Rapid turnover and recurrent structural variation at the 17q21.31 locus in modern and ancient human genomes, and primates

Samvardhini Sridharan[1,2], Runyang Nicolas Lou[3], Scott Ferguson[3], Joana L Rocha[3], Matthew W Mitchell[4], Alison N Kililia[1], Victor Borda[5], Santiago G. Medina-Munoz[6], Simon Gravel[7], CAAPA2 PopGen Working Group, Brenna Henn[8], Peter H Sudmant[1,2,3]

1 Department of Molecular and Cell Biology, University of California, Berkeley, CA, USA
2 Center for Computational Biology, University of California, Berkeley, CA, USA
3 Department of Integrative Biology, University of California, Berkeley, CA, USA
4 Coriell Institute for Medical Research, Camden, NJ 08103
5 University of Maryland School of Medicine, Baltimore, MD, USA
6 CINVESTAV, Mexico City, Mexico
7 Department of Human Genetics, McGill University, Montreal, Canada
8 Department of Anthropology, University of California, Davis, CA, USA

Correspondence should be addressed to Peter H Sudmant: psudmant@berkeley.edu

ABSTRACT

The 17q21.31 locus in humans harbors several complex structural haplotypes including a ~970kb inversion. Different inversion haplotypes have been associated with susceptibility to microdeletions causing Koolen-de Vries syndrome and variation in fecundity and recombination rates. Here, using 210 haplotype-resolved human genome assemblies and pangenome graph-based approaches we characterize 11 distinct structural haplotypes, several of which have not been previously described. Extending our analyses to a set of haplotype-resolved great-ape genomes, we characterize the structure of a recurrent inversion in chimpanzees which extends an additional 750kb and is ~2 million years younger than the human inversion. We further determine that gorillas exhibit an independent duplication of the *KANSL1* gene, disruption of which in humans is the cause of Koolen-de Vries syndrome. Using short read sequencing data we characterize 17q21.31 haplotype diversity worldwide in ~5174 individuals from 107 populations finding increased frequencies of *KANSL1* duplication-containing haplotypes in both European and South Asian populations as well as 8 double recombination events between inverted and non inverted haplotypes ranging in size from 20-180kb. Finally, using 626 ancient Eurasian human genomes we show haplotypes containing *KANSL1* duplications have increased ~6-fold over the past 12 thousand years in Europe. Together, our results highlight the dynamics,

complexity, and recurrent, independent evolution of a medically relevant locus across humans and great apes.

INTRODUCTION

Complex structural variants (SVs) play critical roles in human disease[1], diversity[2,3], and evolution[4,5]. Inversions are a particularly intriguing and understudied class of SV which have historically been challenging to assay as they are often flanked by large, complex, segmental duplications[1,6]. Nevertheless, across taxa inversions are powerful drivers of genetic and phenotypic diversity influencing traits such as behavior[7], fitness[8,9], and morphology[8,10]. Inversions can also act as barriers to gene flow between populations, promoting speciation and adaptation to different ecological niches and environmental pressures[11–13].

One remarkable inversion in humans is found at the 17q21.31 locus spanning ~970kb and disrupting patterns of linkage disequilibrium over more than a megabase[14]. Haplotypes at this locus group into two major clades: H1 haplotypes which are the most common in humans, and H2 haplotypes, which are inverted with respect to the H1 haplotype. Since its discovery, this locus has been associated with signatures of positive selection[14], disease[15], and recurrent structural rearrangement ("toggling")[16] throughout the great ape lineage, highlighting its complex evolutionary history.

The 17q21.31 locus is of particular medical importance harboring several genes associated with human disease. These include *MAPT,* which encodes the tau protein found in aggregates in Alzheimer's dementias[17], and *KANSL1,* disruption of which results in Koolen-de Vries Syndrome (KdVS). KdVS is characterized by developmental delay, intellectual disability, hypotonia, epilepsy, distinctive facial features, and congenital malformations across multiple organ systems[18]. While rare truncating mutations of *KANSL1* can cause KdVS[15] it is more commonly caused by 17q21.31 microdeletions which occur at a frequency of ~1/16,000 individuals. These microdeletions encompass 5 genes and are mediated by non-allelic homologous recombination (NAHR) between directly oriented segmental duplications which are present on the H2 (inverted) haplotype, but not the H1 haplotype. H2 carriers are thus at risk for KdVS. However, H2 haplotypes, which are at particularly high frequency in European populations, are also associated with increased fecundity in some[14] and thus may be the target of positive selection[18].

In addition to the non-inverted (H1) and inverted (H2) haplotypes, recent studies have identified additional structural complexity at the 17q21.31 locus[19,20], including two independent partial duplications of the *KANSL1* gene on the H1 and H2 haplotypes, as well as copy number variation of the *NSF* gene. Here, we set out to characterize the recent history and full extent of structural complexity at the 17q21.31 locus using both humans and non-human primate long read haplotype-resolved genome assemblies, extensive human population diversity sequencing, and ancient genomes.

RESULTS

Resolving complex genome structures at the 17q21.31 locus with 210 haplotype resolved genome assemblies

Recently generated long read genomic resources provide an unprecedented opportunity to uncover structural diversity at complex loci such as the 17q21.31 locus. We collected 210 diverse haplotypes including the recently assembled T2T human genome to examine the structural complexity of this region. We first used PGRTK to annotate the repeat architecture of these genomes (**Fig 1A, B, S1**). PGRTK identifies clusters of paralogous and homologous sequence within and between genomes enabling us to compare structures among haplotypes. Previous literature has established that alongside the major inversion haplotypes H1 and H2, several additional regions of copy number variation exist on different haplotypes. These include distinct and independent duplications of the *KANSL1* gene on the H2 and H1 haplotypes, referred to as α and β respectively. Additionally partial duplications of the *NSF* gene, on either the H1 or H2 haplotype, are referred to as γ duplications.

Alongside our PGRTK-based classification of structures we used a pangenome based approach (PGGB) to cluster haplotype architectures (see methods) identifying a total of 24 unique structures (**Fig S1**). Many of these structures were highly similar differing only subtly in the composition of the segmentally duplicated *LRRC37A* gene. We thus further clustered structures based on their pairwise jaccard distance into 11 unique summary structures (**Methods**). Structural differences between these haplotypes were confirmed by plotting a multiway structural alignment using SVbyEye (**Fig 1C**). These clusters exhibit finer resolution than previously developed nomenclature. For example, we identified three structures which can be classified as H1.β1.γ1, based on their single copies of the *KANSL1* and the *NSF* duplications, which we designate these with additional numeric suffixes. We further identified two distinct H1.β1.γ2 haplotypes and three H1.β1.γ3 haplotypes. These distinct structures are mostly differentiated by copy number variation at the *LRRC37A2* gene, which flanks the inversion. Overall, the most commonly observed H1 structures were H1.β1.γ1.1 (37% of H1), H1.β1.γ2.2 (30% of H1), H1.β1.γ3.1 (19% of H1) and H1.β2.γ1 (8% of H1) - making up a total proportion of 94% of H1s (88% overall). Most of the variation in structural diversity we observed is located on H1.β1 haplotypes, which are the highest frequency. Only a single H1.β2 and H2.α2 structure were observed in our dataset and several previously described haplotypes were not present in our long read assemblies (H1.β3.γ1, H2.α1.γ2, H2.α2.γ1), however, these are at lower frequency in the population and were observed in our subsequent analyses of structural variation using short reads. Together, these data recapitulate previously described structures in higher detail and demonstrate additional complexity present in humans.

Recurrent but distinct inversions in humans and chimpanzees

The recent completion of several haplotype resolved T2T ape genomes provides the opportunity to contextualize the evolutionary history of this locus in great apes. To explore the population-level diversity in these non-human species we sequenced and assembled two gorilla genomes (four haplotypes) and assessed a set of recently assembled chimpanzee and bonobo genomes (Rocha et al in prep). Previous work has demonstrated that the H2 orientation is the ancestral

118 architecture of the locus. However, fluorescent *in-situ* hybridization of a panel of several apes
119 has suggested that the 17q21.31 inversion is segregating in orangutans, chimpanzees, and
120 bonobos [16] and has emerged independently in these species (i.e. inversion *toggling*).
121
122 We extracted the orthologous 17q21.31 sequence across five non-human great ape species
123 and constructed structural alignments using SVbyEye (**Fig. 2A,B**). These alignments
124 demonstrate the increasing complexity of segmental duplications in African great apes in
125 comparison to orangutans, as has been previously described [16]. To understand how these
126 structures relate to the human architecture we complemented these analyses by projecting
127 human-identified PGRTK sequence blocks onto the primate genomes (**Fig. 2C**). These
128 projections lose resolution as a function of genetic distance particularly in rapidly evolving
129 sequence such segmental duplications but still highlight major architectural features like the
130 inversion and flanking regions.
131
132 Our analyses revealed several striking, lineage specific structural features at this locus. In
133 gorillas we identified a 500kb duplication overlapping to the *KANSL1* gene. This is an
134 independent duplication of the gene distinct from that observed in human haplotypes. This
135 duplication corresponds to an annotated gene sequence in the recent T2T gorilla genome
136 assembly, though it is unclear if it is a functional duplication. We also identified two inverted
137 chimpanzee haplotypes. These inversions are in the same orientation as the human H1
138 haplotype. We refer to this haplotype as the H1p haplotype, and the ancestral primate haplotype
139 as H2p. The chimpanzee H1p/H2p inversion spans ~1.4Mbp (**Fig 2B**). Alignments between
140 human and chimpanzee haplotypes revealed that while the human and chimpanzee inversions
141 exhibit ~750kb of overlapping sequence, the chimpanzee inversion is distinct, containing
142 approximately 650kb of sequence that is not found in the human inversion (**Fig 2B**). Thus, not
143 only is the chimpanzee inversion the result of an independent event, but it is a completely
144 distinct structural variant. This provides further evidence that the H2 orientation is the ancestral
145 structure as while the H1 and H1p haplotypes differ structurally by hundreds of kb the H2 and
146 H2p haplotypes exhibit highly similar structures.
147
148 We next sought to compare the timing of the different inversion events by aligning ~400kbp of
149 unique (non-duplicated) sequence within the shared inversion block across humans and great
150 apes (**Fig. 2D, methods**). The human H1/H2 inversion exhibits a coalescence time of
151 approximately to 2.38 million years ago, confirming a relatively deep coalescence of the two
152 haplotypes [19]. The chimpanzee H1p/H2p inversion however emerged much more recently
153 ~330,000 years ago. These analyses also revealed extensive incomplete lineage sorting across
154 humans, gorillas, chimpanzees, and bonobos. 54% of trees exhibited a topology with humans
155 as an outgroup to the *pan* and gorilla lineages with the canonical topology represented 46% of
156 the time. Together, these results demonstrate that the 17q21.31 locus is highly dynamic across
157 great apes with distinct but overlapping structural variants occurring independently in several
158 taxa over vastly different timescales.

