Supplementary Note, Tables, and Figures

for

Structural haplotypes and recent evolution of the human 17q21.31 region

Linda M. Boettger, Robert E. Handsaker, Michael C. Zody, Steven A. McCarroll

Table of Contents

Supplementary Note

Section S1. Identification of CNV segments

To identify the genomic span of each CNV in the 17q21.31 locus, we used a combination of array and sequence data to triangulate on the breakpoints of each CNV.

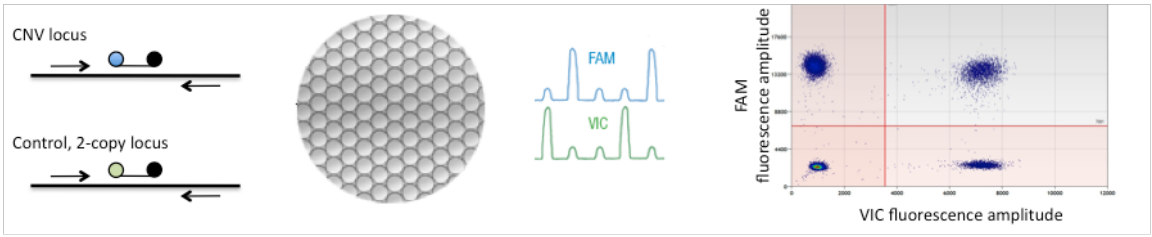
1. As a first step, we identified (at about kilobase resolution) the approximate span of CNV segments, using array-based data. Copy-number measurements from multiple array-based platforms (Illumina 1M array, SNP6.0 array, OMNI 2.5M array), data were normalized using a median polish algorithm and then used to generate (for each genomic site interrogated) a "population profile" consisting of each site's distribution of copy-number measurements across the HapMap / 1000 Genomes samples. These profiles were then clustered to identify several discrete segments of 17q21.31 each having the property that population profiles were strongly correlated within-segment but not between-segment. These segments are Regions 1, 2, and 3 in **Fig. 1a**. The boundaries of these segments were further refined (to about 100 bp resolution) by comparing read-depth profiles. This analysis used low-coverage sequence data across 946 individuals from the 1000 Genomes Project, pooled across individuals with shared high or low copy number and then compared between these two groups.

3. Precise breakpoints of these rearrangements were then identified by searching the 1000 Genomes data, using the Genome Analysis Toolkit[1]for read pairs for which one read mapped to the region identified above, and either (i) the other read mapped to the other side of the breakpoint (typically a "read pair" that offered higher-resolution information about breakpoint localization) or was unmapped (often a "split read" containing breakpoint sequence). For Region 1 and Region 2, we were able to derive putative breakpoints from the low coverage 1000 Genomes data. For the right breakpoint of Region 1, which lies in repetitive sequence, we also used high coverage sequencing data for one HapMap sample (NA12878). The breakpoints for Region 2 were confirmed by alignment to the H2 alternate haplotype assembly in GRCh37.

Section S2. Analysis of CNVs using droplet-based digital PCR

To determine integer copy number of these CNV segments in populations of genomes, we employed a molecular approach based on PCR in digitally counted droplets. Each assay involved simultaneous interrogation of the CNV locus ("locus X") and an invariant two-copy control locus ("locus Y"). For each locus, we designed a pair of PCR primers

and a dual-labeled fluorescence/FRET oligonucleotide probe (using the FAM fluorophore for locus X, and the VIC fluorophore for locus Y) that fluoresces in proportion to the accumulation of its corresponding PCR product. 20-microliter PCR reactions (each containing 5 ng genomic DNA and each primer at 900 nM and probe at 250 nM concentration) were each emulsified into approximately 20,000 droplets in an oil/aqueous emulsion, using a microfluidic droplet generator (QuantaLife). Droplet generation produces microodroplets that are uniform in size, 1 nl in volume, and that compartmentalize the PCR reaction, such that each droplet contains zero, one, or very few template molecules from each locus; the greater the copy number of a locus in the genome analyzed, the more droplets contain a PCR template for it. PCR was performed on these emulsions, which were then analyzed using a droplet reader (QuantaLife) to count the droplets that were positive and negative for each fluorophore. The droplets were resolved into four classes – (i) FAM-VIC-, (ii) FAM-VIC+, (iii) FAM+VIC-, (iv) FAM+VIC+ (Supplementary Fig. 4). Note that detection need only distinguish between positive and negative droplets; measurement of copy number comes not from precise quantitative fluorescence measurements, but from counting the numbers of positive and negative droplets for each fluorophore. By applying Poisson statistics (to account for the possibility that a droplet may have contained multiple copies of the template locus), the absolute number of copies of both genomic loci in the initial reaction was estimated. By comparing this molecular count between locus X and the control, two-copy locus Y, the integer number of copies of locus X in each diploid genome was evaluated (Fig. 1a-c).



Droplet-based digital PCR assays are listed in Supplementary Table 16.

Section S3. Analysis of CNVs using WGS read depth (Genome STRiP)

As a second method for determining the integer copy number of these CNV segments in populations, we generalized the Genome STRiP genotyping method[2] to analyze duplications in low coverage sequencing data from 1000 Genomes Phase 1. For each CNV segment, the number of observed sequenced reads falling within the CNV segment was counted for each sample, requiring a minimum mapping quality of 10, and

compared to the expected number of fragments per copy at the locus. The expected number of fragments was estimated based on the genome-wide sequencing coverage, correcting for the alignability of the segment and for sequencing bias based on the GC content of the segment. Alignability was estimated by mapping overlapping 36-mers from the reference genome back to the reference. GC-bias was estimated by counting the number of aligned reads in overlapping 400bp windows binned by GC fraction compared to a set of selected reference windows having no evidence of copy number variability.

The vectors of observed and expected read counts were fitted to a constrained Gaussian mixture model with two parameters (m_1 and m_2) and genotype classes corresponding to potential copy numbers of 0-10. The means of each genotype class were constrained to be proportional (m_1) to the copy number and the variances were constrained to be proportional (m_2) to the copy number (or to a small constant k = 0.2 for the copy number zero class). After using an expectation maximization (EM) algorithm to determine the most likely values for m_1 and m_2, the relative likelihood of the observed read depth given each potential genotype class was calculated for each genome.
We compared ddPCR and WGS read depth by calculating the average distance for each copy number dosage from the median value for each cluster. For Region 1, Read depth and ddPCR yielded similarly compact clusters; for Region 2 ddPCR data yielded clusters 43% more compact than read depth data; for Region 3, read depth data yielded clusters 25% more compact than ddPCR. However, the precision of both methods can vary. WGS Read Depth is affected by region length, sequencing depth, read length, and GC content, while ddPCR is affected by the molecular sequences used in each assay.

Section S4. Determination of haplotypic contributions to diploid copy number, and heuristic phasing in trios

In 67% of trios, only one combination of haplotypic copy numbers was possible under the assumptions that duplication α is present only on the H2 background, duplication β is present only on the H1 background, and no haplotype has a copy number of zero in the α, β or γ regions. In other trios, multiple potential combinations were consistent with the experimental data; we initially assigned an equal likelihood to all possible combinations of hapotypic copy numbers (that were consistent with the three trio members'

experimentally determined diploid copy numbers). We then created the following expectation-maximization (EM) loop: from the probabilistic (often certain) inferences of haplotypic copy number in each trio, we estimated an allele frequency for each copy-number allele; we then re-estimated the relative likelihood of each combination of haplotypic copy numbers in each trio, given the population-level allele frequency. (This has the effect of eliminating haplotypic combinations that are theoretically possible but extremely unlikely given the apparent frequencies of copy-number alleles as estimated from the rest of the population.) The revised probabilistic estimates then allowed a new population-level estimate of copy-number allele frequencies. We repeated this EM loop process until estimates of allele frequency converged, generally requiring only 1-3 loops of the EM.

Section S5. Statistical phasing of structural and fine-scale variation in populations

The full features of imputation algorithms such as Beagle are today available primarily for biallelic variants. To apply imputation software to multi-allelic CNVs such as those analyzed here, we encoded the structural genotypes of each sample by creating a series of surrogate biallelic markers with 0/1 alleles as a form of binary code. The states of the multi-allelic polymorphisms were thereby encoded (and subsequently decoded) as combinations of multiple biallelic markers. Once the CNV state is encoded, Beagle is allowed to phase the surrogate markers independently. While it was technically possible for Beagle to infer a phased haplotype with an “illegal” encoding, this happened only rarely (and never in the 1000 Genomes panel) suggesting that the input genotypes have high accuracy and that Beagle was able to successfully ascertain the relationships among these surrogate markers from population-level data. The structural states were encoded in 12 biallelic surrogate markers as follows:

ID	Encoding
H1H2	0 for H1 allele, 1 for H2 allele
R1H1_1	1 for H1 allele with Region 1 CN >= 1
R1H1_2	1 for H1 allele with Region 1 CN >= 2
R1H1_3	1 for H1 allele with Region 1 CN >= 3
R2H2_1	1 for H2 allele with Region 2 CN >= 1
R2H2_2	1 for H2 allele with Region 2 CN >= 2

R3H1_1	1 for H1 allele with Region 3 CN >= 1
R3H1_2	1 for H1 allele with Region 3 CN >= 2
R3H1_3	1 for H1 allele with Region 3 CN >= 3
R3H1_4	1 for H1 allele with Region 3 CN >= 4
R3H2_1	1 for H2 allele with Region 3 CN >= 1
R3H2_2	1 for H2 allele with Region 3 CN >= 2

The following table illustrates this encoding for the nine observed CNV haplotypes.

Haplotype	**H1/H2**	**R1H1 (β)**	**R2H2 (α)**	**R3H1 (γ on H1)**	**R3H2 (γ on H2)**
H1.β1.γ1	0	1 0 0	0 0	1 0 0 0	0 0
H1.β1.γ2	0	1 0 0	0 0	1 1 0 0	0 0
H1.β1.γ3	0	1 0 0	0 0	1 1 1 0	0 0
H1.β1.γ4	0	1 0 0	0 0	1 1 1 1	0 0
H1.β2.γ1	0	1 1 0	0 0	1 0 0 0	0 0
H1.β3.β1	0	1 1 1	0 0	1 0 0 0	0 0
H2.α1.γ2	1	0 0 0	1 0	0 0 0 0	1 1
H2.α2.γ1	1	0 0 0	1 1	0 0 0 0	1 0
H2.α2.γ2	1	0 0 0	1 1	0 0 0 0	1 1

To evaluate the ability to phase and impute these structural haplotypes we created two reference panels, one using SNP genotypes from 1000 Genomes Phase 1 within +/- 300Kb of the CNV locus and one using SNP genotypes from Hapmap3 using SNPs within +/- 1Mb of the CNV locus. In each panel, we discarded all SNP genotypes within the CNV region from chr17:44165000-44785000 and replaced these with the encoded genotypes for the surrogate CNV markers. These panels used different subsets of the 467 individuals in our pool based on the availability of SNP genotypes in each data set. Using the Hapmap3 panel, we also evaluated the ability to impute these structural haplotypes using only SNPs from the Illumina 1M and the Affymetrix 6.0 chips. The composition of each panel is shown below.

Panel	Samples			Flanks	SNPs	
	1000G	Trios	Total		Reference	Imputation
1000 Genomes	373	63	436	1Mb	6302	6302
Hapmap3	211	89	300	300Kb	814	814
Illumina 1M	211	89	300	300Kb	814	559
Affymetrix 6.0	211	89	300	300Kb	814	355

To evaluate the utility of these reference panels as an imputation resource, we performed a series of “leave one out” trials. For each trial, we constructed a new panel of unphased markers where we set to “missing” the genotypes for the twelve surrogate CNV markers for one of the trio founders and then used Beagle to phase the resulting panel and impute the missing markers. Beagle was run with default parameters plus two additional settings (nsamples=20 and niterations=50). Both of these settings increase accuracy at some additional computational cost (each leave-one-out trial in the larger 1000 Genomes panel took between one and two hours). By forcing Beagle to re-phase the panel each time, we avoided any potential problems with data from the left-out individual impacting the phasing of the reference panel.

After running Beagle, the imputed surrogate markers were decoded to recover the phased CNV haplotypes for the target sample. For each reference panel, we performed this experiment separately for each trio founder. The imputed structural haplotypes were compared to the known phased haplotypes determined from ddPCR analysis of the trios. We evaluated the genotype concordance at each surrogate biallelic marker and to the CNV genotypes formed by decoding multiple markers (e.g. copy number on each haplotype for regions α, β and γ). We also measured the allelic r^2 (correlation between the estimated dosage for each sample compared to the dosages for the true genotypes) (Supplementary Tables 13-14).

Using the Hapmap3 panel, we performed two additional sets of leave-one-out trials to evaluate the power to impute the CNV state using only SNPs from the Illumina 1M chip or the Affymetrix 6.0 chip. These trials were performed identically to the Hapmap3 trial, except that for the left-out individual we supplied only the SNP genotypes corresponding to the target chip (for the remaining reference individuals we used all Hapmap3 SNP genotypes).

Section S6. Estimation of CNV dosages

Dosages for composite markers, including the diploid copy number for alpha, beta and gamma, were computed by summing the Beagle dosages for the constituent surrogate markers. For example, dosage for region alpha was computed as the sum of the dosages for R2H2_1 and R2H2_2, the dosage for region beta was computed as the sum of the dosages for R1H1_1, R1H1_2 and R1H1_3, etc.

