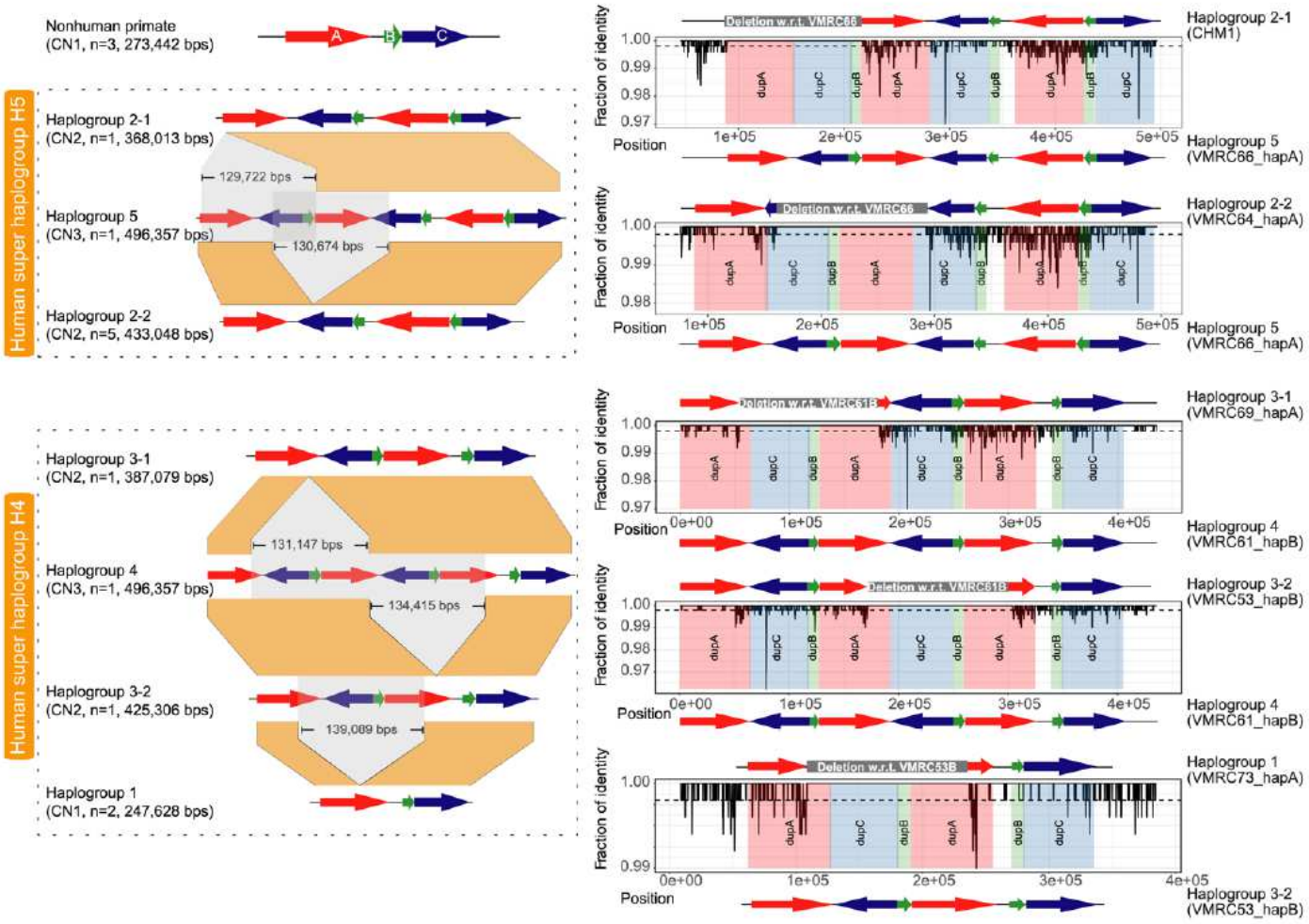

Figure 2

Complex structural haplotype diversity at the 7q35 TCAF locus in humans. Left panel: Schematic representations of the major- (super haplogroups H4 and H5) and sub-haplogroups identified from 15 targeted BAC long-read assemblies. Colored arrows are TCAF SDs, and the gray and orange areas indicate putative deletion events and orthologous sequences between haplogroups. For example, despite the similarity in structure, Haplogroup 2-1 and Haplogroup 2-2 are likely resulted from different deletion events from Haplogroup 5 based on comparative sequence data. Right panel: The comparative sequence analysis between individual haplogroups shows pairwise sequence identities (black lines) over 500 bp windows (sliding by 100 bp). With each plot, colored rectangles correspond to TCAF SDs, while the white region embedded between a DupA (red) and a DupB (green) blocks represents a 12.3 kbp single-copy unique sequence. Gray rectangles above each sequence identity plots indicate the locations of putative structure variants (Table S3).