159 **Timing and recurrence of human 17q21.31 structural haplotypes**

160 Long-read haplotype-resolved assemblies phase genome structures with linked SNPs which
161 can be used to trace the evolutionary history of complex genome architectures. Extended
162 suppression of recombination between the H1 and H2 haplotypes enables dating of the
163 coalescent time of the inversion to ~2.38 million years. However, recombination between H1
164 haplotypes will disrupt patterns of linkage disequilibrium (LD) making it challenging to assess
165 the emergence of specific H1 substructures. We sought to identify high LD blocks adjacent to
166 the distal region of H1 structural variation (containing the *KANSL1* and *NSF* duplications). Such
167 regions can be used to trace the history of these structures. We identified 100kb of sequence
168 adjacent to the structurally variable portion of H1 with a consistent tree topology indicating high
169 LD and reduced recombination (**Fig 3A, B**). Using this sequence we constructed a coalescent
170 tree across 210 human haplotypes and the T2T chimpanzee genome as the outgroup. This tree
171 reveals that H1 *KANSL1* duplications (i.e. H1.β2) largely fall into two clusters, suggesting at
172 least two independent origins of this duplication. In contrast, *NSF* duplications (γ2, γ3) exhibit
173 several independent origins throughout the tree. One particularly large cluster of γ3 duplications
174 was identified, within which all observed γ4 duplications were also found. We also find
175 independent origins of the sub-haplotypes cataloged in **Fig 1A** including multiple origins of the
176 H1.β1.γ1.3 structure. Together, these results highlight that the structure of H1 haplotypes at the
177 17q21.31 locus is highly mutable and predisposed to recurrent structural variation. This includes
178 large-scale gene duplications such as *KANSL1* and *NSF* as well as smaller SVs including copy
179 number changes of *LRRC37A.*

180 **Population diversity of the 17q21.31 locus worldwide using short read data**

181 Despite their immense power, existing long-read datasets remain limited in sample size and
182 population diversity. Nevertheless, large scale short read sequencing datasets from diverse
183 human populations have recently been generated. To characterize patterns of 17q21.31
184 variation across the world we analyzed short read sequencing data from 5174 samples
185 encompassing 107 populations representative of human diversity worldwide (see **Methods,**
186 **Figs S2-5**). Using SNPs tagging the inversion as well as read-depth based copy number
187 quantification we estimated haplotype frequencies assigning individuals to one of four groups:
188 H1.β1, representing H1 haplotypes without *KANSL1* duplications; H1.β2+, representing H1
189 haplotypes with *KANSL1* duplications; H2.α1 representing H2 haplotypes without *KANSL1*
190 duplications; and H2.α2+ representing H2 haplotypes with *KANSL1* duplications (**Fig 4A-D**).
191
192 Worldwide, the most common haplotype is H1.β1,confirming previous work using more limited
193 genome sequencing datasets[19,20]. However, we observe extreme population stratification of
194 complex haplotypes at this locus. European populations exhibit the highest worldwide frequency
195 of H2.α2+ haplotypes (10%-30.58%) and intermediate haplotypes frequencies (10.9%-31.5%) of
196 H1.β2+. In contrast, South Asian populations, which have been historically underrepresented in
197 sequencing datasets, exhibit the highest worldwide frequency of H1.β2+ haplotypes (14.6%-
198 66.8%) and intermediate H1.β1 haplotype frequencies (28.7%-79.2%) with lower levels of
199 H2.α2+ (1.4%-10.9%). East Asian populations are essentially fixed for H1.β1 haplotypes. We
200 also assessed whole genome sequencing data from 368 African genomes from the CAAPA2
201 consortium demonstrating limited haplotype diversity in Africa, with most populations exhibiting

202 H1.β1. However, we observed H2.α1 haplotypes at low frequency in Khomani San, Nama,
203 Mbuti, Biaka, Bench, and Sheko populations. H2.α2+ haplotypes were also present at some low
204 frequencies in some African populations. These results confirm previous analyses highlighting
205 the presence of rare H2.α1 haplotypes almost exclusively in African populations [19].
206
207 While short read sequencing data do not enable characterization of phased *NSF* duplications
208 (i.e. γ2+) we calculated *NSF* copy number genotype frequencies worldwide confirming the
209 highest *NSF* copy numbers in East Asia (**Fig 4E**). We find that the overwhelming majority of
210 *NSF* duplications occur on the H1.β1 background. However, we do observe evidence of both
211 *NSF* duplications and deletions happening at low frequencies on H2.α2 background (**Fig 4F**).
212 Together, these results provide the most comprehensive catalog of 17q21.31 diversity in
213 humans to date and highlight the extreme stratification of haplotypes between continental
214 populations.

215 **Double recombinants and complex gene conversion events between structural**
216 **haplotypes**

217 The suppression of recombination by inversions can disrupt patterns of linkage disequilibrium
218 (LD) across the genome. To understand how the 17q21.31 inversion has impacted patterns of
219 recombination we used SNPRelate, a windowed PCA-based analysis, to cluster genotypes
220 across 3135 individuals (**Fig 5A, B**). Individuals clustered into three groups corresponding to
221 H1/H1, H1/H2, H2/H2 genotypes with LD extending ~80kb beyond the proximal inversion
222 breakpoint. Despite the persistence of this long 900kb LD block, we identified three instances of
223 recombination at the proximal end of the extended LD block (**Fig 5C**) occurring primarily in
224 African individuals (32/56, 57%). We also noted several individuals exhibiting 20-180 kb
225 stretches showing switching between genotype groups within the inversion region, indicating
226 double crossover recombination events, or possibly, long stretches of interlocus gene
227 conversion (**Fig 5A-C**). Of the 8 identified switch such events, two exceeded 140 kb, occurring
228 in a single European individual and a single South Asian individual, and likely represent cases
229 of double crossover recombination. The remaining events range from 20-70 kb and were found
230 almost exclusively in African populations (51/54, 94%). We also noted that the windowed PCA-
231 analyses identified more complex clustering patterns at the proximal and distal ends of the
232 inversion locus (**Fig 5B**). We found that these corresponded strongly with the complex structural
233 haplotypes which we describe thus demonstrating that complex structural variation can be
234 detected from short read-based SNP patterns in some cases. Together these results
235 demonstrate that despite extensive inversion-driven recombination suppression, substantial
236 exchange has happened between vastly divergent haplotypes.

237 **Ancient genomes track the rise in frequency of *KANSL1* duplications in Europe**

238 To determine how 17q21.31 structural haplotype frequencies have changed in Europeans over
239 recent history we analyzed 626 ancient genomes spanning ~10 thousand years of human
240 history[21] **(Fig 6A)**. European Hunter Gatherer populations, dating from ~5000-10,000 years
241 before present (BP) exhibited extremely low haplotype diversity with H1.β1 making up >95% of
242 haplotypes **(Fig 6B)**. Higher frequencies of H2.α2+ haplotypes appear ~7 to ~5 kya BP in Early
243 and Neolithic Farmers (25%-31%) . Steppe Pastoralists also exhibit H2.α2+ haplotypes (10%)

as well as high frequencies of the H1.β2+ haplotype (30%). To understand the temporal dynamics of haplotype frequency changes over the last 10 kya we fit a multinomial logistic regression model to these ancient genomic derived data **(Fig 6C)**. This model demonstrates that while H1.β1 was the dominant haplotype ~11 thousand years ago with a frequency of >90%, it has subsequently declined ~1.6-fold. Concomitantly, the H1.β2+ and H2.α2+ haplotypes have increased from ~1% and ~7% to 29% and 21% respectively. Together, these data demonstrate that 17q21.31 haplotype diversity in Europe today is the result of a massive shift in frequencies of the *KANSL1* duplication in both H1 and H2 haplotypes over recent history.

DISCUSSION

Here we leverage long-read haplotype-resolved human genomes, newly assembled T2T and population-scale long-read sequenced great ape genomes, diverse human short-read data, and ancient DNA to assess the diversity and recent evolution of the 17q21.31 locus. Together, our results provide the most comprehensive assessment to date of this remarkable region of the genome providing new insights and clarifying previous works. For instance, while it has been established that the H2 haplotype is the ancestral orientation of the 17q21.31 locus, the presence of the inversions in chimpanzees, bonobos, and orangutans has been interpreted as recurrent “toggling” of the inversion across great apes. Here, by using population-scale haplotype resolved assemblies from chimpanzees we resolve the structure and sequence of the chimpanzee 17q21.31 inversion. We find that the chimpanzee inversion is substantially larger than the human inversion.While it overlaps the human inversion across ~750kb of sequence it extends an additional ~650kb. Furthermore, we find that this inversion occurred far more recently than the human inversion, ~330,000 years ago, in contrast to previous estimates[16]. Thus, “highly mutable and recurrent” might be a more apt description of the inversion locus than “toggling” as independent inversion events with distinct breakpoints have occurred on different lineages. However, additional population-scale long-read sequencing of primate genomes will be required to determine the sequence and structures of the bonobo and orangutan inversions and discern their evolutionary history.