In our encoding scheme, a copy number variant with maximum haploid copy number *N* is represented as *N* surrogate bi-allelic markers m_i with alleles labeled arbitrarily as A (zero) and B (one) where a haploid copy number of *k* is encoded as the first *k* markers having allele B and the last *N-k* markers having allele A. Beagle will generate dosages for these surrogate markers based on their posterior likelihoods:

$$\boldsymbol{dosage}(\boldsymbol{m_i}) = \mathbf{0} * \mathbf{Pr}(\boldsymbol{m_i} = \boldsymbol{AA}) + \mathbf{1} * \mathbf{Pr}(\boldsymbol{m_i} = \boldsymbol{AB}) + \mathbf{2} * \mathbf{Pr}(\boldsymbol{m_i} = \boldsymbol{BB})$$

which can also be interpreted as

$$\boldsymbol{dosage}(\boldsymbol{m_i}) = \mathbf{0} * \mathbf{Pr}(\boldsymbol{CN}(\boldsymbol{m_i}) = \mathbf{0}) + \mathbf{1} * \mathbf{Pr}(\boldsymbol{CN}(\boldsymbol{m_i}) = \mathbf{1}) + \mathbf{2} * \mathbf{Pr}(\boldsymbol{CN}(\boldsymbol{m_i}) = \mathbf{2})$$

where $CN(m_i)$ is the diploid count of B alleles at m_i. Under our encoding scheme, we define the dosage for a CNV region as a whole in a compatible way as

$$\boldsymbol{dosage}(\boldsymbol{CNV}) = \sum_{\boldsymbol{i-0}}^{\boldsymbol{2N}} \boldsymbol{i} * \boldsymbol{Pr}(\boldsymbol{CN}(\boldsymbol{CNV}) = \boldsymbol{i})$$

where *CN(CNV)* is the diploid copy number of the CNV region. The likelihood that *CN(CNV)* = *X* can be expressed as the sum of the joint likelihoods over all possible genotypes for the m_i markers that sum to *X*:

$$\boldsymbol{Pr}(\boldsymbol{CN}(\boldsymbol{CNV}) = \boldsymbol{i}) = \sum_{\boldsymbol{c_1}+\cdots+\boldsymbol{c_N}=\boldsymbol{i}} \prod_{\boldsymbol{j}=\mathbf{1}}^{\boldsymbol{N}} \boldsymbol{Pr}(\boldsymbol{CN}(\boldsymbol{m_j}) = \boldsymbol{c_j})$$

Using the fact that

$$\sum_{\boldsymbol{j}=\mathbf{0}}^{\mathbf{2}} \mathbf{Pr}(\boldsymbol{CN}(\boldsymbol{m_j}) = \boldsymbol{j}) = \mathbf{1}$$

it can be shown algebraically that

$$\boldsymbol{dosage}(\boldsymbol{CNV}) = \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{N}} \sum_{\boldsymbol{j}=\mathbf{0}}^{\mathbf{2}} \boldsymbol{j} * \mathbf{Pr}(\boldsymbol{CN}(\boldsymbol{m_i}) = \boldsymbol{j}) = \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{N}} \boldsymbol{dosage}(\boldsymbol{m_i})$$

and thus we can calculate the estimated dosage for the CNV region by simply summing the dosages for the surrogate markers m_i that encode that region.

Section S7. Identification of *KANSL1* fusion gene transcripts

Genomic breakpoints for the H1-polymorphic *KANSL1* duplication (duplication β) were analyzed and primers were designed to amplify a predicted fusion transcript, which we had obtained from mRNA clone BC006271. (This mRNA is currently annotated as a potential ligation/fusion artifact, but its sequence is consistent with our model of duplication β as a tandem duplication, as this mRNA fuses the 5' exons of *KANSL1* to sequence on the other side of the β duplicon. Using cDNA from derived from cell line GM10854 (H1.β1.γN, H1.β2.γN), a fusion transcript was amplified and sequenced (Supplementary Fig. 6a). We observed this fusion in cDNA from individuals whose genomic DNA indicated that they carried the β duplication, but never in other individuals.

In order to predict a potential fusion transcript for the H2-polymorphic *KANSL1* duplication (duplication α), we mined RNA-seq data for the presence of *KANSL1* fusion transcripts in published CEU RNA sequence data[3]. Paired-end reads found in NA11920 (H2.α2.γ2, H1.β1.γ3) revealed two separate breakpoints:

GCGTCATGTACCCCTAGACGTGGGAACAACGCAAGTC
TCCTGTGGTCTGCTGGGAAGTTTTTCCAGTTCAACGG

This information was used to design primers to amplify and sequence the transcript. cDNA derived from GM20806 (H2.α2.γ2, H2.α2.γ2) was used as template (Supplementary Fig. 6b). We observed amplification of this fusion in cDNA from individuals whose genomic DNA indicated that they carried the α duplication, but never in other individuals.

See Supplementary Table 16 for primer sequences.

Section S8. Dating the coalescence of duplication-containing chromosomes

Using Agilent SureSelect capture probes and Illumina sample barcoding, we generated targeted sequencing of the unique inverted region (UIR) plus flanking duplications and ~150 kb of flanking unique sequence (43.41 Mb to 44.85 Mb on GRCh37) for 4 individuals with β duplication (diploid copy number 4 in Regions 1 and 2, H1.β2.γ1) from Gujarati Indians from Houston, TX (GIH, population frequency of H1 β duplication 36%)

and Utah residents (CEPH) with Northern and Western European ancestry (CEU, population frequency of H1 β duplication 26%) (GIH: NA21106, NA21113, NA21137; CEU: NA11894). We sequenced these samples to greater than 100-fold coverage of single copy regions (except NA21137, which was only covered to slightly over 50-fold) using Illumina MiSeq with 150 bp paired end reads.

We used a custom genotyping program that examined only sites where all of the remaining samples had sufficient coverage (12-fold for single copy regions) to produce a confident genotype to call SNP genotypes across the unique inverted region (GRCh37 coordinates 43,705,166 to 44,165,259, once the β duplication was excluded). We then compared these calls to 1000 Genomes low coverage genotypes for H1/H1 individuals with diploid copy 2 for regions 1 and 2 (lacking the β duplication). We determined that all SNPs segregating at minor allele count > 1 among the captured H1.β2.γ1 chromosomes matched high frequency SNPs in the unduplicated population, indicating that these were evidence of recombination between the duplicated chromosomes and the unduplicated chromosomes and not variation arising on the duplicated background since the duplication. Based on this, we identified one individual (NA21113) who carried two unrecombined chromosomes across the entire UIR, two additional individuals who had no recombination within ~185 kb of the duplication (NA21137, NA11894), and a fourth with no recombination within ~25 kb of the duplication (NA21106). We note in support of this identification that these appear to be common haplotypes resulting from historic recombinations and that the coordinates match the recombination loci identifiable from the major clusters of the β duplication haplotypes derived from trio-based phasing (**Figure 3**). We also note that combining our analysis with the haplotypes from phasing and imputation, we observe only these two recombinations within the 460 kb region, suggesting that the region has undergone minimal recombination since the most recent common ancestor of all duplication containing chromosomes. Within each of these regions (full UIR, most distal 185 kb of UIR, most distal 25 kb of UIR) we computed average pairwise diversity among the non-recombinant chromosomes and between the non-recombinant chromosomes and the clone-based chimpanzee assembly of the region[4], excluding from consideration bases without a cleanly aligned orthologous chimpanzee base. By taking the ratio of the average duplicated chromosome diversity to the average chimp divergence and assuming human-chimpanzee speciation at 6 Mya, we estimated the average coalescence of the duplicated chromosomes. Using either

one sequence over the full region or three sequences over 185 kb yields the same estimate of 11.9 kya (there were no polymorphisms in the 4 sequences over the most distal 25 kb, so the estimate was 0). Because we are selectively removing recombinant chromosomes (which might have a deeper-than-average coalescent with other chromosomes), we risk introducing bias, but we note that both regions with different sets of chromosomes produce the same estimate and that using the largest number of chromosomes directly adjacent to the duplication yields no diversity at all, suggesting that we are not unfairly biasing when using fewer chromosomes over longer stretches. Because these estimates arise from a small number of polymorphic sites, we performed 1000 bootstraps by selecting with replacement on the columns of the alignments and computed the 95% confidence intervals for diversity, divergence, and human/chimpanzee diversity/divergence ratio for each of the three datasets (Supplementary Table 12). Both samples with any diversity yield similar ranges, with the outer bounds being 4.1 to 20.6 kya. We note that our technique only estimates the age of the most recent common ancestor of the sampled chromosomes, and that the MRCA of all duplicated chromosomes may be older; however, as our sampled chromosomes come from two diverged populations (Europeans and Indians) and our sequences form a perfect star phylogeny, we infer that the common ancestor of our sampled chromosomes is likely close to the overall ancestor. Of course the duplication event itself is almost certainly older than the MRCA.

To attempt to date the duplication itself, we took the capture data from these four sequences across the portion of the beta duplication which is not part of any other known segmental duplication (in the reference genome) and is not overlapping annotated high copy repeats. We then adjusted our genotyping program to require 36-fold coverage of all samples at each site and allowed for 3:1 as well as 2:2 polymorphisms, because each individual would have 4 alleles of beta. This resulted in an estimate for the age of the beta duplication of 19.6 kya (95%CI: 6.0 - 35.1) based on observing 38,804 nucleotide positions. We note that this diversity estimate averages the age of the duplication (comparison of either chromosome's proximal copy to either chromosome's distal copy) with the age of coalescence (comparison of the proximal and distal copies within either chromosome). Using the coalescence data of the chromosomes determined from the UIR of 11.9 kya, the duplication might be 27.3 k years old, although we note that the confidence intervals on both estimates overlap

extensively, suggesting that the beta duplication may not predate the expansion of the haplotype by very much.

The above analyses capture only the observed diversity in extant sequences, but not other variation present in the population but not sampled.

To analyze coalescence of chromosomes containing the alpha duplication, we performed a similar analysis for 3 individuals (2 Tuscan and 1 Gujarati) homozygous for H2.α2.γ2 (TSI: NA20770 and NA20768 and GIH: NA20890), all covered at more than 100-fold by MiSeq data. We used all three samples across the full UIR, analyzing 310,276 sites (we were not concerned about recombination with other haplotypes because of the inversion and the extremely low frequency of non-duplicated H2 chromosomes in these populations), from which we estimated an average pairwise diversity of 0.0028% (95%CI: 0.0017-0.0038%), estimating an age of 16.5 kya (10.6-22.6 kya). Again, this only estimates the time to most recent common ancestor of the six sampled chromosomes, but we again observe a star-like tree from diverged populations and thus infer that we are close to inferring the overall MRCA.

However, the alpha duplication itself is certainly much older. Comparing the two copies of the reference sequence across the unique portion of the alpha duplication, we get a divergence of 0.214%, compared to 0.804% between human and chimpanzee (average of the divergence estimates from H1 and the 2 H2 copies, whose divergences from chimpanzee were not significantly different (χ^2)) and a divergence between H1 and H2 (average of the two H2 copies, not significantly different in their divergence from H1) of 0.418%. Depending on whether we calibrate to a human-chimpanzee divergence of 6 Mya or an H1-H2 divergence of 2.3 Mya[4], we get either 1.6 Mya or 1.2 Mya. The 1.6 Mya result, based on divergence of humans and chimpanzees does not carry over any potential uncertainty in the H1-H2 dating and is more consistent with the results from resequencing (below).

As an alternate method of dating the alpha duplication, we used our MiSeq data from the three H2.α2.γ2 resequencing individuals and the genotyping method described above for the beta duplication (36-fold minimum coverage and allowing 3:1 allele frequencies). Based on these data (35,419 sites), we infer an age for the alpha duplication of 896 kya

(95%CI: 727-1095). However, we again note that this method averages the age of separation of the two alpha copies with the time to the most recent common ancestor of each copy, which we estimated above to be only 16.5 kya. Thus our result for the resequencing-based dating is consistent with the average of the 1.6 Mya age derived from comparison of the two complete reference copies to chimp and the 16.5 kya age of coalescence of the observed H2.α2.γ2 haplotypes (808 kya, well within the CI for the resequencing-based estimate). Lastly, we note that the portion of the alpha duplication which overlaps a copy of the older segmental duplication (the "A" duplicon) appears to have undergone recent gene conversion with the adjacent A duplicon copy, as it is more similar to that copy of A than to the copy adjacent to the original copy of alpha. Examining only that difference gives a misleadingly young age for alpha.

These observations are consistent with the previous suggestion that the H2.α2.γ2 haplotype may be an old structural type that has recently come to high frequency[4]. One possible explanation for that would be that the haplotype has segregated at very low frequency for over a million years and only recently risen in frequency in non-African populations. An alternate theory would be that the duplication event occurred between heterogeneous chromosomes with very different histories, so that the duplication occurred recently but combined two copies of the sequence that had diverged long ago. Because both alpha copies are equally diverged from H1, we considered H1 unlikely to be a source (although possibly some as yet unsampled H1 chromosomes would have more similarity to the second copy of alpha). Another possibility is the H2.α1.γ2 haplotype, which has not yet been fully sequenced and assembled. To test this, we used HiSeq resequencing data from a single H2.α1.γ2 individual (NA20589; TSI) to construct a list (incomplete) of sites where H2.α1.γ2 differs from the H1 reference. We then looked to see if the H2.α1.γ2 calls were more similar to the duplicated (distal) copy of alpha than the proximal copy. However, we instead found that overwhelmingly the H2.α1.γ2 calls were more similar to the original (proximal) alpha (69 of 69 sites where the two reference alpha copies differ). These data support the H2.α1.γ2 alpha copy being largely identical to the proximal reference (H2.α2.γ2) copy, indicating that this is likely not the origin of the second copy, although this single individual may poorly represent the diversity present in the H2.α1.γ2 population. It might be possible that the proximal reference copy

is the duplicated copy, but this is not consistent with a model of the inversion being a simple event occurring on a background with only a single alpha.

We next considered the possibility that an extinct hominid might have been the source of the distal copy. It has previously been considered and rejected that the entire H2 haplotype derived from an extinct hominid, Neanderthal[5-7],but we considered the possibility that only the alpha duplication arose via introgression. We performed a similar analysis to that described for the H2.α1.γ2 haplotype, aligning reads from both Neanderthal[7] and Denisovan[8] to the H1 reference and identifying sites different from H1. The vast majority of such sites found in both cases represented singletons on the extinct lineage or sequencing errors, as they differed from H1 and both H2 alpha copies (1160/1161 for Neanderthal, 2245/2248 for Denisova). In all other cases, the extinct hominid matched both copies of alpha (being identical at those positions, 1 case for Neanderthal and 3 for Denisova). Again, we cannot rule out the existence of other Neanderthal or Denisovan haplotypes (or other extinct homonids) that could have been the source of the second H2 alpha copy, but currently available data show no support for origin through introgression.