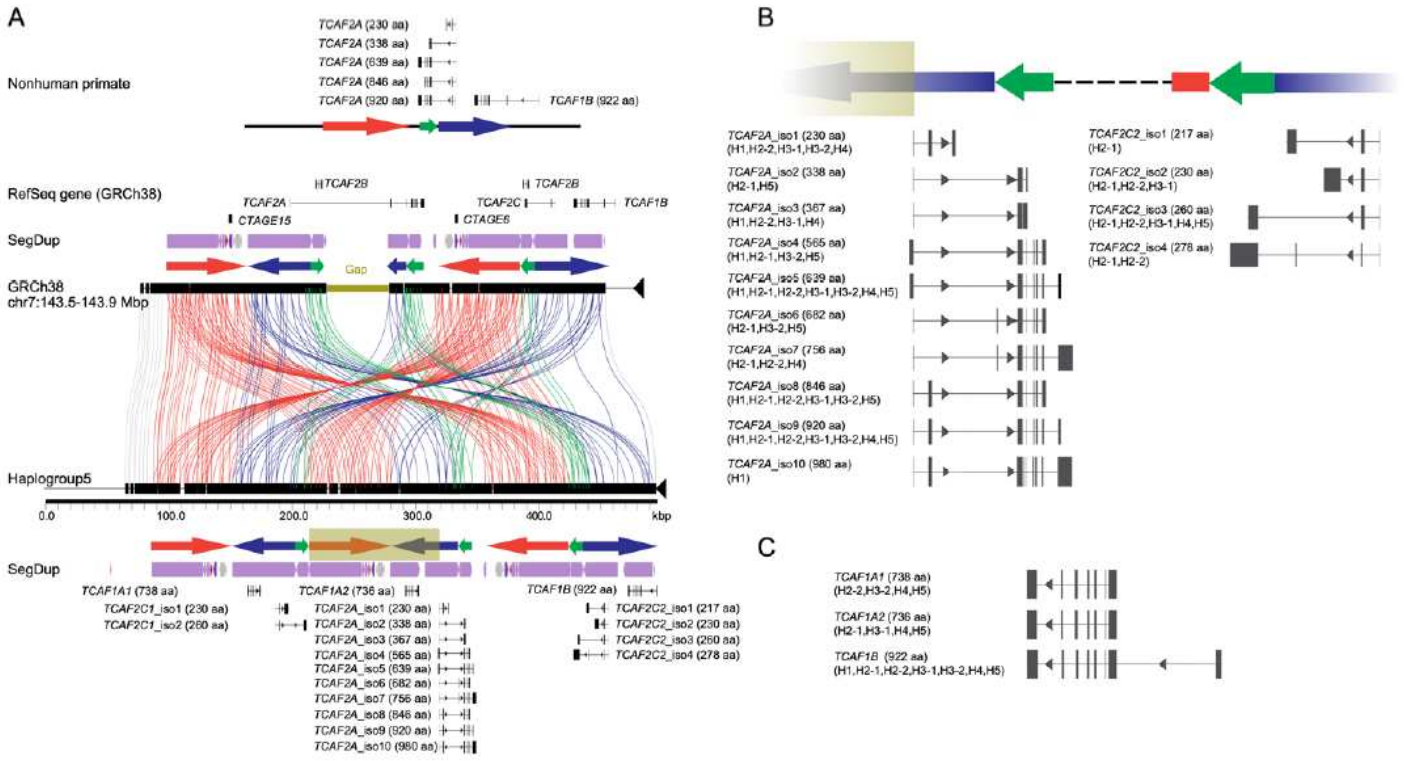

Figure 3

TCAF gene and isoform diversity among human haplotypes. (A) Miropeats analysis reveals a great consistency between Haplogroup 5 and the gapped 7q35 TCAF locus in the human reference genome (GRCh38). Colored arrows are annotated TCAF SDs and lines connecting the sequences show regions of homology. The dark green rectangle in the sequence-resolved, gap-free Haplogroup 5 contig (bottom) represents the region missing in the GRCh38 sequence (dark green bar, top). Additional annotations include a schematic of TCAF sequence structure and gene models/isoforms for nonhuman primate, RefSeq gene track (GRCh38), segmental duplication (SegDup) tracks, BAC clones used in the Haplogroup 5 assembly, and predicted TCAF models and isoforms using full-length non-chimeric transcripts from six human tissues (Methods). Note that the TCAF2A gene model in GRCh38 is incorrect due to the presence of the gap in the middle of the TCAF SD region. Also note that the numbers above overlap between two BAC clones and indicate the percent sequence identity (#identical bases/total bases). For illustration, all predicted gene models and isoforms are aligned to Haplogroup 5 sequence for (B) TCAF2A and (C) TCAF2C, along with annotations for amino acid sequence lengths and haplogroups in which they are observed. For simplicity, we skip sequences between TCAF2A and TCAF2C2 (dashed line). Detailed haplotype-specific and/or tissue-specific gene models and isoforms can be found in Figures S6-S13 and Table S5.