The H2 haplotype has been associated with increased fecundity in Icelandic[14] and Polish[22] populations and has thus been suggested to be the target of selection in Europeans. Indeed, H2 haplotype frequencies are highest in European populations. We used ancient genomes to reconstruct allele frequency trajectories of 17q21.31 haplotypes finding that while H1.β1 haplotypes have declined in frequency markedly over the past 10,000 years, H1.β2+ and H2.α2+ haplotypes have increased in frequency substantially: ~29-fold for H1.β2+ and 3-fold for H2.α2+. Thus, the group of haplotypes containing duplications of the *KANSL1* gene has increased in frequency ~6-fold. Prima facie such an increase in frequency would correspond to a selection coefficient of approximately 0.015, similar to the average selection coefficients for regions such as lactase persistence. Critically, however, this estimate is an extremely crude oversimplification which neglects the small sample sizes in the ancient data and the importance of modelling the complex demographic history of Europe. Nevertheless, these results provide an intriguing alternate hypothesis: that duplications of the *KANSL1* gene may be the targets of selection as opposed to the inversion itself.

287 The availability of increasingly diverse population-scale short-read sequencing of humans
288 enabled us to characterize 17q21.31 haplotype diversity worldwide. Our results confirm previous
289 work highlighting the prevalence of the H1.β1 haplotype in Africa with the ancestral H2.α1
290 haplotypes present at low frequencies in some populations such as the Khomani San. However,
291 our analyses also provide some of the first observations of 17q21.31 diversity in South Asian
292 populations, which have historically been undersampled. We find surprisingly high frequencies
293 of the H1.β2+ and H2.α2+ haplotypes across South Asia. These results mirror those of Europe
294 with the two *KANSL1* duplication-containing haplotypes showing increased frequencies
295 compared to African and East Asian populations, albeit with differing relative frequencies of the
296 H1.β2+ and H2.α2+. These results potentially suggest that a recent increase in duplicated
297 haplotypes may have also occurred in South Asia mirroring the increase in Europe, though
298 additional analyses are needed to confirm that hypothesis.
299
300 Finally, we discovered several double recombination events inside the 17q21.31 inversion
301 locus. Intriguingly, these events are overwhelmingly found in African populations, many of which
302 have very low H2 haplotype frequencies. Therefore, these double recombination events are
303 likely quite old, having occurred when H2 haplotypes were at higher frequencies in Africa. We
304 also identified two other much longer double recombination events present in two non-African
305 individuals. In addition, we identified a region intersecting the *CRHR1* gene where H1 and H2
306 haplotypes are completely homogenized and individuals cannot be assigned to one haplotype
307 or the other (**Fig 5A-C**). This finding has previously been reported and explained as the result of
308 an ancient double recombination event[19]. Crucially, unlike the other double recombinants that
309 we identified, this event has likely been fixed in H2 haplotype, and the original sequence
310 associated with the H2 haplotype has been lost in all human populations. Except for these rare
311 double recombination events, however, we found that recombination is completely suppressed
312 across the inversion. An additional consequence of segregating inversions is that by
313 suppressing recombination they can promote the accumulation of deleterious variation[9]. It
314 remains an unanswered question how the increased frequency of 17q21.31 inversion in specific
315 human populations may affect mutational load.

316 The importance of the 17q21.31 locus in human disease, human diversity, and evolution has
317 been the focus of many studies since its discovery. Nevertheless, long read sequencing, ancient
318 genomes, and additional human population datasets are continuing to increase our
319 understanding of this region. Our insights synthesize previous work with these new advances in
320 genetics and genomics highlighting both new insights and posing new avenues for future
321 research.

322 METHODS

323 Datasets

324 Short-read sequencing data were compiled from high-coverage resequencing of 5174
325 individuals from the 1,000 Genomes Project (1KG), the Human Genome Diversity Panel
326 (HGDP) and the Simons Genome Diversity Panel (SGDP). We also obtained unpublished short-
327 read sequencing data of 318 individuals from the CAAPA2 Data consortium. In total we were

able to infer the complex genotypes of 4585 individuals with high confidence from these contemporary samples. Among these 697 are trios from 1KG and the remainder 3888 are unrelated individual samples compiled from 1KG, HGDP, SGDP, and CAAPA2. The trios are only used for method validation purposes. Genotypes from 626 historical samples from between -12kya and 1000 ya were obtained from [23]. Historical samples were assigned to populations based on their original publications [5,21]. 216 long-read haplotype assemblies were compiled from the HPRC and HGSVC [3]. In addition we included the GRCh38 genome and CHM13 reference genomes to our dataset[24]. Haplotype assemblies were only included if they spanned the 17q21.31 locus. This excluded three assemblies `HG0015#hap1, HG00320#hap2, and NA21093#hap1` from our dataset, although they were used in characterization of complex structures in the locus. These three assemblies that were excluded appear to have been misassembled across the locus as they were either discontiguous, or had a partial *KANSL1* duplication that could not be cross-validated with short-read sequencing data. We obtained primate genome assemblies for orangutans, gorillas, bonobos, and chimpanzees from three sources: GenomeArk (eight haplotypes, two per species), and 57 newly generated and unpublished gorilla, chimpanzee, and bonobo genomes. For the purposes of this study, we include the unpublished assemblies for the 17q21.31 region, with full genome releases planned alongside separate publications currently in preparation.

Characterization of complex structures at the 17q21.31 locus in humans

We extracted the 17q21.31 region including flanking sequences from the GRCh38 reference genome (chr17: 45275488-46921902). We then used minimap2[25] (-x asm5) to identify the homologous sequences in all the samples in the HGSVC and HPRC genomes. We performed a principal bundle decomposition using PanGenomic Research Tool Kit (PGR-TK) for 216 of the HGSVC and HPRC genome assemblies with the parameters of `-w 80 -r 12 --min-span 32 --bundle-length-cutoff 10000 --min-branch-size 16 --min-cov 0`[26]. We examined the results, and excluded five assemblies that were potentially misassembled. We further excluded bundles that were unique to just one haplotype, or those whose length were in the lowest quantile across all the remaining bundles (i.e. < 23kb). We grouped the haplotypes into representative structures based on the order and orientation of the principal bundles. The position of genes along haplotypes was determined by mapping coding sequences to haplotypes using minimap2[25] (`-c -x asm5`). We chose one haplotype within each unique structure and used PGGB to build a pangenome graph[27]. We then used odgi similarity to generate a matrix of jaccard distance between each unique haplotype structure and made a neighbor joining tree to cluster these haplotype structures. We ran minimap2[25] (`-x asm5 -c --eqx -D -P --dual=no`) on pairs of assemblies to create sequence alignments, which were then inputted to SVByEye[28]for visualization.

Characterization of complex structures at the 17q21.31 locus in primates

We used minimap2[25] (-x asm20) to identify the orthologous sequences in the great ape assemblies to the human 17q21.31 locus in GRCh38 reference genome (chr17:45275488-46921902). Because we observed an inversion in chimpanzees that spans a longer region than the inversion in humans, we then manually extended these sequences by 2Mb on either flank in both great apes and all human haplotypes, so that these sequences span the entirety of the

370 chimpanzee inversion. We then reran PGRTK on these extended human sequences with the
371 same settings and projected the principal bundles to the great ape sequences. Within each
372 species, we categorized these haplotypes by their inversion, *KANSL1* and *NSF* duplication
373 status. We then selected the one representative haplotype for each unique structure in each
374 species for visualization. We ran minimap2[25] and SVByEye[28] on pairs of great ape and human
375 haplotypes with the same settings except -x asm20 was used as a flag in minimap2[25] instead of
376 -asm5. To estimate the divergence time of structural haplotypes across great apes, we first
377 identified the largest region of shared homology within the inverted segment (corresponding to
378 the blue bundle in Fig. 1A). We performed pairwise alignments using minimap2[25], with the
379 CHM13 human reference sequence as the query, against 17 primate assemblies: all great ape
380 haplotypes from GenomeArk (including 4 orangutan, 2 gorilla, 2 bonobo, and 2 chimpanzee
381 haplotypes), 4 additional gorilla haplotypes, the inverted chimpanzee haplotype, and
382 representative human haplotypes H1.β1.γ1, H1.β2.γ1, and H2.α2.γ2. Multiple sequence
383 alignment was performed using kalign[29]. Phylogenetic trees were inferred using IQ-TREE with
384 the Jukes-Cantor substitution model and 1000 bootstrap replicates Divergence times were
385 estimated using the orangutan–human/chimpanzee/gorilla split `18.13` mya as a calibration
386 point[30]. Trees were visualized using the ggtree R package[31].

387 **Evolutionary history of structural variation at the 17q21.31 locus in humans**

388 To date the emergence of different H1 haplotype structures we first needed to find a unique
389 sequence at the 17q21.31 locus that is physically linked to the *KANSL1* and *NSF* duplications
390 and not broken down by recombination. The largest region of shared homology within the
391 inverted segment (corresponding to the blue bundle in Fig. 1A) in CHM13 is also the closest in
392 proximity to the *KANSL1* duplications. Therefore, we used similarities in phylogenetic tree
393 topology as a proxy for linkage disequilibrium across this region.This region was divided into
394 nine non-overlapping 50 kb windows using seqkit `(sliding -s 50000 -W 50000)`[32]. All 210
395 human haplotypes that were included in previous analyses were included to construct the trees,
396 as well as 2 GenomeArk chimpanzee haplotypes. For each of the nine windows, sequences
397 were aligned using kalign[29], and maximum likelihood trees were inferred with IQ-TREE[33] using
398 the Jukes–Cantor model and 1000 ultrafast bootstrap replicates `(-m JC -bb 1000 -nt AUTO)`.
399 Trees were rooted using the two chimpanzee haplotypes as an outgroup. To quantify
400 topological similarity between trees, we computed pairwise Robinson–Foulds (RF) distances.
401 An RF distance matrix was then constructed for all pairwise tree comparisons, and the windows
402 with the lowest distance to the right-most window (i.e. the window that is closest to the *KANSL1*
403 duplication) were selected. We then merged the selected windows, and reran IQ-TREE using
404 the chimpanzee-human split time `7.74` mya as a calibration point for the emergence of
405 haplotype structures[33].