Section S9. Analysis of allele frequency differentiation between European and non-European populations

To provide a background distribution with which to evaluate the observed allele frequency differentiation of the α and β duplications, we used SNP data from the 1000 Genomes phase 1 release for chromosome 17, excluding the region from chr17:43165000-45785000 (one megabase upstream and downstream from the structurally polymorphic 17q21.31 locus).

We had observed the α duplication segregating at allele frequency 19% in CEU and appearing only once among the 942 non-European chromosomes sampled by 1000 Genomes (in a Japanese individual on whose chr17 the H2.α2 form presented on a 2 Mb segment that an ancestry Hidden Markov Model classified as apparently European in origin, even when we excluded the SNPs within the 17q21.31 inversion region itself; we nonetheless treat this as a real observation in a non-European genome.) Of 7,013 SNPs with allele frequency between 18% and 20% in Europe, 37 (0.53%) were observed 0-1 times among the non-European population samples.

Evaluating the β duplication (allele frequency 26% in CEU, and not observed at all among the 942 CHB, CHS, JPT, LWK, and YRI chromosomes sampled in 1000 Genomes Phase 1), out of 6,127 SNPs with allele frequency between 25% and 27% in Europe, 4 SNPs (0.065%) were monomorphic in the non-European populations.

One variable in this analysis is the genotyping error rate in the 1000 Genomes data for these sites, particularly errors in the non-European populations where the SNPs being used for comparison are rare. Supplementary Figure 5 shows the cumulative fraction of these SNPs observed with low allele counts in the 1000 Genomes non-European populations. Considering SNPs that are at 19% frequency in Europe with 0-5 observed minor alleles in the non-European populations (corresponding to a 1% genotyping error rate), the observed frequency of such SNPs is 1.93%. For SNPs that are at 26% frequency in Europe with 0-5 minor alleles in the non-European populations, the observed frequency of such SNPs is 0.49%.

Section S10. Analysis of allele frequency differentiation within Europe

To provide a comparison for the observed differences in allele frequency for β duplication across Europe, we used SNP data from the 1000 Genomes phase 1 release for chromosome 17 excluding the region from chr17:43165000-45785000 (one megabase upstream and downstream from the structurally polymorphic 17q21.31 locus) and including only SNPs with minor allele frequency > 5% in both the TSI and CEU populations. For each SNP, we calculated F_{ST} between the TSI (n=85) and CEU (n=98) populations. Of these 148,964 SNPs, 1082 (0.73%) had F_{ST} higher than F_{ST} for the β duplication (0.0545).

Section S11. A structural history of the 17q21.31 region

Here we describe the structural evolution of the 17q21.31 locus using a combination of clone-based sequence data, digital droplet PCR copy quantitation (Fig. 1), and low pass sequencing data from the 1000 Genomes project (Fig. 1).

We start with the known complete or nearly complete haplotypes. Based on prior work[4,6], we can currently establish with high confidence two structural types segregating in the human population distinguished primarily by an inversion of approximately 590 kb of sequence that is locally unique on the reference genome. In Supplementary Figure 1 we

mark these two sequences as the H1 reference type (H1.β1.γ2) and H2 reference type (H2.α2.γ2). The "unique inverted region" (UIR) is flanked on either side in the H1 reference genome by inverted repeat elements we term "A" duplicons (in cyan on the bottom line of each haplotype diagram in Supplementary Fig. 1). All variants of the "A" duplicons within the region appear to derive from a single ancestral sequence which was copied intact in inverted form from one end of the inversion region to the other sometime between the divergence of orangutan (which has only one "A" element) and the divergence of human and chimpanzee (which has two)[4]. The extant human copies have diverged both through point mutation and variable deletion of a large stretch of Alu-rich sequence, but also show evidence of having undergone recombination or gene conversion in recent human history. Supplementary Figure 7 depicts the sequence contained in each of the A duplicons noted on these two reference chromosomes compared to the full length H2A1 copy.

In the H1 reference the flanking A duplicons are both approximately 130 kb in length and each contains a gene of the *LRRC37A* family and a large amount of non-coding Alu-rich DNA. The proximal (H1A0) and distal (H1A1) copies of A differ from each other by slightly overlapping deletions of approximately 30 kb each in the Alu-rich region, most likely resulting from differential loss of sequence due to NAHR between Alu elements. These differential sequences are represented by a dark blue box (for the H1A0-specific sequence) and a magenta box (for the H1A1-specific) on the middle lines of the haplotype diagrams (Supplementary Fig. 1). Multiple lines of evidence support those deleted sequences all being present in an ancestral version of the A duplicon, most notably the presence of both deleted regions in their entirety in the proximal (H2A1) duplicon of the H2 reference haplotype (Supplementary Figure 7; more detail below). To the distal end of the A1 duplicon lies a stretch of approximately 140 kb containing several exons of the gene *NSF* which we term the "B" duplicon (in green on the bottom line of each haplotype diagram Supplementary Fig. 1). On the H1 reference type, the "B" duplicon and ~75 kb of the "A" duplicon are tandemly duplicated. We refer to these sequences as "B2" and "A2". We collectively refer to this duplication as the γ duplication (noted in various shades of green in the top line of the haplotype diagrams in Supplementary Fig. 1 and as the combined green and orange boxes in Fig. 2; note that we cannot tell with certainty which copy is the original, i.e., whether the new copy was added to the distal end of B1 or inserted into the middle of the original A1 copy). We

note, however that the most proximal 20 kb of “A2” consist of sequence present in the reference “A0” but not the reference “A1”, suggesting that this is not a tandem duplication, a point we will refute later. This reference structure is also referred to as H1.β1.γ2, meaning H1 orientation with 1 copy of the β duplication and 2 copies of the γ duplication. This copy is marked with base pair positions of feature boundaries in GRCh37 coordinates.

As previously observed, the H2 reference has a much more complex structure. The inverted region is again flanked by A elements, but they are structurally different. The proximal flanking A, designated H2A1 because it is more similar to H1A1 than the positionally orthologous H1A0, is 165 kb long and, as noted, contains the full sequence of both H1A0 and H1A1 (depicted by the adjacent dark blue and magenta boxes in the middle line of the haplotype diagram in Supplementary Fig. 1, also shown in Supp. Fig. 7). Proximal to the H2A1 is a full length “B” element we denote H2B1, due to its positional similarity to H1B1. Proximal to H2B1 is a truncated “A” element, H2A2, which is structurally identical to H1A2, although the most proximal few kb are more similar at the nucleotide level to H1A0, possibly indicating a breakpoint of inversion, or maybe simply the result of more recent recombination with an H1 chromosome. At the distal end, the sequence is heavily rearranged. Although there is a large gap in the GRCh37 reference in the distal flanking repeat region and a failure to connect back to unique sequence on the distal end, sequencing of several additional unfinished clones combined with population data from 1000 Genomes allows us to infer that the structure must be very close what is represented in Supplementary Figure 1. The defining feature is a duplicative transposition of ~67 kb of the proximal “A” element and the flanking 81 kb of sequence from the UIR. These are arranged in direct orientation w/r/t their proximal copies and flanked by further partial “A” elements. Although it is possible that this sequence originated from an inverted duplication of the orthologous sequence from the distal end of an H1 chromosome, sequence similarity data from the duplicated part of the UIR suggests that it is more likely derived from an H2 chromosome than an H1 (Supplementary Figure 8, Supplementary Text S8). On Supplementary Figure 1, the duplicated portion of the UIR is colored in red in the middle line (on the H2 reference chromosomes and also at the orthologous position on the H1 chromosomes) and the entire duplication, termed the α duplication, is shown on the top line in red (distal copy) and orange (proximal copy). Note that the portion of the duplication derived from the A

element (termed “H2A5”) is in inverted tandem orientation with the remnant of the original distal flanking A (termed “H2A4”) with a short spacer of ~12 kb (depicted in dark blue on the middle bar). Although in theory H2A5 should show greatest sequence similarity to the copy that gave rise to the α duplication, it shows clear evidence of having undergone recent gene conversion from H2A4, most likely by an inverted folding mechanism as observed on the Y chromosome [9,10]. The H2A4 and H2A5 copies show only 23 differences (including 8 unreliable simple sequence length variations) over a stretch of ~55 kb, whereas the UIR parts of the duplicated copies differ by 0.23%, suggesting that the duplication may be very old (though we cannot exclude the possibility that one of the duplicated copies came from a different chromosome, which would partially explain the high divergence) (Supplementary Figure 8). The inverted tandem repeat continues to confound efforts to generate a finished clone-based sequence across this region, but using a combination of deep clone-based sequencing, overlapping Fosmids of a non-identical H2 individual, and population data from 1000 Genomes, we can very closely infer the structure and content of the missing sequence. Several of the uncertain junctions have been validated by direct PCR from H2 individuals (other than the reference donor, whose DNA is not directly available). Distal to the α duplication lies a short (~15 kb) piece of “A” duplicon termed H2A6 followed by a longer element termed H2A3 which shows substantial structural and sequence homology to H1A0. Distal to that is a second “B” element. Although the current reference ends there, draft sequencing of a clone that links out to unique distal sequence shows that the H2 reference has a single distal B (H2B2). We refer to this structural type as H2.α2.γ2, for H2 orientation with 2 copies of the α duplicon and 2 copies of the γ duplicon.

Our last known haplotype is the chimp type. Constructed from BAC clones from the reference donor Clint[4], the clone-based chimp assembly is in H2 orientation. We note that the whole genome shotgun chimp assembly lacks assembly of the A duplicon regions and is oriented in H1 order off the human reference, although the distal end is to the left in the chromosome because of a large pericentric inversion that is fixed between humans and chimpanzees[11]. The clone-based chimp assembly has a much simpler duplication structure with only one “A” element at each end of the region and a single “B” element on the distal end of the region. The chimp “A” elements have undergone substantial deletion within the Alu-rich regions that is different from that which has occurred in the human “A” copies, and further, the chimp copies may have been

structurally swapped during the independent inversion that occurred in the chimp lineage[4], so ancestral inference about "A" element structure from the chimp sequence is not certain. We can also infer from the chimp and orangutan, which has only a single A and a single B[4], that the ancestral human structure, regardless of orientation, was likely flanked by one "A" copy at either end and a single "B" at the distal end (in addition to the chimp evidence, a single copy "B" would be functionally constrained to be adjacent to the 3' end of NSF).

We do observe such a structure currently present at high frequency in all populations (Fig. 2 and Supplementary Fig. 1), labeled H1 ancestral (H1.β1.γ1). Although we do not have the complete haplotype sequence of any such individual, we see 137 H1 homozygous individuals in 1000 Genomes low coverage samples with only 2 diploid copies of β and 2 diploid copies of γ (based on genomeSTRiP analysis). We propose that this represents the ancestral structure of human H1. Surprisingly, in these samples, we also see elevated copy number in the region of H1A0 that is deleted from both H1A1 and H1A2, consistent with 2 haploid copies of that region. This indicates that this sequence is present in an additional copy on the ancestral H1 chromosome. Most likely it is present in the H1A1 copy, making that copy structurally identical to the H2A1 copy. This also explains the previously mysterious presence of this H1A0 sequence at the end of the H1A2 copy that should have been derived from H1A1, and indicating that despite the appearance to the contrary in the reference chromosome, the γ duplication occurred as a direct tandem event. We designate this ancestral distal A copy H1A1' to distinguish it from the reference H1A1.

From this ancestral H1, we can generate two separate paths representing the two H1 duplications we observe. The first path is simply the one we have traced backwards from the reference to the ancestral. A tandem duplication (the γ duplication) of the distal part of H1A1' and all of H1B1 occurs to form H1A2 and H1B2 (although we cannot say whether A1B1 or B2B2 is the "original" copy). We also observe individuals with haploid γ duplication copy number of 3, 4, and possibly 5, presumably arising from unequal crossing over between elevated γ copy chromosomes. These are depicted in Supplementary Fig. 1 as H1.β1.γ3 and H1.β1.γ4 (1 copy of β and 3 and 4 copies of γ). We have no complete assemblies of these haplotypes but observe them in read depth data and digital droplet PCR data. Read coverage data suggest that the elevated copies

look more like the reference H1B1 copy than H1B2 (the A portion of the γ duplication has too many copies elsewhere to be cleanly differentiated by depth of read coverage). However, we cannot rule out the possibility that gene conversion has occurred between different copies of the duplication. At some point after the original γ duplication, a subunit of H1A1' deletes to form H1A1 (the dark blue box in the middle line of haplotype diagrams). We know that this happens no earlier than the initial γ duplication, but may be concurrent with it. We do not see evidence in current chromosomes of the presence of this sequence in A1 copies in any H1 chromosomes with elevated γ copy number (>1), but cannot rule out that it still exists. We do not currently infer a date for this original duplication, but as it is found in all human populations examined, we expect it to be quite old.

At a much later date (best estimate, less than 20,000 years ago, see Supplemental Text S8), the ancestral H1 experienced another tandem duplication spanning part of the *KANSL1* gene, all of H1A1', and a small part of B. We term this the β duplication. This event is only seen in populations with European ancestry. We do not have a complete sequence of such a chromosome, but we can confirm the breakpoints of this event at single base resolution with breakpoint-spanning reads from NA12878[1]. Again, depth of coverage data from 1000 Genomes low coverage sequencing shows that copies of the H1A0/H1A1' sequence deleted from the reference H1A1 tracks linearly with copies of the β duplication. This shows that the β duplication derived from an ancestral H1 with a full length H1A1'. This is depicted as H1.β2.γ1 (2 β, 1 γ) in Supplementary Figure 1. We also observed a smaller number of individuals who had 3 copies of the β duplication (H1.β3.γ1), apparently arising from unequal crossing over between the two copies of the β duplication on H1.β2.γ1 chromosomes.