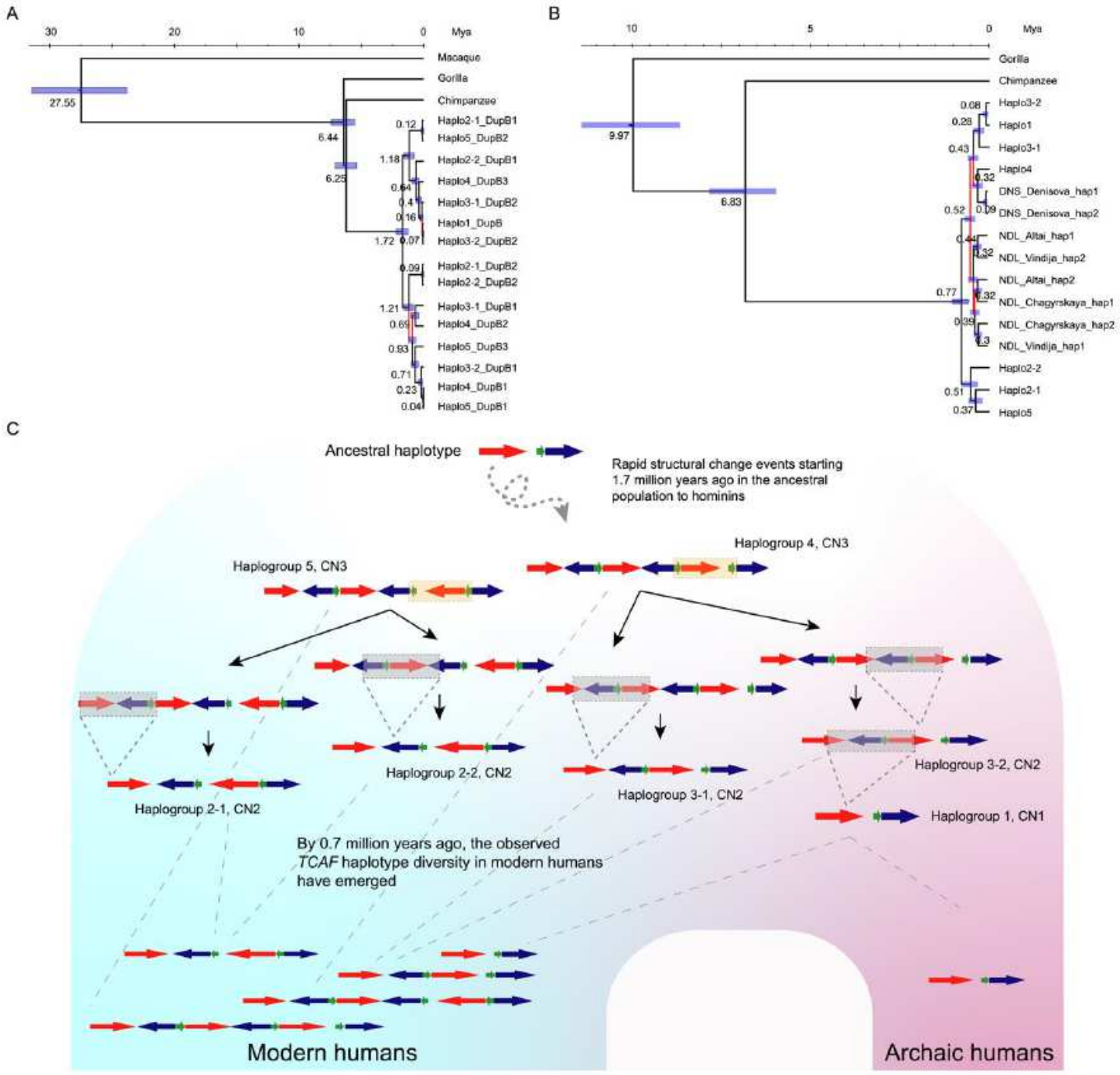

Figure 4

Evolutionary reconstruction of TCAF structural diversity. (A) Phylogeny of the haplogroups was inferred using TCAF DupB sequences and BEAST (v.2.6.2) with five independent runs of 10 million iterations of Markov Chain Monte Carlo (Methods). Numbers and horizontal bars at internal nodes indicate point estimates and 95% highest posterior density intervals for the divergences (in million years ago, Mya), respectively. Branches with posterior probabilities <90% are colored in red. See Figure S14-S16 for results of other SD sequences. (B) Inferred phylogeny of the modern human haplogroups and archaic hominin haplotypes using the 12 kbp unique sequences embedded within TCAF SDs (Figure 1). Haplotypes of archaic samples were generated using high-confident single-nucleotide variants (SNVs) called within the

unique diploid region. Phylogenetic inference was performed similarly as described above. (C) Schematic model for the evolution of TCAF haplotypes in humans based on phylogenetic inferences (Figures 4A-4B and S14-S17). Colored arrows are TCAF SDs; orange and gray areas indicate relative inversion and deletion events between haplogroups, respectively. Short dashed lines indicate putative breakpoints of structural changes between haplogroups, while the long-dashed lines illustrate lineage sorting.