406 **Using contemporary short read genomes to understand 17q21.31 population diversity in**
407 **modern human populations**

408 *Tag SNP based inversion genotyping*

409 In order to determine the genotype status of individuals in our dataset, we used a total of 1271
410 tag SNPs to genotype the 17q21.31 locus based on the method used in[23] For each individual,

411 we constructed a genotype vector by coding observed genotypes at these SNPs as 0
412 (homozygous reference), 1 (heterozygous), or 2 (homozygous alternate). Then, Euclidean
413 distance was calculated between the observed genotype vector and expected genotype vectors
414 for H1/H1, H1/H2, and H2/H2. Each individual was assigned the inversion genotype with the
415 minimum Euclidean distance to their observed genotype.

416 *Read-depth based copy number genotyping at NSF* and *KANSL1*

417 Copy number genotypes for the 1000 Genomes, HGDP, and SGDP, and CAAPA2 individuals
418 were estimated using read depth as described in [5]. In brief, read depths were quantified from
419 CRAM (CAAPA2) or BAM (1000G, HGDP, SGDP) files. CRAM files from CAAPA2 were based
420 on a different version of the GRCh38 reference genome, and to ensure consistency across
421 datasets, we remapped this data to the same version of the reference genome as the other
422 three datasets (GRCh38_full_analysis_set_plus_decoy_hla). We generated d4 files from these
423 BAM files. From the d4 files we calculated the average read depths in 1,000 bp sliding windows
424 in 100 bp steps across the 17q21.31 locus. These depths were normalized to the average read
425 depth within a control region chr17:42800000-46000000 within the 17q21.31 locus for each
426 individual, in which no copy number variation was observed in more than 3,000 individuals (i.e.
427 individuals from the 1KGP and HGDP samples). To determine the copy number of the α and β
428 segments of the *KANSL1* duplication, we averaged the copy number of windows that
429 overlapped with the unique region of each duplication (α: chr17:46143000-46238000 and β:
430 chr17:46095000-46123000). In the case of *NSF* copy number, since the GRCh38 reference
431 genome has two copies of *NSF*, we summed the normalized read depth across two target
432 regions corresponding to the two *NSF* copies (chr17:46,336,376–46,489,410 and
433 chr17:46,564,311–46,707,123). CAAPA2 samples were excluded from *NSF* analysis because of
434 difficulties in aligning the sequencing reads to this complex region. Lastly, we removed samples
435 whose copy number estimates were too large to be realistic (i.e. copy number > 20 for *KANSL1*
436 and > 50 for *NSF*).

437 *Determining haplotypes from 17q21.31 genotypes*

438 We were able to associate the β and α duplications with the inversion status of their haplotype
439 background because the β duplication is always associated with the H1 haplotypes and that the
440 α duplication is always associated with the H2 haplotypes. Therefore, it was possible to
441 integrate the inversion genotypes (H1/H1, H1/H2, H2/H2) and copy numbers (based on read
442 depth in sector calls) to assign complex genotypes representing both inversion and *KANSL1*
443 duplication status as described in[23] **(Fig. S4)**. We then used R to plot the average copy number
444 across the β and α regions, and assigned genotypes based on their inversion status and copy
445 number. The samples clustered into duplication genotypes with the exception of a handful of
446 samples which did not match their inversion calls. In cases where multiple interpretations of
447 duplication status are possible, the more probable assignment is listed (e.g., H1D/H1D vs.
448 H1/H1DD we chose the former because H1.B3 is very rare). Since there is more diversity in
449 *NSF* copy numbers compared to *KANSL1* especially on the H1.β1 background, we were unable
450 to phase *NSF* duplications and only provide diploid copy number estimates. Complex genotypes
451 for the historical short-read genomes were obtained from Irving-Pease[23].

452 *Sliding window PCAs reveal regions of double recombination*

453 In order to determine recombination or gene conversion events within the inversion locus and
454 extended LD of 100kb, PC1 values were computed in non-overlapping 10kb windows for all
455 individuals using SNPRelate[34]. Tag SNP based inversion genotyping revealed three clusters:
456 H1 homozygous (bottom), H2 homozygous (top), and H1/H2 heterozygous (middle). We
457 computed the mean, median, and multiple quantile ranges of PC1 for individuals of each
458 inversion genotype in every window. To identify noisy or ambiguous positions, we evaluated
459 whether the variance for different genotypes overlapped at each position. Positions with
460 overlapping variance were flagged and excluded from downstream interpretation. The region of
461 interest was further refined to non-duplicated regions except for the unique *KANSL1*. To refine
462 outlier classification, we focused on runs of at least two adjacent outlier windows within the
463 same sample. For these regions, we reassigned the genotype based on directional deviation of
464 PC1: H1H2 genotypes with PC1 below the genotype's 5th percentile were reassigned to H1H1,
465 while those above the 95th percentile were reassigned to H2H2. Homozygous genotypes were
466 reassigned to H1H2 if their PC1 values deviated substantially in either the positive or negative
467 direction respectively. Further manual curation allowed us to determine single recombinant
468 events and double recombinant events (**Table S4**).
469

470 **REFERENCES**

471 1. Porubsky, D. *et al.* Recurrent inversion polymorphisms in humans associate with genetic
472 instability and genomic disorders. *Cell* **185**, 1986–2005.e26 (2022).
473 2. Plender, E. G. *et al.* Structural and genetic diversity in the secreted mucins MUC5AC and
474 MUC5B. *Am. J. Hum. Genet.* **111**, 1700–1716 (2024).
475 3. Logsdon, G. A. *et al.* Complex genetic variation in nearly complete human genomes.
476 *bioRxiv* (2024) doi:10.1101/2024.09.24.614721.
477 4. Scheer, K. *et al.* Adaptive increase of amylase gene copy number in Peruvians driven by
478 potato-rich diets. *bioRxiv* (2025) doi:10.1101/2025.03.25.644684.
479 5. Bolognini, D. *et al.* Recurrent evolution and selection shape structural diversity at the
480 amylase locus. *Nature* **634**, 617–625 (2024).
481 6. Ebert, P. *et al.* Haplotype-resolved diverse human genomes and integrated analysis of
482 structural variation. *Science* **372**, (2021).
483 7. Thomas, J. W. *et al.* The chromosomal polymorphism linked to variation in social behavior
484 in the white-throated sparrow (Zonotrichia albicollis) is a complex rearrangement and
485 suppressor of recombination. *Genetics* **179**, 1455–1468 (2008).
486 8. Hager, E. R. *et al.* A chromosomal inversion contributes to divergence in multiple traits
487 between deer mouse ecotypes. *Science* **377**, 399–405 (2022).
488 9. Jay, P. *et al.* Mutation load at a mimicry supergene sheds new light on the evolution of
489 inversion polymorphisms. *Nat. Genet.* **53**, 288–293 (2021).
490 10. Wang, H. *et al.* Chromosomal inversion polymorphisms shape human brain morphology.
491 *Cell Rep.* **42**, 112896 (2023).
492 11. Le Moan, A. *et al.* Coupling of twelve putative chromosomal inversions maintains a strong
493 barrier to gene flow between snail ecotypes. *Evol. Lett.* **8**, 575–586 (2024).
494 12. Harringmeyer, O. S. & Hoekstra, H. E. Chromosomal inversion polymorphisms shape the
495 genomic landscape of deer mice. *Nat. Ecol. Evol.* **6**, 1965–1979 (2022).
496 13. Jamsandekar, M. *et al.* The origin and maintenance of supergenes contributing to
497 ecological adaptation in Atlantic herring. *Nat. Commun.* **15**, 9136 (2024).

14. Stefansson, H. *et al.* A common inversion under selection in Europeans. *Nat. Genet.* **37**, 129–137 (2005).
15. Koolen, D. A. *et al.* Mutations in the chromatin modifier gene KANSL1 cause the 17q21.31 microdeletion syndrome. *Nat Genet* **44**, 639–641 (2012).
16. Zody, M. C. *et al.* Evolutionary toggling of the MAPT 17q21.31 inversion region. *Nat Genet* **40**, 1076–1083 (2008).
17. Pedicone, C., Weitzman, S. A., Renton, A. E. & Goate, A. M. Unraveling the complex role of MAPT-containing H1 and H2 haplotypes in neurodegenerative diseases. *Mol. Neurodegener.* **19**, 43 (2024).
18. Karamik, G. *et al.* The clinical phenotype of Koolen-de Vries syndrome in Turkish patients and literature review. *Am. J. Med. Genet. A* **191**, 1814–1825 (2023).
19. Steinberg, K. M. *et al.* Structural diversity and African origin of the 17q21.31 inversion polymorphism. *Nat Genet* **44**, 872–880 (2012).
20. Boettger, L. M., Handsaker, R. E., Zody, M. C. & McCarroll, S. A. Structural haplotypes and recent evolution of the human 17q21.31 region. *Nat Genet* **44**, 881–885 (2012).
21. Allentoft, M. E. *et al.* Population genomics of post-glacial western Eurasia. *Nature* **625**, 301–311 (2024).
22. Clark, J. S. C. *et al.* Association studies between chromosomal regions 1q21.3, 5q21.3, 14q21.2 and 17q21.31 and numbers of children in Poland. *Sci Rep* **12**, 18923 (2022).
23. Irving-Pease, E. K. *et al.* The selection landscape and genetic legacy of ancient Eurasians. *Nature* **625**, 312–320 (2024).
24. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
25. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
26. Chin, C.-S. *et al.* Multiscale analysis of pangenomes enables improved representation of genomic diversity for repetitive and clinically relevant genes. *Nat. Methods* **20**, 1213–1221 (2023).
27. Garrison, E. *et al.* Building pangenome graphs. *Nat. Methods* **21**, 2008–2012 (2024).
28. Porubsky, D. *et al.* SVbyEye: a visual tool to characterize structural variation among whole-genome assemblies. *Bioinformatics* **41**, (2025).
29. Lassmann, T. Kalign 3: multiple sequence alignment of large data sets. *Bioinformatics* **36**, 1928–1929 (2019).
30. Yoo, D. *et al.* Complete sequencing of ape genomes. *Nature* **641**, 401–418 (2025).
31. Guangchuang Yu [aut, cre, cph], Tommy Tsan-Yuk Lam [aut, ths], Justin Silverman [ctb]. *Ggtree*. (Bioconductor, 2017). doi:10.18129/B9.BIOC.GGTREE.
32. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS One* **11**, e0163962 (2016).
33. Minh, B. Q. *et al.* IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
34. Zheng, X. *et al.* A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328 (2012).