The H1 resident β duplication is clearly distinct from the H2 resident α duplication in both structure and phylogeny. It includes the full length H1A1', while the α duplication includes only about half of an A1. The β duplication also includes almost 50 kb of additional sequence from the UIR. On Supplementary Figure 1, the part of the β duplication that is shared with α is colored red in the middle bar and the part that is unique to the β duplication is colored in orange.

Switching back to the H2, we also note a second, rarer type, lacking the complex duplication of the reference H2. Based on digital droplet PCR and also on 1000 Genomes read depth, we infer that this structure has also 2 copies of the γ duplication (which appears to be shared between H1 and H2), which we infer to be on either end of the region as in the H2 reference (although we cannot rule out the possibility that these copies are in tandem on the simpler H2, as they are in H1), and only one copy of α. Based on depth of coverage, we also infer that the distal "A" element, which we term H2A0', is full length and structurally homologous to the proximal H2A1. Lastly, there appears to be a third H2 type, present at very low frequency, which lacks the proximal copy of the γ duplication (shown on Fig. 2 but omitted on Supplementary Figure 1) and is likely to have resulted from a recombination between H2 and H1 chromosomes in pairing of H2A1 and H1A0. Further investigation will be required to confirm this.

To derive the duplicated H2 from the simpler and presumably ancestral H2 requires a complex rearrangement event, as described by the FoSTeS or MMBIR models[12,13]. Roughly, we suppose that a break occurs in the distal H2A0' region, giving rise to the A3, A4, and A6 regions of the H2 reference genome. Replication jumps to the H2A1 copy and reads into the inverted region to generate the α duplication, the end of which is rejoined to the A6 segment by a non-homologous repair mechanism. This results in a direct dispersed duplication of approximately 150 kb flanked by several fragmentary "A" element copies. As noted, the A5 piece presumably derived from H2A1 has undergone recent gene conversion to become almost identical to A4.

Lastly, we address the question of whether the H1 sequence or the H2 sequence is truly ancestral. Previously, in the absence of a clear sequence of structural mutations, patterns of SNP variation (in HapMap) were used to infer that the H2 form was more likely to be ancestral[4], based on an apparently greater proportion of H1-polymorphic SNPs than H2-polymorphic SNPs being in the ancestral state on the inversion type on which they were monomorphic. We further examined this earlier inference using the more-complete set of SNPs available from the 1000 Genomes Project phase 1. This analysis revealed equal frequencies of ancestral SNPs among H1 variants fixed in H2 (91%) and H2 variants fixed in H1 (93%) ($p = 0.66$, $\chi 2$). By contrast, the structural variation seems to favor a model in which H1 was the ancestral form: the H2.α1.γ2 structure allows one to reconstruct inversional paths with either H1 or H2 as the

ancestral type, but the H1 to H2 path appears to require fewer steps to explain all the observed and inferred modern haplotypes. We can generate a valid recombinational path through a pair of mismatched H1.β1.γ2 chromosomes to generate an H2.α1.γ2 chromosome (**Supplementary Figure 4**, top panel). This event would be consistent with the relative similarity of the "A" elements, as the H2 copies are all more similar over most of their length to the H1A1 copy than they are to the H1A0. We can also make a valid path through an H2.α1.γ2 chromosome to generate either an H1.β1.γ1 (H1 ancestral) or an H1.β1.γ2 (H1 reference) chromosome, depending on where the path exits on the distal end (**Supplementary Figure 4**, bottom panel).

However, several structural points argue in favor of the H1 being ancestral. First, in the chimp, and presumably in the ancestral human, there is only one "B" element, and it is distal to all copies of A. This matches the H1.β1.γ1 structure. In order to get to the H2.α1.γ2 state from that ancestral state, a duplicative transposition event would be required which would put the A2 portion of the H2A0' and all of B on the proximal end of the region inserted into the middle of the H1A1 repeat to create the proximal A-B-A sequence. Furthermore, if this inverted to form H1.β1.γ1, it would then have to duplicate again at exactly the same breakpoints to make H1.β1.γ2 (H1 reference). If, as seems more plausible, the H2.α1.γ2 inverted to form H1.β1.γ2, a deletion of the γ duplication would be required to get to H1.β1.γ1. Furthermore, since we know that the H1A1' element in the H1.β1.γ1 chromosomes contains sequence not present in the H1.β1.γ2, but duplicated in the very recent β duplication event, there does not seem to be any plausible way to work backwards structurally from an H2-derived H1.β1.γ2 to the all the observed H1 structures.

However, the sequence of the unique inverted regions of the H1 and H2 references diverge more (0.45%) than any of the duplicated sequences on the distal end of H1, suggesting that the H1-H2 split must be older, whereas the scenario outlined above would not only make H1 ancestral but would actually place the entire H2 lineage within the extant H1 types (as it would derive from H1.β1.γ2 while there are still chromosomes with H1.β1.γ1 structure that must have split from the H1.β1.γ2 types prior to the inversion that formed the H2 lineage [caveat: this is true unless modern H1.β1.γ1 chromosomes are more recently derived from NAHR between the γ copies of an H1.β1.γ2

chromosome]). Although this seems inconsistent, it may not be. Since the inverted region of H2 has been effectively recombinantly isolated from H1 since the inversion, it has not coalesced with any H1 haplotypes, and its lineage traces as far back as the haplotype on which the inversion occurred. On the other hand, even if they are structurally older, the H1 chromosomes have been free to recombine with each other. As suggested by the SNP ancestry data, all the H1 chromosomes may share a common ancestor more recently than the inversion, but that does not mean that the inversion is older than all the H1 structures, only that it is older than the observed variation on the H1 chromosomes.

Supplementary Tables

Supplementary Table 1. Breakpoints and supporting evidence for α, β, γ duplications

Note that bases indicates the actual count of bases in the region that met the criteria

Dup	Copy	Prox. Break	Support	Dist. Break	Support
α	Proximal	499,261	clone	650,665	clone
α	Distal	1,245,098	clone	1,381,792	clone
α	H1 reference projection	44,212,781	clone	44,366,715	clone
β	reference	44,165,260	HiSeq	44,433,878	HiSeq
γ	H1 proximal	44,369,527	clone	44,566,775	clone
γ	H1 distal	44,566,776	clone	44,784,489	clone
γ	H2 proximal	263,940	clone	481,883	clone
γ	H2 distal	1,397,032	clone	n/a	not in assembly

Breakpoints for H1 events are given in GRCh37 reference assembly coordinates.
Breakpoints for H2 events are given in GRCh37 ALT_REF_LOCI_9 coordinates.
H2 α breakpoints are shown projected to H1 coordinates for purposes of demonstrating the sequence content relationship between α and β, both spanning part of *KANSL1*.
Clone support indicates that we were able to determine breakpoints by completely aligning multiple duplication copies present in large insert clones from the same haplotype.
HiSeq support indicates that we have assembled data from deep whole genome sequencing of an individual (NA12878) containing a β duplication. Breakpoint (//) spanning sequence =
ACCAGCCTGGCCAACAGGGTAAAACTCCGTCTCCACTAATAATACAAAAATTAGCC
GGGTGTGGTGGCGTGCACCTGTAATCCCAGCTTCTCAG//CTGTCATGCTACCCCAA
CATGGGCTTCCCTAACATT

Supplementary Table 2. H1/H2 states by population (per diploid genome)

Haplotype orientation was determined as described in section S4 of the Methods.
A blank cell indicates that a genotype or haplotype is never observed in our sample from that population.

	H1,H1	**H1,H2**	**H2,H2**
CEU (N=108)	0.60	0.39	0.01
FIN (N=73)	0.75	0.19	0.04
GBR (N=68)	0.50	0.47	0.03
TSI (N=98)	0.43	0.43	0.14
ASW (N=48)	0.83	0.17	
LWK (N=79)	1.00		
YRI (N=100)	1.00		
CHB (N=78)	1.00		
CHS (N=91)	1.00		
JPT (N=78)	0.99	0.01	
CLM (N=47)	0.64	0.34	0.02
MXL (N=54)	0.69	0.24	0.07

Supplementary Table 3. H1/H2 states by population (per haplotype)

	H1	**H2**
CEU (N=216)	0.80	0.20
FIN (N=146)	0.86	0.14
GBR (N=136)	0.74	0.26
TSI (N=196)	0.64	0.38
ASW (N=96)	0.92	0.08
LWK (N=158)	1.00	
YRI (N=200)	1.00	
CHB (N=156)	1.00	
CHS (N=182)	1.00	
JPT (N=-156)	0.99	0.01
CLM (N=94)	0.81	0.19
MXL (N=108)	0.81	0.19

Supplementary Table 4. Region 1 copy number (per diploid genome)

Measured from whole-genome sequence data (1000 Genomes Project phase 1) as described in Methods S3.

Copy Number	**2**	**3**	**4**	**5**
CEU (N=108)	0.57	0.33	0.09	0.01
FIN (N=73)	0.72	0.26	0.02	
GBR (N=68)	0.62	0.34	0.04	
TSI (N=98)	0.75	0.19	0.05	0.01
ASW (N=48)	0.90	0.10		
LWK (N=79)	1.00			
YRI (N=100)	1.00			
CHB (N=78)	1.00			
CHS (N=91)	1.00			
JPT (N=78)	1.00			
CLM (N=47)	0.68	0.21	0.11	
MXL (N=54)	0.83	0.17		

Supplementary Table 5. Region 1 copy number (estimated allele frequencies)

Estimated from whole-genome sequence data (1000 Genomes Project phase 1) as described in Methods S3, S5 and S6.

Copy Number	**1**	**2**	**3**
CEU (N=216)	0.74	0.25	0.01
FIN (N=146)	0.86	0.14	
GBR (N=136)	0.79	0.21	
TSI (N=196)	0.86	0.12	0.02
ASW (N=96)	0.95	0.05	
LWK (N=158)	1.00		
YRI (N=200)	1.00		
CHB (N=156)	1.00		
CHS (N=182)	1.00		
JPT (N=-156)	1.00		
CLM (N=94)	0.83	0.13	0.04
MXL (N=108)	0.92	0.08	

Supplementary Table 6. Region 2 copy number (per diploid genome)

Measured from whole-genome sequence data (1000 Genomes Project phase 1) as described in Methods S3.

Copy Number	**2**	**3**	**4**	**5**
CEU (N=108)	0.30	0.49	0.20	0.01
FIN (N=73)	0.54	0.34	0.12	
GBR (N=68)	0.31	0.46	0.23	
TSI (N=98)	0.31	0.45	0.22	0.02
ASW (N=48)	0.75	0.25		
LWK (N=79)	1.00			
YRI (N=100)	1.00			
CHB (N=78)	1.00			
CHS (N=91)	1.00			
JPT (N=78)	0.99	0.01		
CLM (N=47)	0.43	0.40	0.13	0.04
MXL (N=54)	0.59	0.30	0.11	

Supplementary Table 7. Region 2 copy number (estimated allele frequencies)

Estimated from whole-genome sequence data (1000 Genomes Project phase 1) as described in Methods S3, S5 and S6.

Copy Number	**1**	**2**	**3**
CEU (N=216)	0.55	0.44	0.01
FIN (N=146)	0.72	0.28	
GBR (N=136)	0.54	0.46	
TSI (N=196)	0.55	0.43	0.02
ASW (N=96)	0.87	0.13	
LWK (N=158)	1.00		
YRI (N=200)	1.00		
CHB (N=156)	1.00		
CHS (N=182)	1.00		
JPT (N=156)	0.99	0.01	
CLM (N=94)	0.65	0.30	0.05
MXL (N=108)	0.77	0.20	0.03

Supplementary Table 8. Region 3 copy number (per diploid genome)

Measured from whole-genome sequence data (1000 Genomes Project phase 1) as described in Methods S3.

Copy Number	2	3	4	5	6	7	8
CEU (N=108)	0.28	0.29	0.23	0.17	0.03		
FIN (N=73)	0.26	0.19	0.27	0.14	0.14		
GBR (N=68)	0.22	0.37	0.26	0.09	0.06		
TSI (N=98)	0.20	0.37	0.35	0.06	0.02		
ASW (N=48)	0.19	0.38	0.35	0.08			
LWK (N=79)	0.16	0.42	0.32	0.08	0.02		
YRI (N=100)	0.20	0.39	0.31	0.10			
CHB (N=78)	0.17	0.20	0.31	0.14	0.15	0.03	
CHS (N=91)	0.19	0.12	0.31	0.15	0.17	0.02	0.04
JPT (N=78)	0.19	0.14	0.37	0.10	0.17	0.03	
CLM (N=47)	0.34	0.32	0.23	0.09	0.02		
MXL (N=54)	0.20	0.37	0.19	0.13	0.09	0.02	

Supplementary Table 9. Region 3 copy number (estimated allele frequencies)

Estimated from whole-genome sequence data (1000 Genomes Project phase 1) as described in Methods S3, S5 and S6.

Copy Number	1	2	3	4
CEU (N=216)	0.52	0.3	0.17	0.01
FIN (N=146)	0.49	0.19	0.31	0.01
GBR (N=136)	0.47	0.39	0.11	0.03
TSI (N=196)	0.45	0.45	0.10	
ASW (N=96)	0.41	0.51	0.08	
LWK (N=158)	0.41	0.5	0.08	0.01
YRI (N=200)	0.43	0.49	0.08	
CHB (N=156)	0.41	0.22	0.33	0.04
CHS (N=182)	0.43	0.15	0.30	0.12
JPT (N=156)	0.45	0.13	0.39	0.03
CLM (N=94)	0.58	0.28	0.13	0.01
MXL (N=108)	0.46	0.38	0.08	0.08

Supplementary Table 10. Structural haplotypes inferred from phasing in trios

Demined by Methods S2 and S7.
Haplotypes in each trio are denoted as "-T" for transmitted or "-U " for untransmitted.