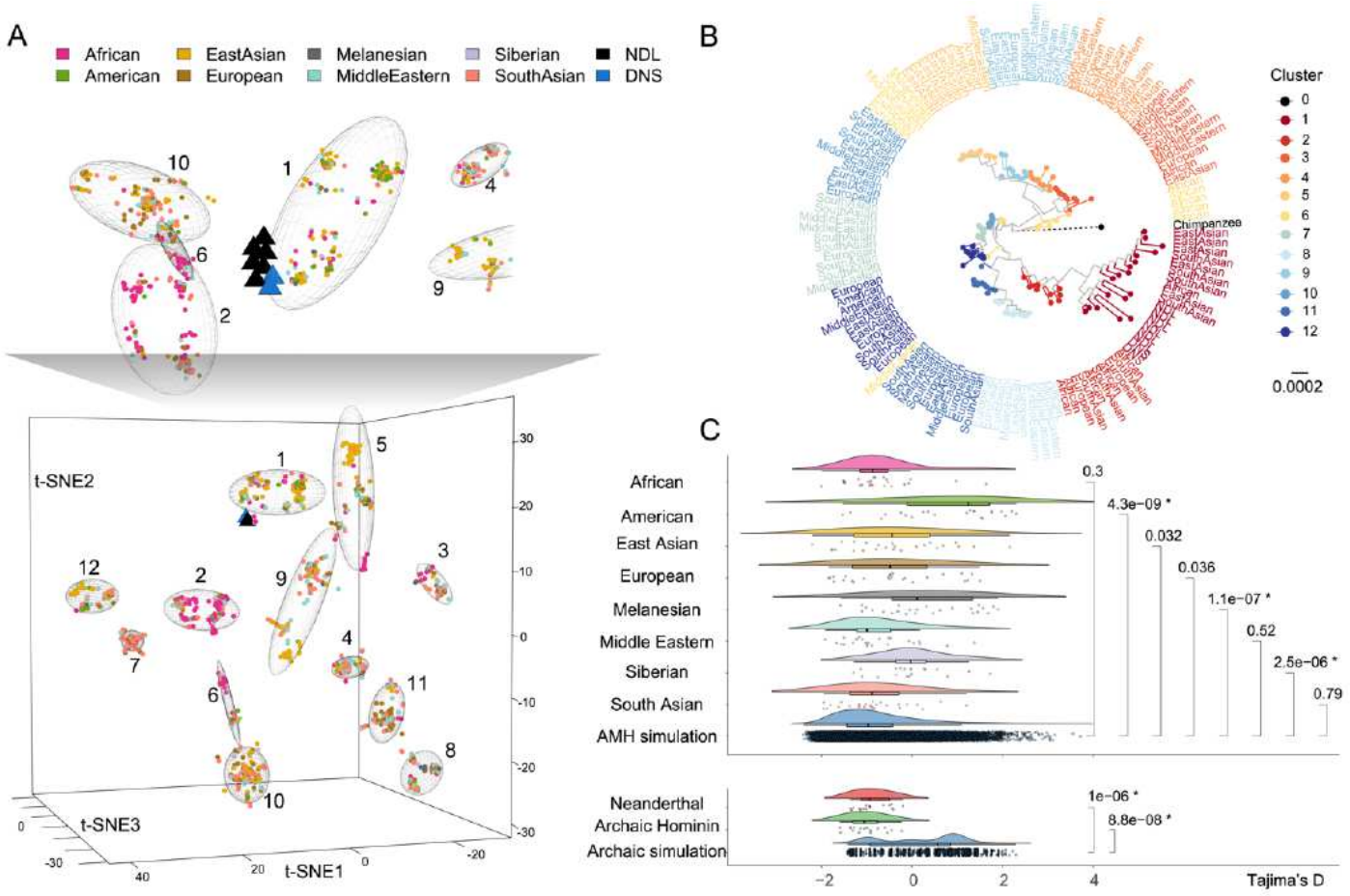

Figure 5

Archaic hominin versus human haplotype diversity. Haplotypes were inferred using 1,275 SNVs in the three unique diploid sequences around the TCAF SD region (Figures 1 and 2). (A) Haplotype-based principle component analysis was performed, followed by haplotype clustering and cluster visualization using the dimension-reduction technique, t-SNE (Methods). On the t-SNE plot, each dot/triangle is a haplotype and colored according to population/species origin. NDL and DNS refer to Neanderthal (black triangle) and Denisovan (blue triangle) haplotypes. Numbers and ellipses in the 3D t-SNE plots indicate individual clusters (see also Figures S24-S29). The zoom-in above the 3D t-SNE shows that all archaic haplotypes are in tight proximity to each other. (B) The maximum likelihood phylogeny was constructed using 10 randomly selected haplotypes from the 12 inferred clusters, in addition to eight archaic and one chimpanzee haplotypes. Note that the branch length of chimpanzee (dashed line) is truncated by 90% of its actual length for the purpose of illustration. (C) Distributions of Tajima's D statistic computed using 2 kbp windows sliding over the three unique diploid regions. Significance of natural selection signals was

determined using coalescent simulations based on 1,000 different demographic models (Methods). The asterisks indicate significant tests after the Bonferroni correction.