DECLARATION OF INTERESTS

The authors declare no competing interests.

ACKNOWLEDGEMENTS

We thank R Nielsen, BM Henn, P Moorjani, CT Miller, JM Vazquez, R De Kayne for helpful discussion. This work was supported by NIH National Institute of General Medicine award

R35GM142916 to PHS and NIH National Human Genome Research Institute award R01HG013017 to PHS and a Leakey Foundation award to SS.

AUTHOR CONTRIBUTIONS

PHS conceived of the study. SS, RNL, SF, JLR, and PHS performed all analyses and wrote the manuscript. MWM and ANK provided primate cell lines and tissue culture. VB, SGMM, SG, and BH provided data.

DATA AND CODE AVAILABILITY

All data used in this project are publicly available and described in the 'Datasets' section of the Methods and at the following github link https://github.com/sudmantlab/17q21.31_SV_and_evolution which is archived in Zenodo: https://doi.org/10.5281/zenodo.16883463 (Sridharan 2025). Copy numbers, genotypes, and structural haplotypes can be found in the Supplementary tables. The HPRC data can be obtained at: https://humanpangenome.org/data/. The HGSVC data can be obtained at: https://www.hgsvc.org/resources, The 1KG and HGDP data can be obtained at: https://www.internationalgenome.org/data/. The SGDP data can be obtained at: https://www.simonsfoundation.org/simons-genome-diversity-project/. The ancient data are available on the European Nucleotide Archive under the accession codes PRJEB64656 and PRJEB50857. The unpublished data for bonobos, chimpanzees and gorillas can be obtained at: https://github.com/sudmantlab/17q21.31_SV_and_evolution. The GenomeArk long read haplotype resolved assemblies for primates can be obtained at: https://www.genomeark.org/

FIGURE TITLES AND LEGENDS

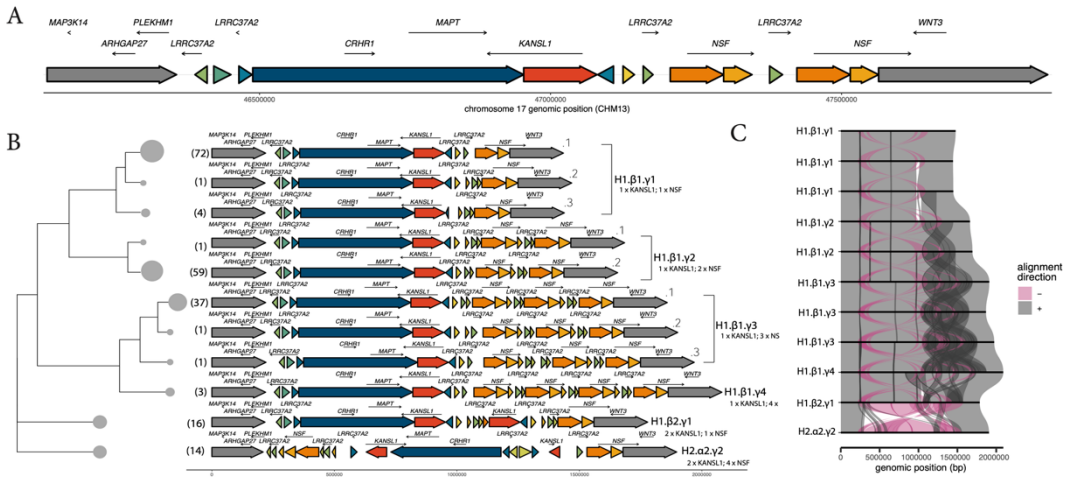


Figure 1 – Distinct 17q21.31 structural haplotypes from human long read genome assemblies: a) Schematic of haplotype structure at the 17q21.31 locus and flanking regions using PGRTK for the reference CHM13 genome. Large colored arrows represent principal bundles, which are individual units of repeats. Bundles of the same color represent homologous sequences. The directionality of the arrow represents the bundle orientation. Unique flanking regions are indicated in greyscale. Smaller labeled arrows denote genes and gene orientations. **b)** Schematic of 11 unique structural haplotypes found within 210 long-read haplotype-resolved assemblies. Each row represents a unique structural haplotype. Structural haplotypes are arranged using hierarchical clustering based on jaccard distance, with tip sizes scaled by the number of assemblies sharing each haplotype. **c)** Structural diversity of 11 unique structural haplotypes visualized using SVbyEye. Alignments are colored by orientation: grey denotes alignments between two structural haplotypes in the forward orientation and pink denotes alignment in the reverse direction. Each row represents a unique haplotype structure arranged in the same order as in **b**.

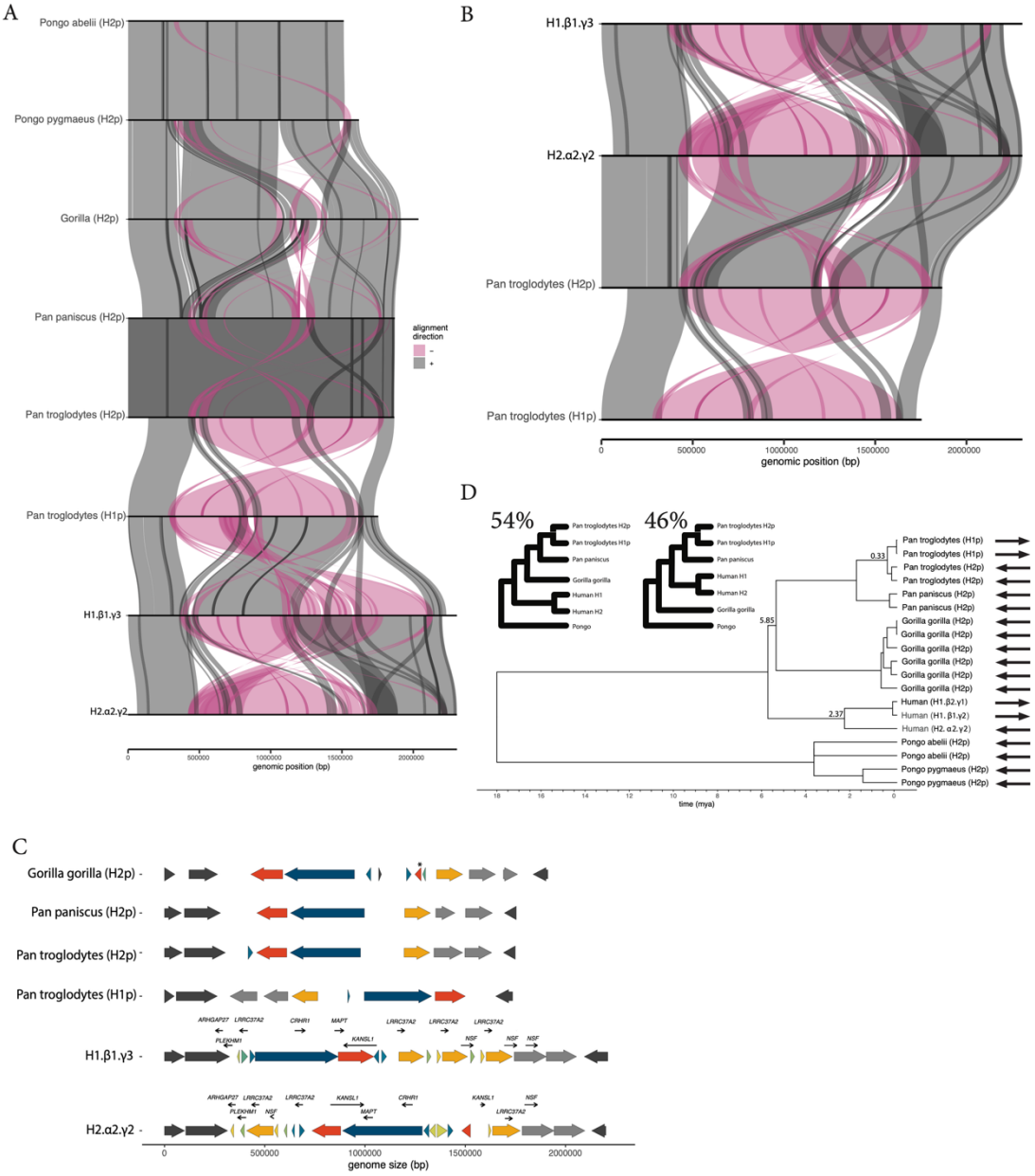


Figure 2 – Characterizing the evolutionary context of the 17q21.31 inversion in the great ape lineage: a) Visualization of sequence alignments spanning the 17q21.31 locus and flanking regions in great apes. The sequences are extended from **Fig 1** by 2 Mb on both ends to characterize the shared homology of flanking regions across all species. Each row represents a representative haplotype from each of the primates. Alignments are colored by orientation (gray = forward, pink = reverse). **b)** Alignment visualization comparing representative human haplotypes H1.β1.γ3 and H2.α2.γ2 and chimpanzee haplotypes H1p and H2p (as seen in **a**), reordered to highlight the distinct, and longer inversion found in chimpanzees. **c)** Unique structural haplotypes in great apes with two representative human structural haplotypes shown as reference. Human haplotypes are plotted along with gene labels and gene orientation. Sequences are visualized with PGRTK trained on human structures with extended sequence as