Source individual	Trio relationship	H1/H2	Region 1 copy number	Region 2 copy number	Region 3 copy number	Haplotype
NA12872	father-T	H2	1	2	2	H2.α2.γ2
NA12872	father-U	H1	1	1	1	H1.β1.γ1
NA12873	mother-T	H1	1	1	1	H1.β1.γ1
NA12873	mother-U	H1	2	2	1	H1.β2.γ1
NA12891	father-T	H1	2	2	1	H1.β2.γ1
NA12891	father-U	H1	2	2	1	H1.β2.γ1
NA12892	mother-T	H1	2	2	1	H1.β2.γ1
NA12892	mother-U	H1	1	1	3	H1.β1.γ3
NA12874	father-T	H1	2	2	1	H1.β2.γ1
NA12874	father-U	H2	1	2	1	H2.α2.γ1
NA12875	mother-T	H1	1	1	1	H1.β1.γ1
NA12875	mother-U	H1	1	1	1	H1.β1.γ1
NA12812	father-T	H1	1	1	1	H1.β1.γ1
NA12812	father-U	H1	3	3	1	H1.β3.γ1
NA12813	mother-T	H1	1	1	1	H1.β1.γ1
NA12813	mother-U	H2	1	2	2	H2.α2.γ2
NA12760	father-T	H1	1	1	1	H1.β1.γ1
NA12760	father-U	H1	1	1	1	H1.β1.γ1
NA12761	mother-T	H2	1	2	2	H2.α2.γ2
NA12761	mother-U	H1	2	2	1	H1.β2.γ1
NA12750	father-T	H1	1	1	2	H1.β1.γ2
NA12750	father-U	H1	1	1	1	H1.β1.γ1
NA12751	mother-T	H1	2	2	1	H1.β2.γ1
NA12751	mother-U	H1	2	2	1	H1.β2.γ1
NA12003	father-T	H1	1	1	1	H1.β1.γ1
NA12003	father-U	H1	1	1	2	H1.β1.γ2
NA12004	mother-T	H1	1	1	1	H1.β1.γ1
NA12004	mother-U	H1	1	1	3	H1.β1.γ3
NA12248	father-T	H1	2	2	1	H1.β2.γ1
NA12248	father-U	H1	1	1	1	H1.β1.γ1
NA12249	mother-T	H1	1	1	3	H1.β1.γ3
NA12249	mother-U	H1	1	1	1	H1.β1.γ1
NA11992	father-T	H1	1	1	1	H1.β1.γ1
NA11992	father-U	H2	1	2	2	H2.α2.γ2
NA11993	mother-T	H1	2	2	1	H1.β2.γ1
NA11993	mother-U	H1	2	2	1	H1.β2.γ1
NA12716	father-T	H1	1	1	1	H1.β1.γ1

NA12716	father-U	H1	1	1	3	H1.β1.γ3
NA12717	mother-T	H1	3	3	1	H1.β3.γ1
NA12717	mother-U	H1	2	2	1	H1.β2.γ1
NA11829	father-T	H1	1	1	1	H1.β1.γ1
NA11829	father-U	H1	1	1	1	H1.β1.γ1
NA11830	mother-T	H1	1	1	2	H1.β1.γ2
NA11830	mother-U	H1	1	1	2	H1.β1.γ2
NA12056	father-T	H1	1	1	2	H1.β1.γ2
NA12056	father-U	H1	1	1	3	H1.β1.γ3
NA12057	mother-T	H1	1	1	1	H1.β1.γ1
NA12057	mother-U	H1	2	2	1	H1.β2.γ1
NA07034	father-T	H1	1	1	3	H1.β1.γ3
NA07034	father-U	H1	1	1	2	H1.β1.γ2
NA07055	mother-T	H1	2	2	1	H1.β2.γ1
NA07055	mother-U	H1	1	1	1	H1.β1.γ1
NA06994	father-T	H1	1	1	3	H1.β1.γ3
NA06994	father-U	H2	1	2	2	H2.α2.γ2
NA07000	mother-T	H1	1	1	1	H1.β1.γ1
NA07000	mother-U	H1	1	1	1	H1.β1.γ1
NA12146	father-T	H2	1	2	2	H2.α2.γ2
NA12146	father-U	H1	2	2	1	H1.β2.γ1
NA12239	mother-T	H1	2	2	1	H1.β2.γ1
NA12239	mother-U	H1	1	1	1	H1.β1.γ1
NA12827	father-T	H1	1	1	1	H1.β1.γ1
NA12827	father-U	H1	1	1	1	H1.β1.γ1
NA12828	mother-T	H2	1	2	2	H2.α2.γ2
NA12828	mother-U	H1	1	1	1	H1.β1.γ1
NA12829	father-T	H1	1	1	2	H1.β1.γ2
NA12829	father-U	H1	1	1	1	H1.β1.γ1
NA12830	mother-T	H1	1	1	3	H1.β1.γ3
NA12830	mother-U	H1	1	1	1	H1.β1.γ1
NA12777	father-T	H1	2	2	1	H1.β2.γ1
NA12777	father-U	H1	2	2	1	H1.β2.γ1
NA12778	mother-T	H1	1	1	3	H1.β1.γ3
NA12778	mother-U	H1	2	2	1	H1.β2.γ1
NA11930	father-T	H1	1	1	1	H1.β1.γ1
NA11930	father-U	H1	1	1	1	H1.β1.γ1
NA11931	mother-T	H2	1	2	2	H2.α2.γ2
NA11931	mother-U	H1	1	1	3	H1.β1.γ3
NA11917	father-T	H1	2	2	1	H1.β2.γ1
NA11917	father-U	H2	1	2	2	H2.α2.γ2
NA11918	mother-T	H1	1	1	1	H1.β1.γ1
NA11918	mother-U	H1	2	2	1	H1.β2.γ1

NA12272	father-T	H1	1	1	1	H1.β1.γ1
NA12272	father-U	H1	2	2	1	H1.β2.γ1
NA12273	mother-T	H2	1	2	2	H2.α2.γ2
NA12273	mother-U	H1	1	1	2	H1.β1.γ2
NA11893	father-T	H1	2	2	1	H1.β2.γ1
NA11893	father-U	H1	1	1	2	H1.β1.γ2
NA11894	mother-T	H1	2	2	1	H1.β2.γ1
NA11894	mother-U	H1	2	2	1	H1.β2.γ1
NA12413	father-T	H1	1	1	1	H1.β1.γ1
NA12413	father-U	H1	1	1	1	H1.β1.γ1
NA12414	mother-T	H1	1	1	1	H1.β1.γ1
NA12414	mother-U	H1	1	1	1	H1.β1.γ1
NA12399	father-T	H1	2	2	1	H1.β2.γ1
NA12399	father-U	H1	2	2	1	H1.β2.γ1
NA12400	mother-T	H1	1	1	1	H1.β1.γ1
NA12400	mother-U	H1	1	1	1	H1.β1.γ1
NA12546	father-T	H1	1	1	3	H1.β1.γ3
NA12546	father-U	H1	1	1	1	H1.β1.γ1
NA12489	mother-T	H2	1	1	2	H2.α1.γ2
NA12489	mother-U	H1	1	1	1	H1.β1.γ1
NA12342	father-T	H1	2	2	1	H1.β2.γ1
NA12342	father-U	H1	1	1	1	H1.β1.γ1
NA12343	mother-T	H1	1	1	1	H1.β1.γ1
NA12343	mother-U	H1	1	1	2	H1.β1.γ2
NA07435	father-T	H1	1	1	1	H1.β1.γ1
NA07435	father-U	H1	1	1	3	H1.β1.γ3
NA07037	mother-T	H1	2	2	1	H1.β2.γ1
NA07037	mother-U	H1	2	2	1	H1.β2.γ1
NA12347	father-T	H2	1	2	1	H2.α2.γ1
NA12347	father-U	H1	2	2	1	H1.β2.γ1
NA12348	mother-T	H1	1	1	1	H1.β1.γ1
NA12348	mother-U	H2	1	2	2	H2.α2.γ2
NA06984	father-T	H1	2	2	1	H1.β2.γ1
NA06984	father-U	H1	2	2	1	H1.β2.γ1
NA06989	mother-T	H1	1	1	2	H1.β1.γ2
NA06989	mother-U	H2	1	2	2	H2.α2.γ2
NA07347	father-T	H2	1	2	2	H2.α2.γ2
NA07347	father-U	H1	2	2	1	H1.β2.γ1
NA07346	mother-T	H1	1	1	2	H1.β1.γ2
NA07346	mother-U	H1	2	2	1	H1.β2.γ1
NA12762	father-T	H1	1	1	2	H1.β1.γ2
NA12762	father-U	H1	2	2	2	H1.β2.γ1
NA12763	mother-T	H2	1	2	2	H2.α2.γ2

NA12763	mother-U	H1	1	1	3	H1.β1.γ3
NA12005	father-T	H1	2	2	1	H1.β2.γ1
NA12005	father-U	H2	1	2	2	H2.α2.γ2
NA12006	mother-T	H1	1	1	4	H1.β1.γ4
NA12006	mother-U	H2	1	2	2	H2.α2.γ2
NA12264	father-T	H1	1	1	2	H1.β1.γ2
NA12264	father-U	H1	2	2	1	H1.β2.γ1
NA12234	mother-T	H1	2	2	1	H1.β2.γ1
NA12234	mother-U	H2	1	2	2	H2.α2.γ2
NA11994	father-T	H1	1	1	2	H1.β1.γ2
NA11994	father-U	H1	2	2	1	H1.β2.γ1
NA11995	mother-T	H1	1	1	3	H1.β1.γ3
NA11995	mother-U	H2	1	2	2	H2.α2.γ2
NA12043	father-T	H1	1	1	3	H1.β1.γ3
NA12043	father-U	H1	2	2	1	H1.β2.γ1
NA12044	mother-T	H1	1	1	3	H1.β1.γ3
NA12044	mother-U	H1	2	2	1	H1.β2.γ1
NA12144	father-T	H2	1	2	2	H2.α2.γ2
NA12144	father-U	H1	2	2	1	H1.β2.γ1
NA12145	mother-T	H1	1	1	1	H1.β1.γ1
NA12145	mother-U	H1	1	1	3	H1.β1.γ3
NA12889	father-T	H1	1	1	1	H1.β1.γ1
NA12889	father-U	H2	2	2	2	H2.α2.γ2
NA12890	mother-T	H2	1	1	2	H2.α1.γ2
NA12890	mother-U	H1	2	2	1	H1.β2.γ1
NA12775	father-T	H1	1	1	3	H1.β1.γ3
NA12775	father-U	H1	1	1	2	H1.β1.γ2
NA12776	mother-T	H1	2	2	1	H1.β2.γ1
NA12776	mother-U	H1	1	1	3	H1.β1.γ3
NA11919	father-T	H1	1	1	3	H1.β1.γ3
NA11919	father-U	H1	2	2	1	H1.β2.γ1
NA11920	mother-T	H2	1	2	2	H2.α2.γ2
NA11920	mother-U	H1	1	1	3	H1.β1.γ3
NA12286	father-T	H2	1	2	2	H2.α2.γ2
NA12286	father-U	H1	1	1	1	H1.β1.γ1
NA12287	mother-T	H1	2	2	1	H1.β2.γ1
NA12287	mother-U	H2	1	2	2	H2.α2.γ2
NA12748	father-T	H2	1	2	2	H2.α2.γ2
NA12748	father-U	H1	1	1	3	H1.β1.γ3
NA12749	mother-T	H1	1	1	3	H1.β1.γ3
NA12749	mother-U	H1	2	2	1	H1.β2.γ1
NA07357	father-T	H1	1	1	1	H1.β1.γ1
NA07357	father-U	H1	1	1	3	H1.β1.γ3

NA07345	mother-T	H1	2	2	1	H1.β2.γ1
NA07345	mother-U	H1	1	1	3	H1.β1.γ3
NA11891	father-T	H1	2	2	1	H1.β2.γ1
NA11891	father-U	H1	1	1	3	H1.β1.γ3
NA11892	mother-T	H1	1	1	1	H1.β1.γ1
NA11892	mother-U	H1	3	3	1	H1.β3.γ1
NA11881	father-T	H1	2	2	1	H1.β2.γ1
NA11881	father-U	H2	1	2	2	H2.α2.γ2
NA11882	mother-T	H2	1	2	2	H2.α2.γ2
NA11882	mother-U	H1	1	1	3	H1.β1.γ3
NA12045	father-T	H2	1	2	2	H2.α2.γ2
NA12045	father-U	H1	1	1	1	H1.β1.γ1
NA12046	mother-T	H1	1	1	3	H1.β1.γ3
NA12046	mother-U	H1	1	1	1	H1.β1.γ1
NA06993	father-T	H1	1	1	3	H1.β1.γ3
NA06993	father-U	H2	1	2	2	H2.α2.γ2
NA06985	mother-T	H1	1	1	4	H1.β1.γ4
NA06985	mother-U	H2	1	2	2	H2.α2.γ2
NA12842	father-T	H1	1	1	2	H1.β1.γ2
NA12842	father-U	H1	1	1	1	H1.β1.γ1
NA12843	mother-T	H2	2	2	2	H2.α2.γ2
NA12843	mother-U	H1	1	1	3	H1.β1.γ3

Supplementary Table 11. Frequencies of structural haplotypes by population

Haplotype	H1.β1.γ1	H1.β1.γ2	H1.β1.γ3	H1.β1.γ4	H1.β2.γ1	H1.β3.γ1	H2.α1.γ2	H2.α2.γ2	H2.α2.γ1
CEU (N=216)	0.26	0.1	0.17	0.01	0.25	0.01	0.01	0.18	0.01
FIN (N=146)	0.35	0.05	0.31	0.01	0.14			0.14	
GBR (N=136)	0.26	0.13	0.11	0.03	0.21		0.01	0.25	
TSI (N=196)	0.31	0.08	0.10		0.12	0.02	0.06	0.31	
ASW (N=96)	0.35	0.43	0.08		0.05		0.01	0.08	
LWK (N=158)	0.41	0.50	0.08	0.01					
YRI (N=200)	0.43	0.48	0.08	0.01					
CHB (N=156)	0.41	0.22	0.33	0.04					
CHS (N=182)	0.43	0.15	0.30	0.12					
JPT (N=156)	0.45	0.12	0.39	0.03				0.01	
CLM (N=94)	0.41	0.09	0.13	0.01	0.13	0.04	0.02	0.17	
MXL (N=108)	0.38	0.19	0.08	0.09	0.08		0.06	0.12	

Supplementary Table 12. Alignment bootstrap and 95% confidence intervals

Note that bases indicates the actual count of bases in the region that met the criteria of having an orthologous chimpanzee base, having sufficient coverage in all assayed samples, and not being filtered as uncallable due to data quality.