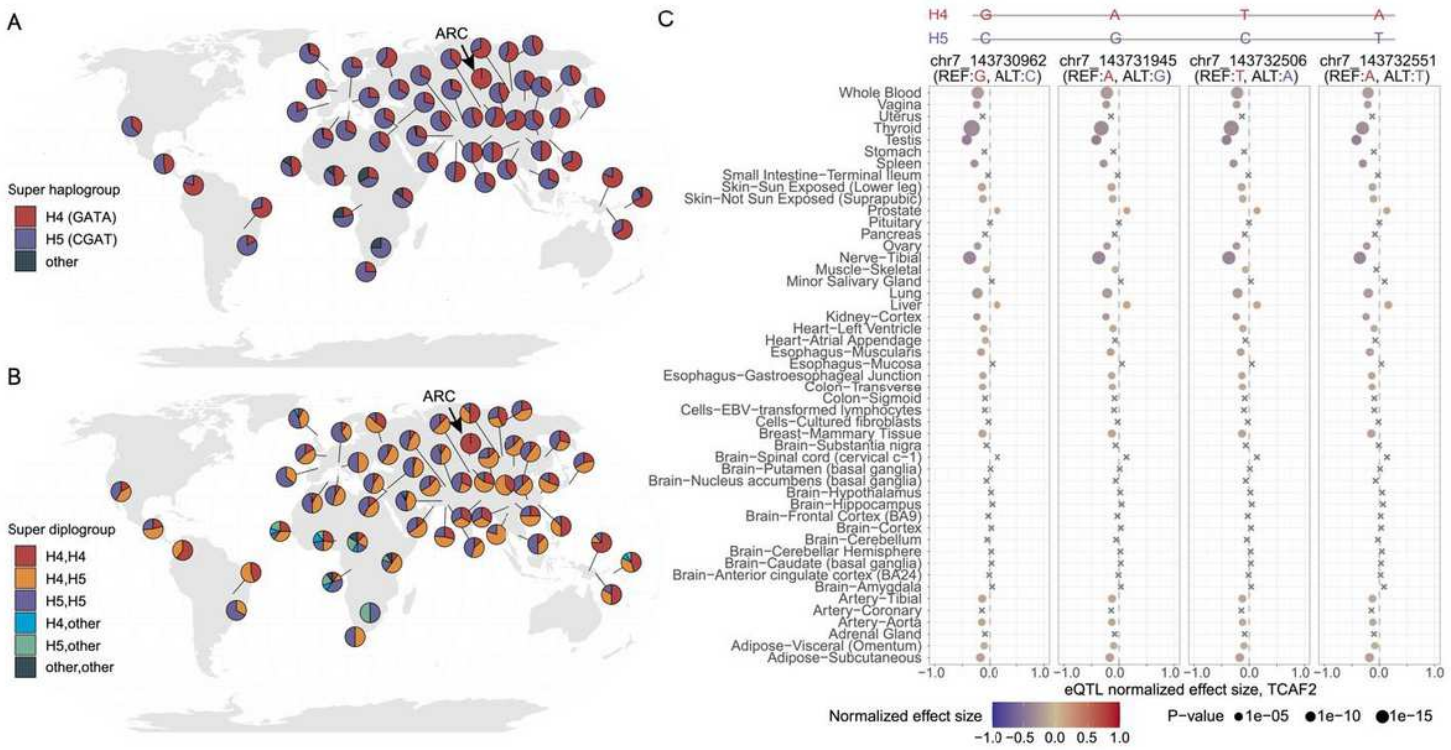

Figure 6

H4 and H5 super haplogroup distribution and eQTL analysis. Four tagging SNVs were first identified by perfectly separating super haplogroups H4 and H5 among the seven BAC haplogroups and confirmed based on patterns of linkage disequilibrium (LD) in samples from the HGDP panel. (A and B) Distributions of super haplogroups H4 (red, 57.1%) and H5 (blue, 40.3%) and their diploid type in the HGDP populations across the world. Other haplotypes were found in ~2.6% of the samples. Note that archaic hominin haplotypes from the Neanderthal (n = 3) and Denisovan (n = 1) samples all carry H4 haplotypes, and as a representation the geographic location of these samples were placed in the Altai Mountains in Siberia (arrows). (C) Multi-tissue eQTL plots show consistent patterns of associations between the four tagging SNVs and expression levels of TCAF2 across 50 tissues. Effect sizes were calculated as the effect of the alternative allele (blue) relative to the reference allele (read) as defined in GTEx (release v8) and are scaled using color. The (unadjusted) p values of eQTL association are represented by dot sizes. Note that crosses (x) indicate insignificant associations. Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TCAFsupplementaryFiguresandTables.v8.pdf
- TableS1VSTpairwiseSuperPop.sorted.v8.xlsx
- TableS7GENECONVIGCtracks.v8.xlsx
- TableS8GeographicCoordinatesHGDPpopsandCNsummary.v8.xlsx