596 seen in **a**. Unique flanking regions are shown in dark grey and lighter grey. Colors are broadly
597 consistent with **Fig 1b** although the longer genetic interval used here results in slight
598 differences. **d)** Dating the divergence of primate and human 17q21.31 haplotypes. Trees were
599 inferred using a large non-duplicated sequence located within the human inversion (i.e. the blue
600 bundle in **c**, **see methods)**. Several important split times are annotated in the tree. The inset
601 shows two main topologies that are observed via bootstrapping and their frequencies.
602

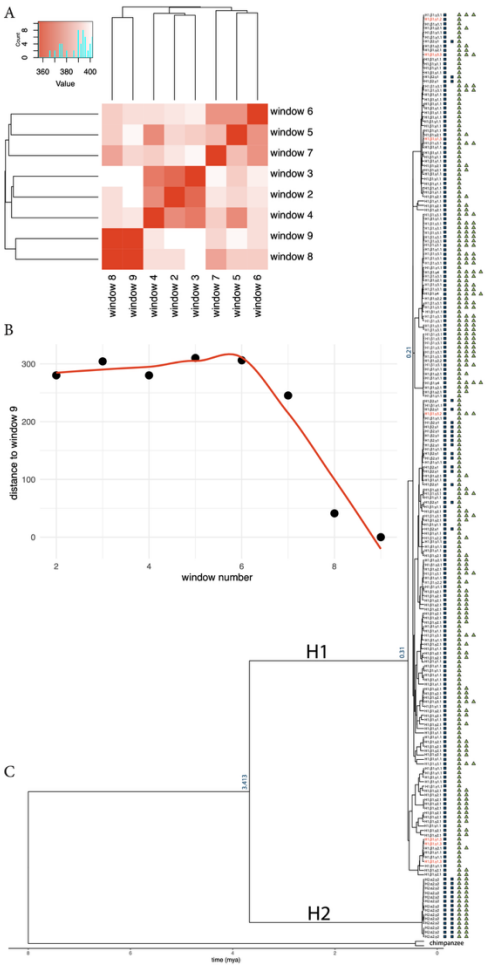


Figure 3 – Timing and recurrence of human 17q21.31 structural haplotypes: a) Robinson-Foulds (RF) distance clustering heatmap across phylogenetic trees constructed with non-overapping 50 kb windows across all human long-read assemblies. Windows are extracted from a large non-duplicated sequence located within the human inversion (i.e. the blue bundle in **Fig 1)**. Darker red indicates greater similarity (lower RF distance) between two trees. Dendogram summarizes the topological similarity across all windows. Inset histogram displays the distribution of RF distance values across all comparisons. **b)** Euclidean distance between the phylogenetic tree constructed at window 9, the closest unique sequence to *KANSL1*, and trees at all other windows. A loess-smoothed curve is shown in red. **c)** Dated maximum-likelihood tree constructed using sequences at windows 8 and 9 and rooted with T2T chimpanzee haplotypes as the outgroup. Human-chimpanzee split-time was used to date the tree. Select internal nodes are labeled with divergence time. Tip symbols indicate copy-number variation of *KANSL1* (blue squares) and *NSF* (green triangles).

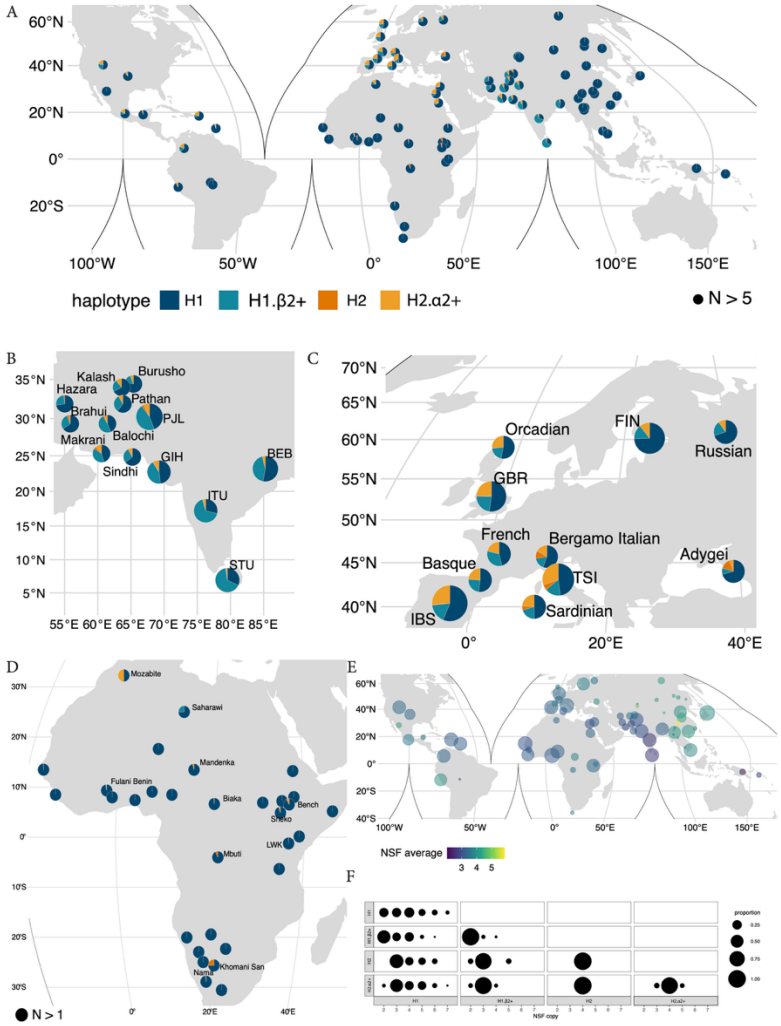


Figure 4 – Population diversity of the 17q21.31 locus worldwide using short read data: a) World map indicating the frequency of direct H1.β1, H1.β2+ and inverted H2.α1 and H2.α2+ 17q21.31 haplotypes in 4585 individuals from 193 populations. Insets show the frequency distribution in South Asia (**b)**, Europe (**c**) and Africa (**d**). Pies are scaled by sample size in **b** and **c**. Only populations with five or more samples are plotted in **a-c**, whereas **d** shows populations with as few as a single sample. In **d**, only populations with H1.β1 frequencies less than 1 are labeled. See **Supplementary Table 2** for exact sample sizes and haplotype frequencies of all populations. **d)** World map indicating the average diploid *NSF* copy number in 3352 individuals from 147 populations. Points are scaled based on sample size, with only populations with five or more samples plotted. See **Supplementary Table 3** for the exact sample sizes and average diploid copy numbers of all populations. **e)** The distribution of *NSF* copy numbers in individuals with different inversion and *KANSL1* duplication genotypes.

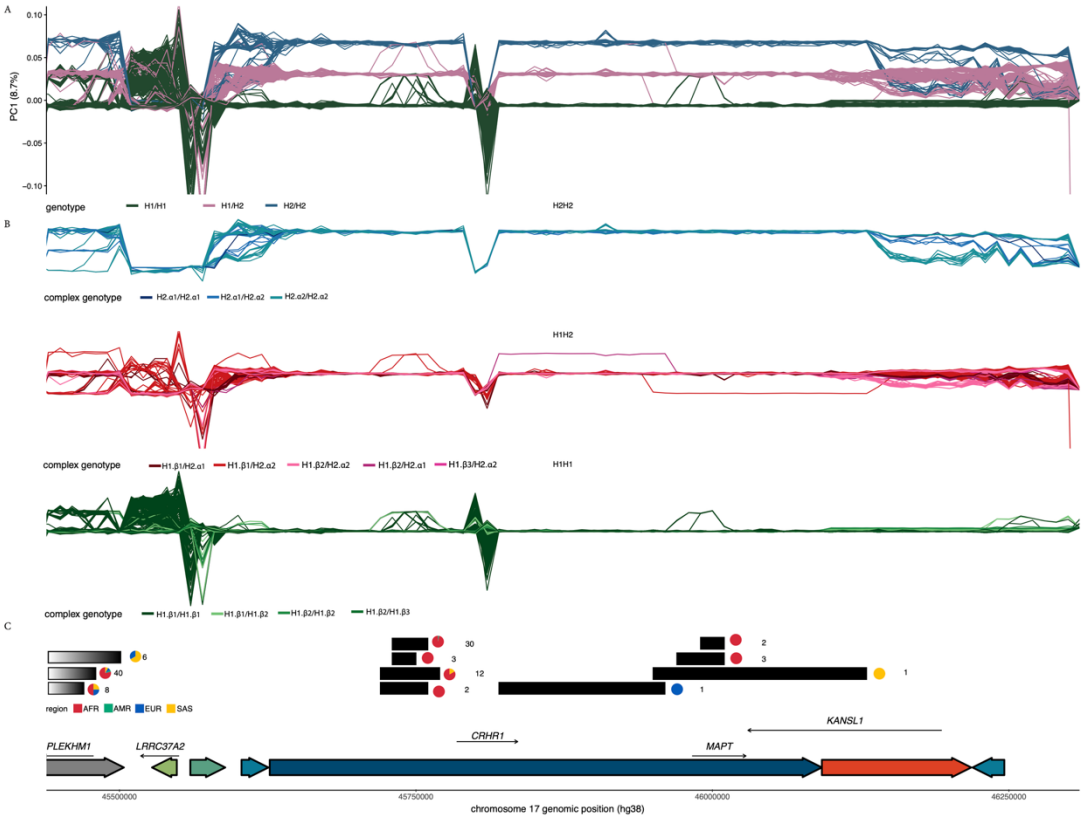


Figure 5 – Population specific instances of recombination events at the 17q21.31 locus: The x-axis represents genomic position across a subset of the 17q21.31 locus including the inversion region and surrounding flanking sequences. A PGRTK-based haplotype structure of the reference genome (hg38) is shown below the figure. **a)** 10kb sliding window PCA. Each colored line represents the variation in the position of an individual on the first principal component axis, and is colored by inversion status. **b)** 10kb sliding window PCA faceted by inversion status and colored by complex genotype. **c)** All recombination events identified in the region of interest. Gradient blocks represent single recombination events wherein the right edge of the block represents the recombination breakpoint. Black boxes represent double recombination events. Numbers next to each block indicate the number of individuals with that recombination event, and pie charts indicate the population of these individuals.