Region	Date (kya)	5% bootstrap	95% bootstrap	Samples (Haplotypes)	Bases
UIR	11.9	5.1	20.2	1 (2)	345,403
185 kb	11.9	4.1	20.6	3 (6)	107.070
25 kb	0	0	0	4 (8)	12,375

Supplementary Table 13. Imputation allelic r^2 values

The following table lists the allelic r^2 values between the imputed samples in the leave-one-out trials and the truth data derived from completely phased trios at the surrogate markers used to encode the CNV alleles as well as dosages for composite markers (Methods, section S9).

Marker	1000 Genomes	Hapmap3	Illumina 1M	Affymetrix 6.0
R1H1_1	1.000	1.000	1.000	1.000
R1H1_2	0.890	0.649	0.657	0.610
R1H1_3	0.061	0.003	0.006	0.001
R2H2_1	1.000	1.000	1.000	1.000
R2H2_2	0.997	1.000	1.000	0.997
R3H1_1	1.000	1.000	1.000	1.000
R3H1_2	0.803	0.700	0.696	0.525
R3H1_3	0.658	0.690	0.692	0.600
R3H1_4	0.000	0.002	0.011	0.005
R3H2_1	1.000	1.000	1.000	1.000
R3H2_2	0.867	0.902	0.902	0.903
Alpha	0.999	1.000	1.000	0.999
Beta	0.928	0.787	0.798	0.770
Gamma	0.843	0.800	0.803	0.683
H1 Gamma	0.884	0.841	0.843	0.734
H2 Gamma	0.966	0.975	0.975	0.975
H1/H2	1.000	1.000	1.000	1.000

Supplementary Table 14. Imputation concordance

The following table lists the concordance between the haplotypes imputed in the leave-one-out trials and the haplotypes determined by trio phasing. For each panel, the nmber of matching haplotypes (or individuals) is shown along with the number tested in each category and the fraction of concordant haplotypes. In the first row, both haplotypes of an individual must match exactly to be counted as a match. The remaining rows are compared on a per-haplotype basis. Asterisks indicate a category of haplotype defined by the prefix and concordance is measured only against the prefix (for example, H1.*.* shows the concordance of the H1/H2 state on the set of all H1 haplotypes in the test individuals).

	1000 Genomes			Hapmap3			Illumina 1M			Affymetrix 6.0		
	Match	Total	%	Match	Total	%	Match	Total	%	Match	Total	%
Indivs	43	63	0.68	48	89	0.54	53	89	0.60	42	89	0.47
Haps	104	126	0.83	132	178	0.74	138	178	0.78	128	178	0.72
H1.*.*	102	102	1.00	144	144	1.00	144	144	1.00	144	144	1.00
H1.1.*	65	67	0.97	79	92	0.86	81	92	0.88	74	92	0.80
H1.β1.γ1	33	35	0.94	33	47	0.70	35	47	0.75	33	47	0.70
H1.β1.γ2	3	11	0.27	1	16	0.06	2	16	0.13	0	16	0.00
H1.β1.γ3	17	20	0.85	23	27	0.85	24	27	0.89	21	27	0.78
H1.β1.γ4	0	1	0.00	0	2	0.00	0	2	0.00	0	2	0.00
H1.2.*	29	32	0.91	43	50	0.86	45	50	0.90	42	50	0.84
H1.β2.γ1	29	32	0.91	43	50	0.86	45	50	0.90	42	50	0.84
H1.3.*	0	3	0.00	0	2	0.00	0	2	0.00	0	2	0.00
H1.β3.γ1	0	3	0.00	0	2	0.00	0	2	0.00	0	2	0.00
H2.*.*	24	24	1.00	34	34	1.00	34	34	1.00	34	34	1.00
H2.1.*	2	2	1.00	2	2	1.00	2	2	1.00	2	2	1.00
H2.α1.γ2	2	2	1.00	2	2	1.00	2	2	1.00	2	2	1.00
H2.2.*	22	22	1.00	32	32	1.00	32	32	1.00	32	32	1.00
H2.α2.γ1	0	2	0.00	0	2	0.00	0	2	0.00	0	2	0.00
H2.α2.γ2	20	20	1.00	30	30	1.00	30	30	1.00	30	30	1.00

Supplementary Table 15. Imputation panel composition

The following table shows the composition of the two reference panels evaluated (the Illumina 1M and Affymetrix 6.0 panels used the same individuals as the HapMap3 reference panel).

Population	1000 Genomes Panel		Hapmap3 Panel	
	Samples from 1000G	Trio Founders	Samples from 1000G	Trio Founders
ASW	21	-	20	-
CEU	7	63	5	89
CHB	32	-	27	-
CHS	32	-	-	-
CLM	22	-	-	-
FIN	31	-	-	-
GBR	29	-	-	-
IBS	2	-	-	-
JPT	27	-	24	-
LWK	42	-	39	-
MXL	25	-	20	-
PUR	23	-	-	-
TSI	33	-	31	-
YRI	47	-	45	-
Total	373	63	211	89

Supplementary Table 16. Primer sequences

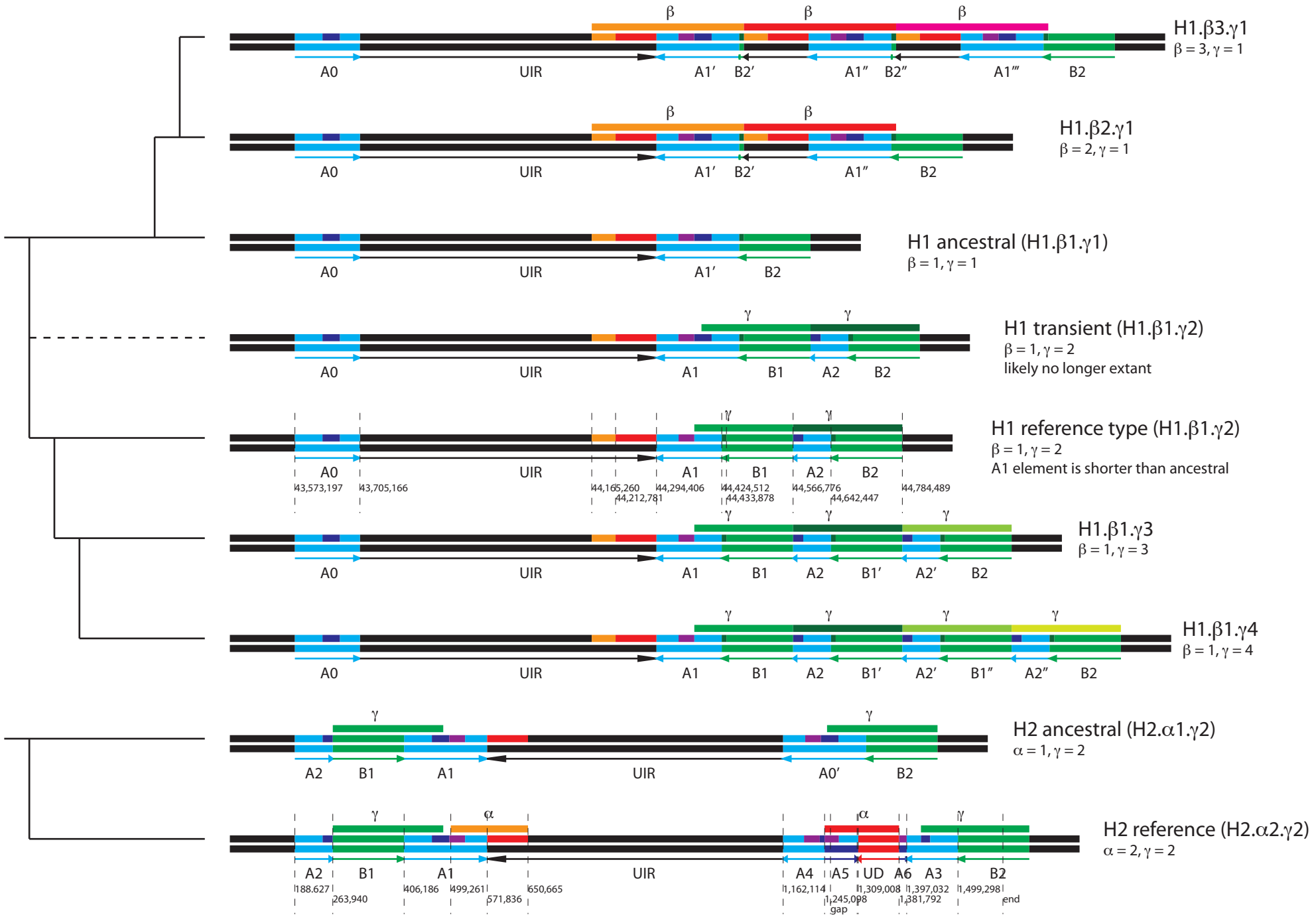
Assay	Primers (5' to 3')
Region 1	ACAGGAACCAGAGACCCAAG CACCCTCTCACCCCTCTACA Probe: FAM-CCAGCATGCACAGCAAAGTGCA
Region 2	CGGGCTGCTCACTATACCTC GTGATGGGAAAGGCTGTTGT Probe: FAM-AAAGCAAAAGGCCTGCCTATGCC
Region 3	GTTGTTGACCATGGCTTCCT GTGAGAAGACGGCCTTTGAG Probe: FAM-CACATGTGTTCTGGAATGCC
Genomic breakpoint: α Duplication	ATGAAATTATAGAGCAATTTGACAGG CGTTAATCTGGGTAAGATGGAGA
Genomic breakpoint: β Duplication	CCTGGCAACGTGGCTATT CATGTTGGGGTAGCATGACA
Transcript created by α duplication	GCAGCGTCATGTACCCCTAG CCCTGTGTTCCTCACCAAGT
Transcript created by β duplication	AGACGCAGGTCAGAATGGAAATGG TGCTGCCACAGAGGTCTGATTT

Supplementary Table 17. Population identifiers

CEU	Utah residents with Northern and Western European ancestry
FIN	Finnish from Finland
GBR	British from England and Scotland
TSI	Toscani in Italia
ASW	African Ancestry in Southwest US
LWK	Luhya in Webuye, Kenya
YRI	Yoruba in Ibadan, Nigeria
CHB	Han Chinese in Beijing, China
CHS	Han Chinese South
JPT	Japanese in Toyko, Japan
CLM	Colombian in Medellin, Colombia

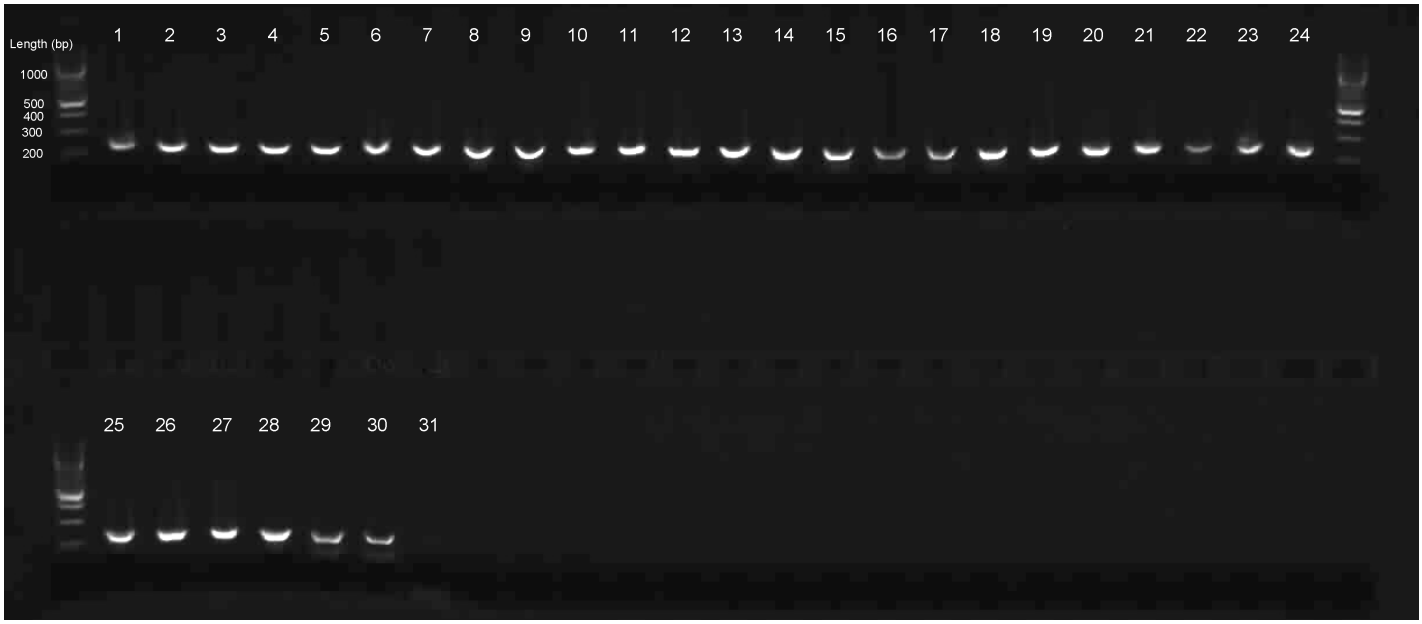
Supplementary Figures

Supplementary Figure 1. Detailed structure and breakpoints
Each set of lines represents one structural haplotype. The defined duplication breakpoints are marked on the H1 haplotype in GRCh37 coordinates and on the H2 reference haplotype in GRCh37 ALT_REF_LOCI_9 H2 assembly. The latter is also marked with the position of the gap in the GRCh37 H2 reference and the end of H2-specific haplotype assembly. Colors have different meanings on each line. Note that this color pattern is intentionally distinct from that used in the more schematic diagrams in the main body to make clear that there is not necessarily a base resolution correspondence between those elements and the slightly more complex high resolution structure. The bottom line for each haplotype shows the high level structure. Black bars represent sequences that are unique in at least one haplotype, cyan bars are copies of "A" duplicons (except on H2.α2.γ2, where some "A" duplicons are dark blue where they appear in tandem with different As), and green bars are "B" duplicons. The middle line for each haplotype represents detailed structures. Copy variant sequences that are part of both α and β duplications are shown in red (this part of the α duplication is depicted on H1 haplotypes even though it truly only appears on H2 haplotypes to indicate the orthologous nature of these duplications). Copy variant sequences that are part of the β duplication but not the α are shown in orange. The "A" duplicons are subdivided with dark blue representing sequences present in H1A0 but not H1A1 and magenta sequences present in H1A1 but not H1A0 (scheme maintained through the A duplicons in all haplotypes). The "B" duplicons are subdivided with a darker green depicting the portion of B overlapped by the β duplication. The top line shows the extent of the α (red shades), β (red shades), and γ (green shades) duplication sequences (only depicted on chromosomes with more than one copy). Details of the haplotypes are described in Supplementary Note S11.

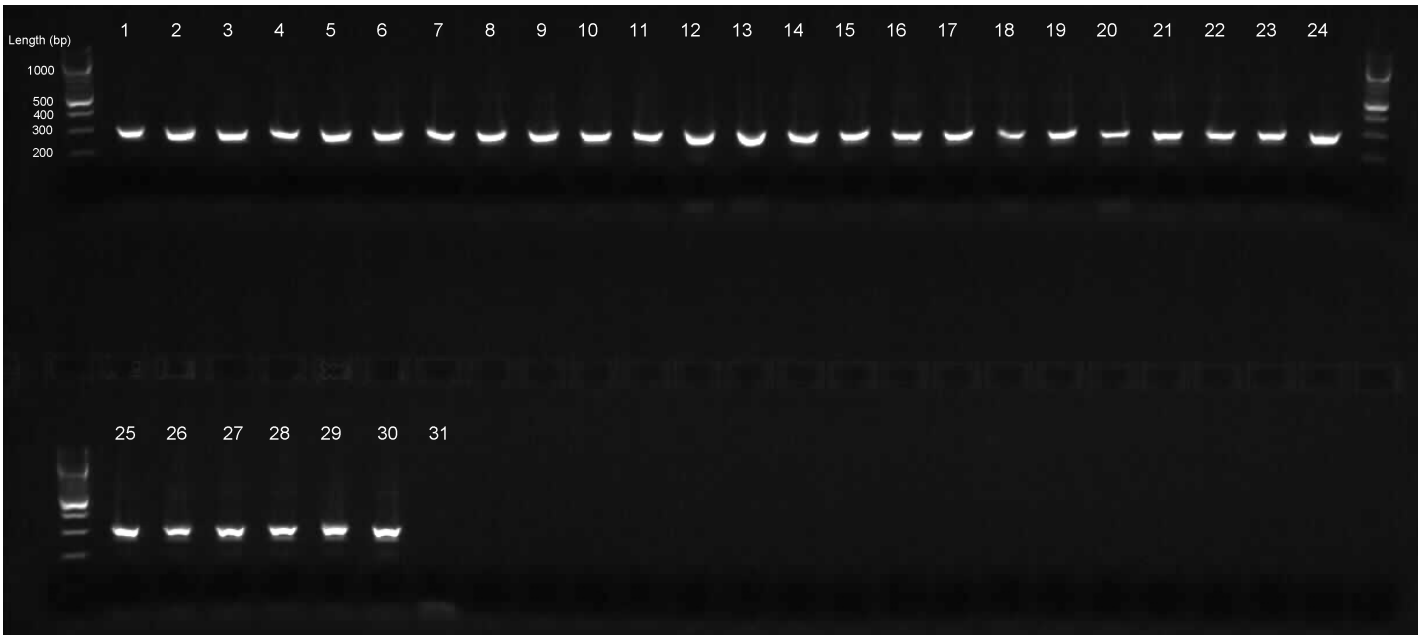


Supplementary Figure 1

Supplementary Figure 2. Verifying consistent breakpoints for duplications α and β across individuals



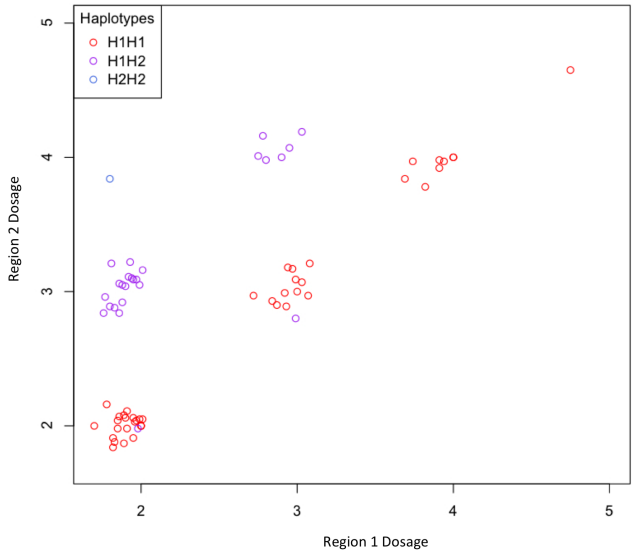
Supplementary Figure 2a. The proximal breakpoint of duplication α was PCR amplified using DNA from 30 unrelated CEU individuals who carry at least one copy of the α duplication (lanes 1-30). A CEU individual lacking the α duplication with haplotypes H1.β2.γ1 and H2.α1.γ2 was used as a negative control (lane 31). All reactions excluding the negative control generated the expected product size of 244 bp. See Supplementary Table 16 for primer sequences.



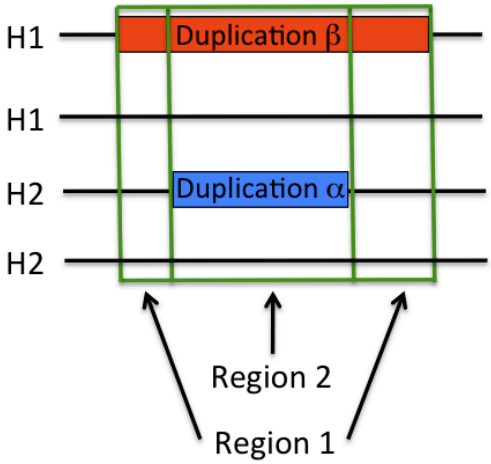
Supplementary Figure 2b. The proximal breakpoint of duplication β was PCR amplified using DNA from 30 unrelated CEU individuals who carry at least one copy of the β duplication (lanes 1-30). A CEU individual lacking the β duplication with haplotypes H1.β1.γ3 and H2.α2.γ2 was used as a negative control (lane 31). All reactions excluding the negative control generated the expected product size of 307 bp. See Supplementary Table 16 for primer sequences.

All PCR reactions were visualized on an agarose gel with the New England Biolabs 100bp ladder.

Supplementary Figure 3. Relating copy number to duplications



a.

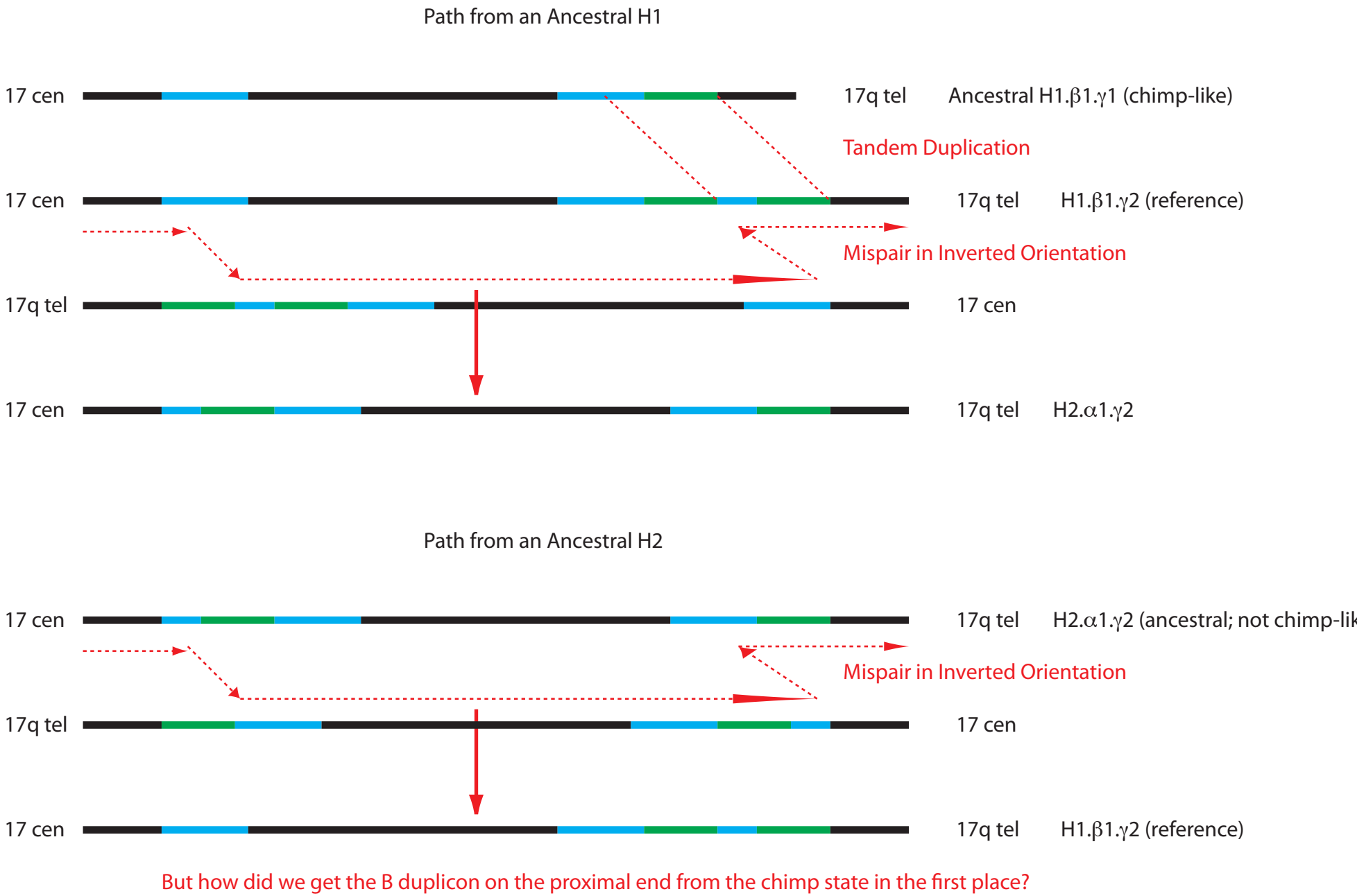


b.

Comparison of copy number in Region 1 and Region 2 in the context of H1/H2 haplotype orientation makes it clear that copy number is influenced by two separate duplication polymorphisms and determines on which haplotype each resides. Three main observations underlie this determination. (1) Copy number of Region 1 was always equal to copy number of Region 2 in H1 homozygotes. (2) Copy number in Region 2 was (in all but two cases, described below) equal to the sum of (i) copy number of Region 1 and (ii) the number of H2 haplotypes that an individual carries. This indicated that a duplication of Region 2 is present on most H2 haplotypes. (3) Copy number in Region 1 was only greater than two in individuals with at least one H1. These data indicated the existence of a long duplication which overlaps Region 1 and Region 2 segregating on the H1 background, and a shorter duplication overlapping Region 2 segregating on the H2 background.

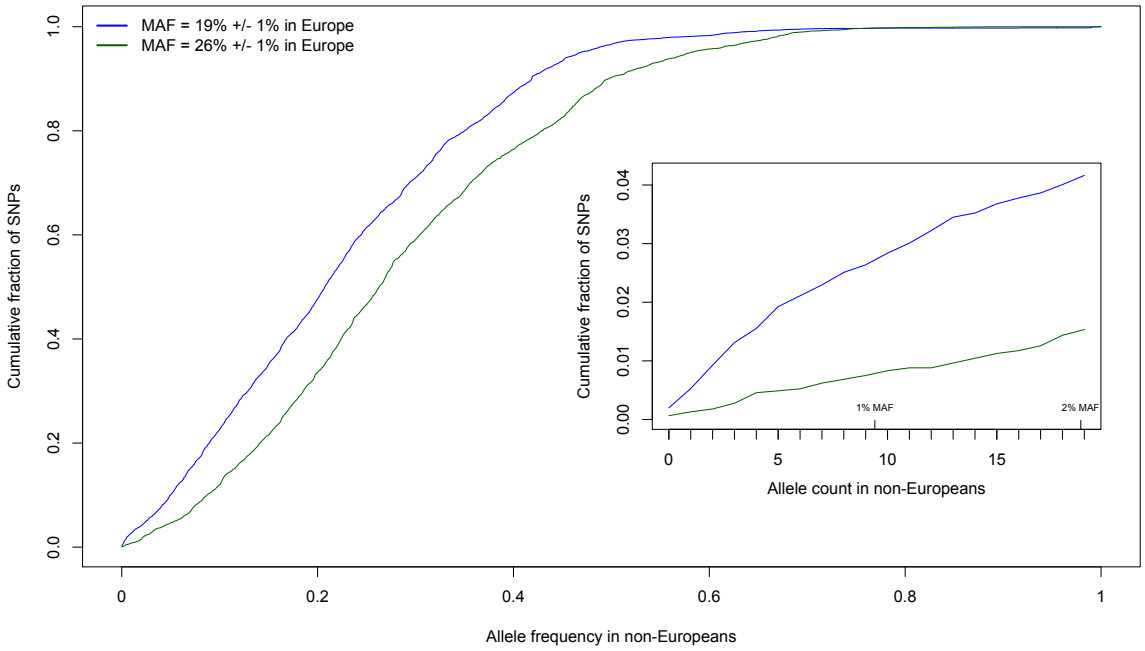
Supplementary Figure 4. Ancestral paths from different haplotypes

In an ancestral population consisting of only H1 chromosomes, we can generate a pattern of non-allelic homologous recombination involving reciprocal crossovers through two inversely paired H1.β1.γ2 chromosome (the transient type shown in Supplementary Figure 1) that would create an H2.α1.γ2 chromosome (top panel). We can also generate a path through two inversely paired H2.α1.γ2 chromosome (bottom panel) that would generate either an H1.β1.γ2 chromosome (depicted) or an H1.β1.γ1 chromosome (by reciprocal crossing in the first "A" duplicon, not shown). However, this requires extra steps to (1) generate the dispersed "B" duplicons on the H2.α1.γ2 without an inversion, and (2) either collapse the γ duplication on the H1.β1.γ2 to create the observed H1.β1.γ1 or reduplicate γ on the H1.β1.γ1 (identically to H2.α1.γ2) to generate H1.β1.γ2. Based on this, an H1 ancestral sequence appears to require at least two fewer events to generate all the observed modern haplotypes (Supplementary Figure 1) than an ancestral H2.