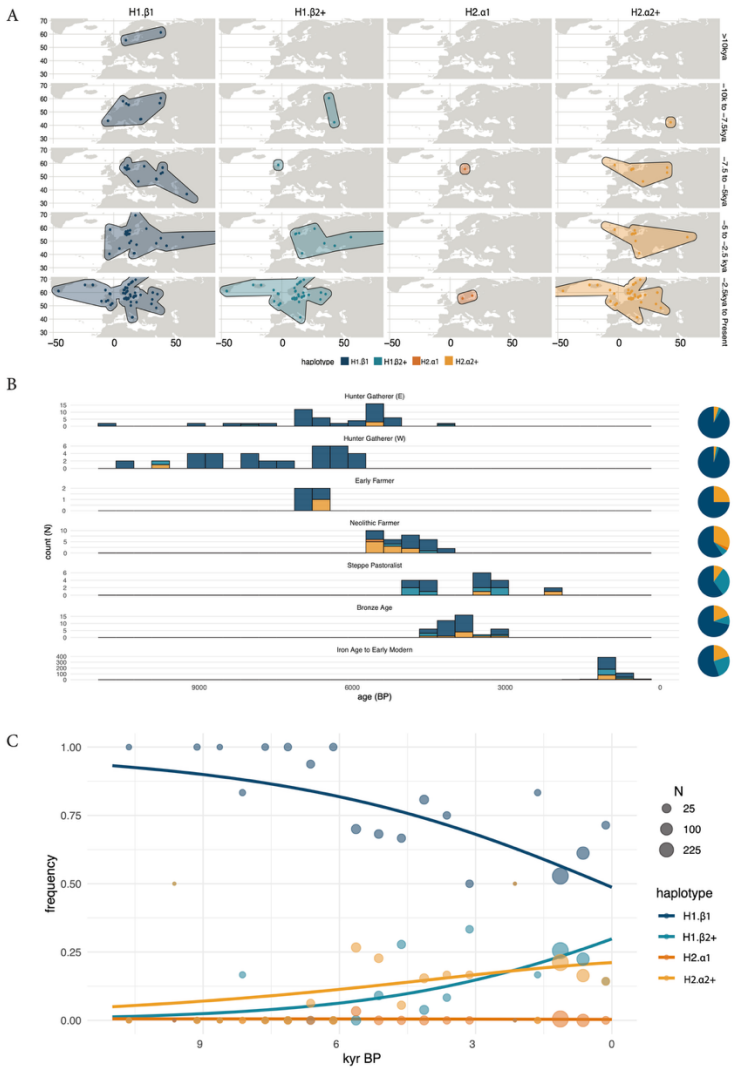
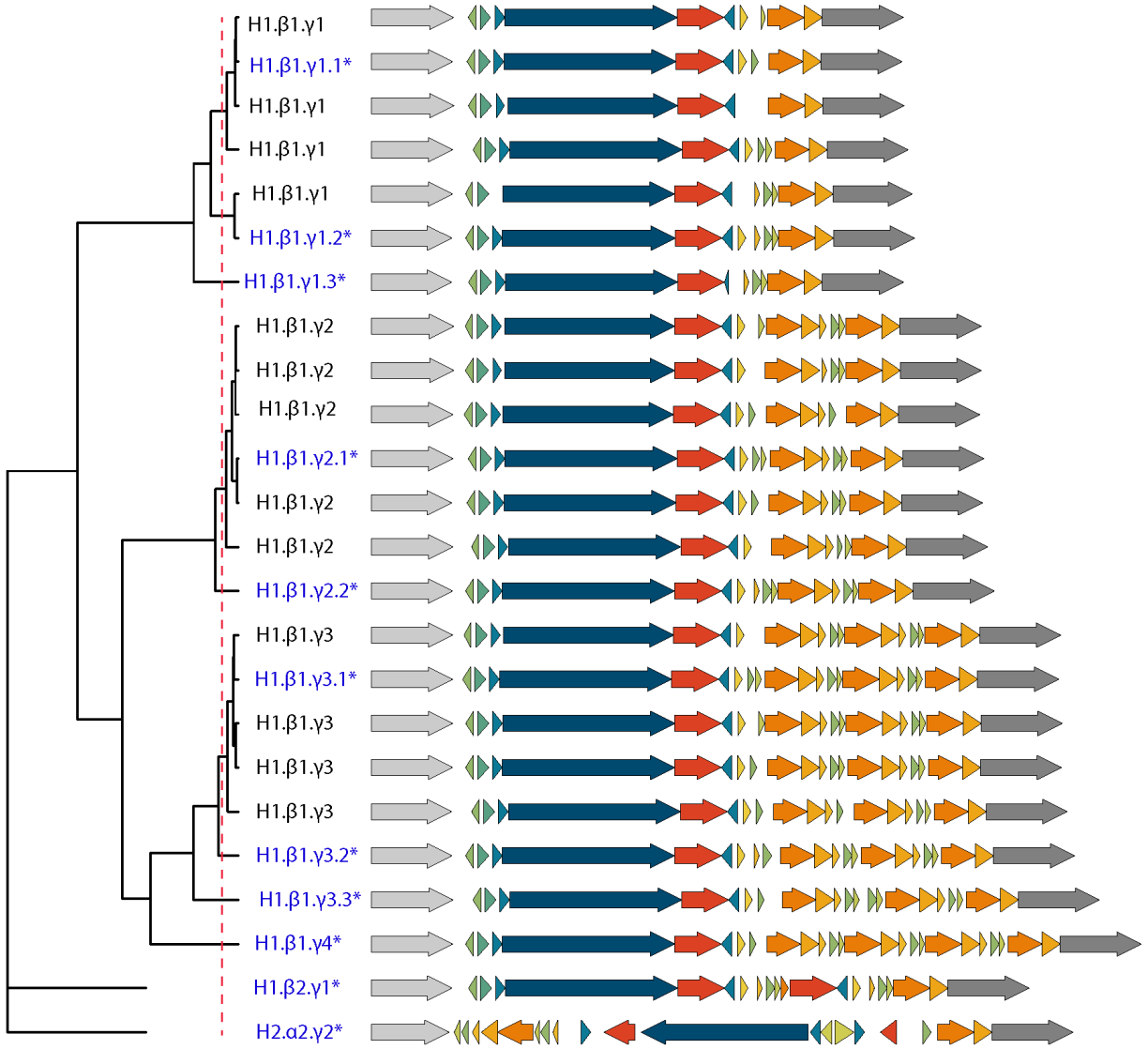
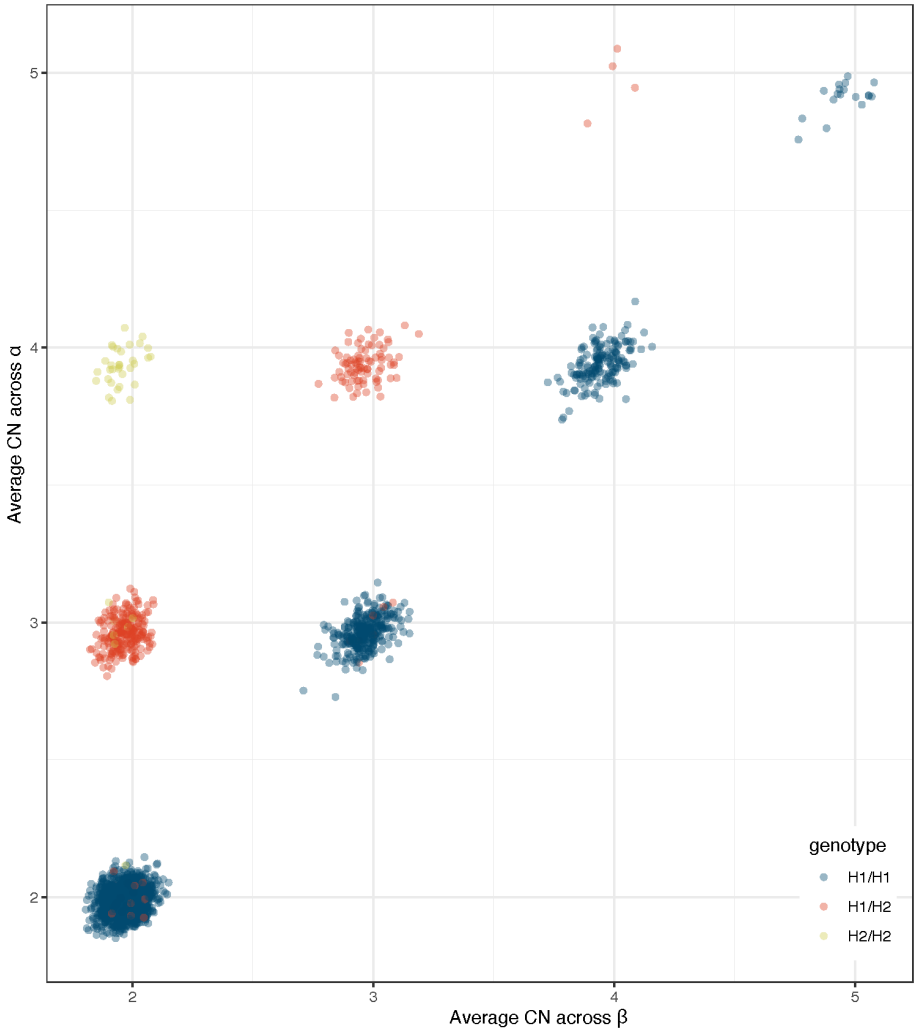