Supplementary Figure 4

Supplementary Figure 5. SNPs with highly differentiated allele frequencies between European and non-European populations in 1000 Genomes phase 1



Minor allele frequency distribution of SNPs in non-Europeans (n=471, populations CHB, CHS, JPT, LWK and YRI) with MAF of 18% - 20% (blue) and 25% - 27% in Europeans (n=379, populations CEU, FIN, GBR, IBS and TSI), corresponding to the allele frequencies we observed for the alpha and beta duplications in CEU. Inset shows the low frequency portion of the same distribution by allele count in the non-European populations. The SNPs were ascertained and genotyped in Phase 1 of the 1000 Genomes Project on chromosome 17, excluding the region 17:43165000-45785000 (+/- 1Mb from the inversion region).

Supplementary Figure 6. Fusion transcripts created from *KANSL1* duplications (α and β)

a. Fusion mRNA created from β duplication breakpoint

TGAGACGCAGGTCAGAATGGAAATGGGCTGCAGACCGGGCAGCTATTGTCAGCCG
CTGGAACTGGCTTCAGGCTCATGTTTCTGACTTGGAATATCGAATTCGTCAGCAAAC
AGACATTTACAAACAGATACGTGCTAATAAGGTTTCTGTGTGGAGACAGTAGAATAT
AAAAATAACACCTTCGCT

KANSL1 (shown in blue) is fused *ARL17* (shown in red). This sequenced breakpoint is also present in cDNA BC006271, likely a complete transcript of this fusion mRNA.

b. Fusion mRNA created from α duplication breakpoint

TACCCCTAGACGTGGGAACAACGCAAGTCCCACTTACAACACTTAAGAACATTCTC
ATGATGACCGTTGAACTGGAAAAACTTCCCAGCAGACCACAGGAGGTTGGCCCCA
GACTCACTGAGTGCCTGCAGCAGCCGTACAGACACAGCATCCTTGGCCACCTCAT
GCCCATCCCGGCCATCTAGGGTCAGCACAACCCAGATGAGGCCGCTGAAGGGCAC
CGGATGCCCAGGAATCACCACCTGGTACCAGAAGCGGTGCCAGCCAGCAGGTCCT
ATGCCCAAACACTTGGTGAGG

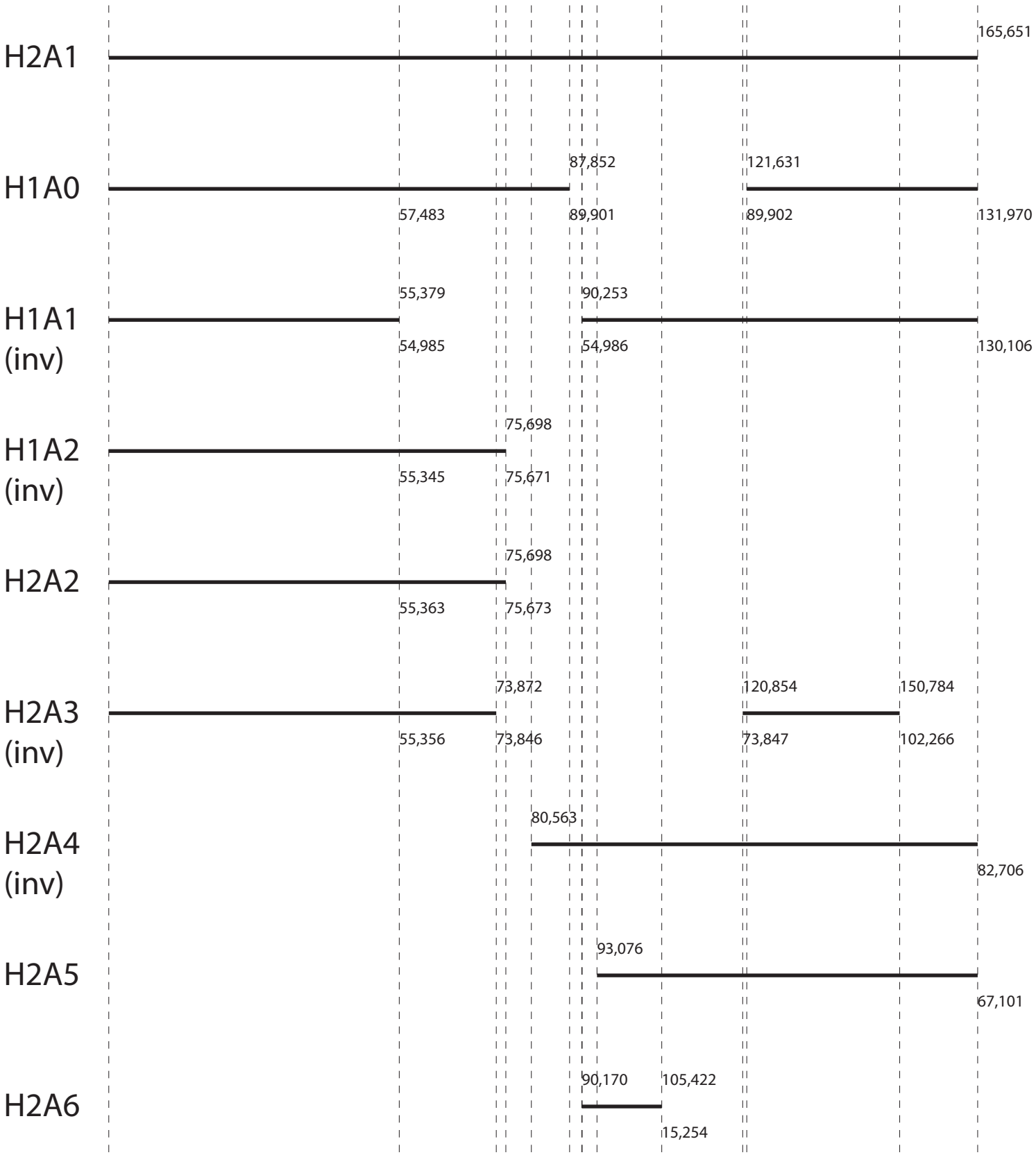
KANSL1 (shown in blue) is fused to *LRRC37A* (shown in green), and yet another fusion occurs with a novel exon (shown in orange).

Protein Domains

KANSL1 contains a coiled-coil domain and a PEHE domain, which is known to directly interact with MOF histone acetyltransferase[14]. Both fusion transcripts retain the coiled-coil domain but lack the PEHE domain. A protein translated from these fusion transcripts could therefore in principle compete with KANSL1 for some protein-protein interactions without recruiting MOF to the resulting complexes.

Supplementary Figure 7. Large structural differences between "A" duplicon copies in the two reference haplotypes

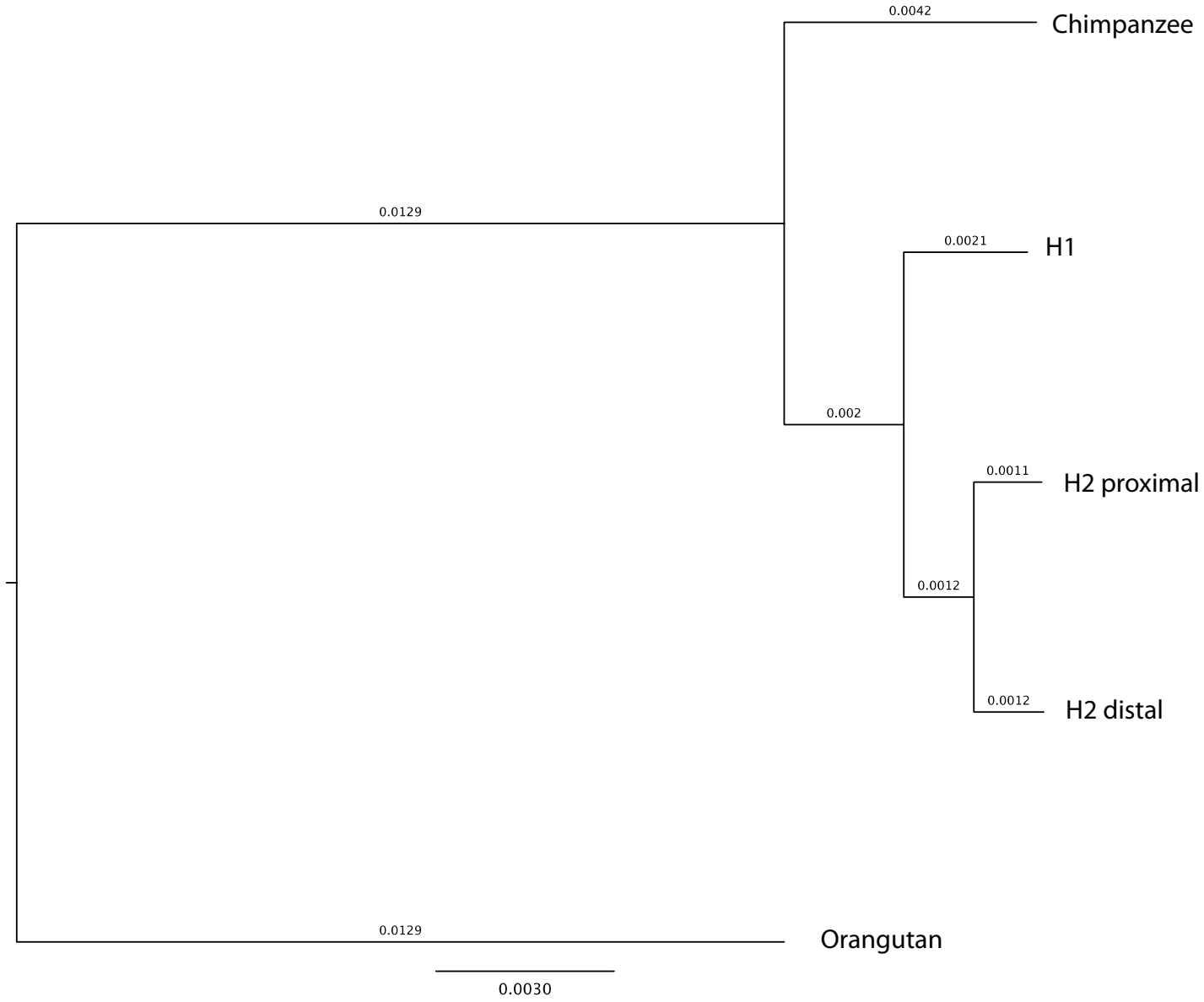
Diagram of all the "A" duplicons on the reference H1.β1.γ2/ H2.α2.γ2 genome. The H2A1 copy is nearly full-length ancestral, and all others are basically subsets of it. There is no part of H2A1 not present in at least one other copy. (Changes smaller than 4 kb are not depicted.) Numbers above the lines are H2A1 coordinates, below are the copy-specific coordinates. All copies start at 1 and end at the right end of H2A1 unless otherwise marked. Coordinates for vertical dotted lines corresponding to right ends of segments in any copy give the coordinate of the last base of the preceding segment, while those for lines corresponding to left ends give the first base of the following segment. Segments that are in inverted orientation on their respective reference haplotype are shown in the H2A1 orientation and marked with (inv). H2A4 and H2A5 are unfinished sequence. The remainder are finished in the GRCh37 reference.



Supplementary Figure 7

Supplementary Figure 8. Phylogenetic tree of the "unique" (non-"A") portion of the α duplicon on H2

Phylogenetic tree constructed by RAxML using GTRGAMMA model with the unique orangutan sequence[4] as an outgroup. With 100 bootstraps, the human grouping is supported is 96% and the H2 at 94%.



Supplementary Figure 8

References

1. DePristo, M.A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491–498 (2011).
2. Handsaker, R.E., Korn, J.M., Nemesh, J. & Mccarroll, S.A. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat Genet* **43**, 269–276 (2011).
3. Montgomery, S.B. et al. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773–777 (2010).
4. Zody, M.C. et al. Evolutionary toggling of the MAPT 17q21.31 inversion region. *Nat Genet* **40**, 1076–1083 (2008).
5. Hardy, J. et al. Evidence suggesting that Homo neanderthalensis contributed the H2 MAPT haplotype to Homo sapiens. *Biochem. Soc. Trans.* **33**, 582–585 (2005).
6. Stefansson, H. et al. A common inversion under selection in Europeans. *Nat Genet* **37**, 129–137 (2005).
7. Green, R.E. et al. A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
8. Reich, D. et al. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053–1060 (2010).
9. Skaletsky, H. et al. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**, 825–837 (2003).
10. Rozen, S. et al. Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature* **423**, 873–876 (2003).
11. Yunis, J.J. & Prakash, O. The origin of man: a chromosomal pictorial legacy. *Science* **215**, 1525–1530 (1982).
12. Lee, J.A., Carvalho, C.M.B. & Lupski, J.R. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell* **131**, 1235–1247 (2007).
13. Hastings, P.J., Ira, G. & Lupski, J.R. A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS Genet* **5**, e1000327 (2009).
14. Li, X., Wu, L., Corsa, C.A.S., Kunkel, S. & Dou, Y. Two mammalian MOF complexes regulate transcription activation by distinct mechanisms. *Molecular cell* **36**, 290–301 (2009).