Figure 6 – Ancient genomes track the rise in frequency of *KANSL1* duplication in Europe: a) Geographic distribution of historical samples across Western Eurasia shown in 2.5 kya time slices from 10 kya to the present. Samples are split based on their haplotype structure (H1.β1, H1.β2+, H2.α1, and H2.α2+) and age in each subplot, with dots marking individual sampling locations. **b)** Temporal distribution of 17q21.31 haplotypes across historical West Eurasian populations. Stacked bar plots show the number of haplotypes present in each population over a time bin. Time is shown on the x-axis in years before present (BP). Pie charts on the right summarize the overall haplotype frequency for each population, with the numbers indicating the number of individuals in each population. **c)** Modeled temporal trajectories of 17q21.31 haplotype frequencies over the past 10 ky using a multinomial logistic regression model. Each point represents the observed haplotype frequency in a time bin (all the populations are pooled), with point size scaled by sample size.



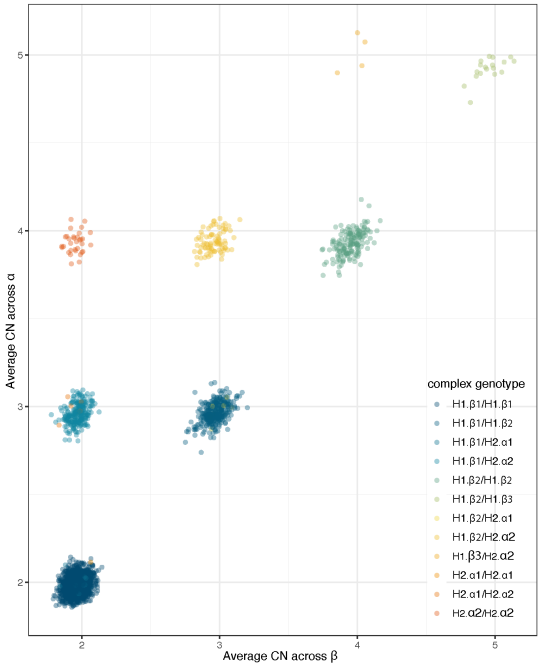
Supplementary Figure 1 – 24 distinct 17q21.31 structural haplotypes from human long read genome assemblies: Schematic of 24 unique structural haplotypes found within 210 long-read haplotype-resolved assemblies. Each row represents a unique structural haplotype. Structural haplotypes are arranged using hierarchical clustering based on jaccard distance, with tip sizes scaled by the number of assemblies sharing each haplotype. The dashed red line indicates where clades were collapsed into a single representative structure as shown in **Figure 1B**.



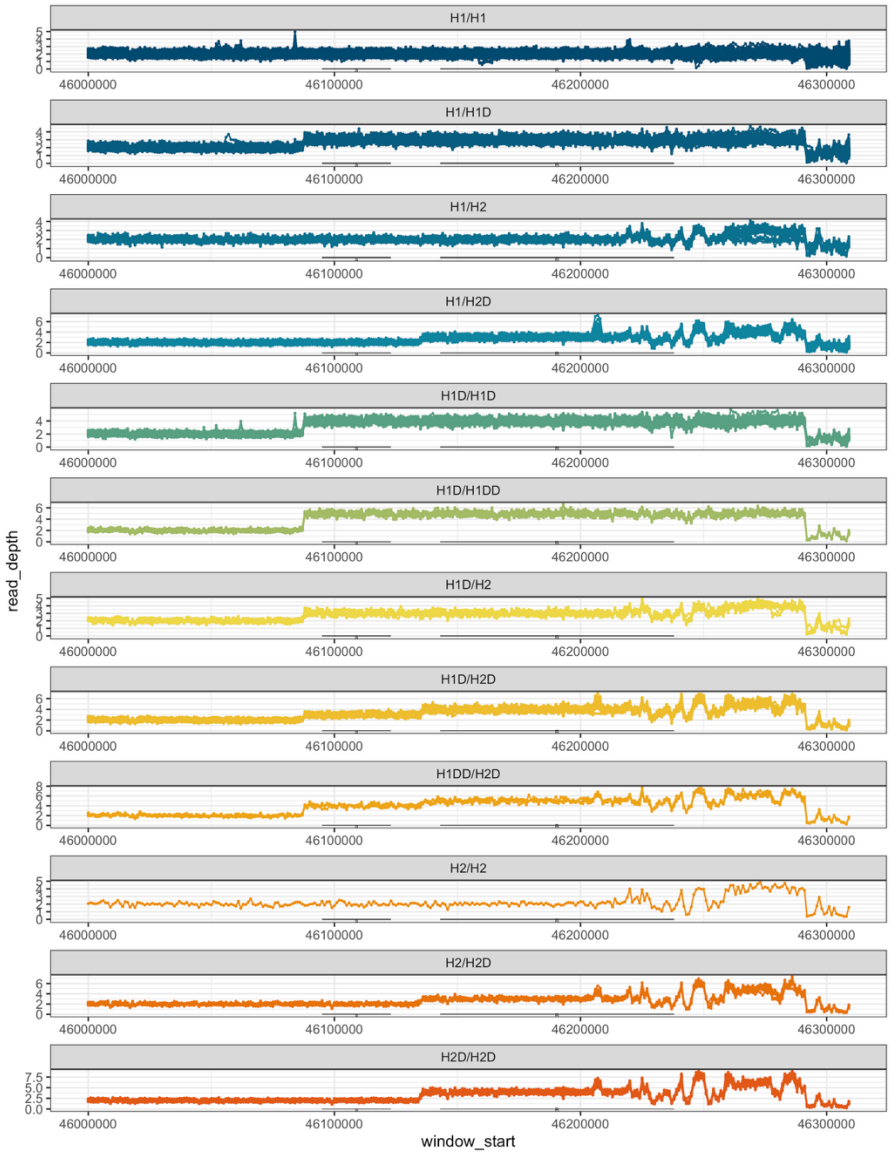
Supplementary Figure 2 – Inversion genotype assignments mapped atop copy numbers in the β and α regions: Each point represents an individual with the x-axis denoting the normalized read depth across β and the y-axis denoting the normalized read depth across α. Points are colored according to inversion genotype, assigned using 1271 tag SNPs. Jitter was applied to both the x and y axes to improve visualization of overlapping points.

	H1.β1	H1.β2	H1.β3	H1.β4	H2.α1	H2.α2	H2.α3
H1.β1	2,2						
H1.β2	3,3	4,4					
H1.β3	4,4	5,5	6,6				
H1.β4	5,5	6,6	7,7	8,8			
H2.α1	2,2	3,3	4,4	5,5	2,2		
H2.α2	2,3	3,4	4,5	5,6	2,3	2,4	
H2.α3	2,4	3,5	4,6	5,7	2,4	2,5	2,6

Supplementary Figure 3 – Table assigning complex genotype status based on α and β regions: Cartesian coordinates (x,y) represent complex genotype assignments with numbers indicating the average copy number as discrete values. The value before the comma corresponds to region β, and the value after the comma corresponds to region α. Red cells denote ambiguous cases which can only be distinguished based on 1271 tag SNPs that are required to differentiate between inversion status. Blue cells represent the most conservative assignment which was used in ambiguous cases within the direct or inverted haplotypes.



Supplementary Figure 4 – Complex genotype at the *KANSL1* region mapped atop copy numbers in the β and α regions: Each point represents an individual with the x-axis denoting the normalized read depth across β and the y-axis denoting the normalized read depth across α. Points are colored according to complex genotype status, assigned based on the *KANSL1* region using the criteria described in Supplementary Figures 2 and 3. Jitter was applied to both the x and y axes to improve visualization of overlapping points.



Supplementary Figure 5 – Read depth across the 17q21.31 locus split by CNV assigned by *KANSL1* calls: Each line represents an individual colored and grouped by complex genotype. The x-axis denotes genome position, and the y-axis shows the average normalized read depth calculated in sliding windows of 1000 bp. Grouping by complex genotype enables visual comparison of CNV patterns across complex genotypes. Horizontal grey lines at the bottom of each plot indicate the positions of the α and β regions, highlighting the differences in read depth between complex genotypes. This visualization complements the discrete copy number genotype assignments shown in Supplementary Figures 3 and 4, illustrating the underlying read depth variation that defines α and β.

TABLES

Supplementary Table 1 - Inversion tagging SNPs

Supplementary Table 2 - Frequencies of H1.β1, H1.β2, H1.β3, H2.α1, and H2.α2 in 1000 Genomes, HGDP, SGDP and CAAPA2 samples

712 **Supplementary Table 3** - Frequencies of diploid *NSF* copy numbers in 1000 Genomes, HGDP,
713 SGDP and CAAPA2 samples
714 **Supplementary Table 4** - Recombinant individuals and recombinant breakpoints in 1000
715 Genomes, HGDP, and SGDP samples
716 **Supplementary Table 5** - Frequencies of H1.β1, H1.β2, H1.β3, H2.α1, and H2.α2 in Ancient
717 European Genomes
718