Supplemental Material

Interchromosomal core duplicons drive both evolutionary and disease instability of the Chromosome 8p23.1 region

Kiana Mohajeri[1†], Stuart Cantsilieris[1†], John Huddleston[1,2], Bradley J. Nelson[1], Bradley P. Coe[1], Catarina D. Campbell[1], Carl Baker[1], Lana Harshman[1], Katherine M. Munson[1], Zev N. Kronenberg[1], Milinn Kremitzki[4], Archana Raja[1,2], Claudia Rita Catacchio[3], Tina A. Graves[4], Richard K. Wilson[4], Mario Ventura[3], Evan E. Eichler[1,2]

†These authors contributed equally to this work.

[1]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, 98195, USA

[2]Howard Hughes Medical Institute, University of Washington, Seattle, WA, 98195, USA

[3]Dipartimento di Biologia, Università degli Studi di Bari "Aldo Moro," Bari 70125, Italy

[4]The McDonnell Genome Institute at Washington University, Washington University School of Medicine, St Louis, MO 63108, USA

Contents

Key words and associated acronyms

Bacterial artificial chromosome (BAC)

Comparative genomic hybridization (CGH)

DA duplication (meaning “large” in Chinese)

Distal repeat unit (REPD)

Expressed sequence tag (EST)

Extended haplotype homozygosity (EHH)

Fluorescence *in situ* hybridization (FISH)

Fosmid end sequences (FES)

Integrated haplotype homozygosity (iHS)

Inversion-associated repeat (IAR)

Million years ago (mya)

Multiple sequence alignment (MSA)

Non-allelic homologous recombination (NAHR)

Observed heterozygosity (oHET)

Open reading frame (ORF)

Proximal repeat unit (REPP)

Segmental duplications (SDs)

Single-molecule, real-time (SMRT) sequencing

Single-nucleotide polymorphism (SNP)

Single-nucleotide variant (SNV)

Singly unique nucleotide k-mers (SUNKs)

Thousand years ago (kya)

Whole-genome shotgun sequence detection (WSSD)

Whole-genome shotgun sequencing (WGS)

Xiao duplication (meaning “small” in Chinese”)

Supplemental Section 1: Sequence, assembly and validation of the Chromosome 8p23.1 inverted haplotype.

Supplemental Section 1.1: Chromosome 8p23.1 inversion structure and sequencing

Given that the Chromosome (Chr) 8p23.1 locus contains a complex architecture of segmental duplications (SDs) and such regions frequently contain sequence that is missing or misassembled in the human reference assembly, we established a high-quality alternate reference for this 6.3 Mbp (mega-base pair) region of Chromosome 8p23.1. We used large-insert clones from the CHORI-17 (CH17) bacterial artificial clone (BAC) library created from a hydatidiform (haploid) mole-derived human cell line, CHM1hTERT. Using a combination of end sequence mapping and BAC filter hybridization protocols (http://bacpac.chori.org/highdensity.htm), we selected a total of 166 CH17 BAC clones for Nextera library preparation and Illumina-based short-read sequencing. We mapped sequence data focusing on SUNKs (singly unique nucleotide k-mers)[1] from these clone inserts to the GRCh37 reference using mrsFAST[2] and created a sequence tiling path across the Chromosome 8p23.1 locus. The short-read-based BAC mappings were utilized in the selection of clones for PacBio single-molecule, real-time (SMRT) sequencing.

Supplemental Section 1.2: BAC assembly and assessment of sequence read depth

We sequenced and assembled 68 large-insert clones (BACs) using PacBio SMRT sequencing (average 550X coverage), and included three additional clones from the NCBI clone repository (http://www.ncbi.nlm.nih.gov/clone/) into our tiling path (71 CH17 BAC clones; Supplemental Table 1). We generated a 6.3 Mbp high-quality alternate reference, assembled using a hierarchical clone-based sequencing approach. To mitigate the challenges associated with sequence and assembly of large blocks of highly identical SDs, we anchored our assembly in 350 kbp (kilo-base pair) of unique sequence flanking the distal and proximal SD clusters (REPD and REPP). To correctly merge allelic overlaps between BACs, we applied a stringent cutoff, requiring a minimum of 20 kbp sequence overlap at 99.98% sequence identity. We then used sequence read depth to identify potential regions of collapsed duplication within individual BAC assemblies. We identified two such regions within the CH17 BACs. The first was a ~15 kbp collapsed triplication at REPD containing *LINC00965* (GRCh37 Chr8:7066083-7283257), which remained unresolved in four BAC clones represented in the CHM1 assembly (CH17-367D14, CH17-195A22, CH17-183F15 and CH17-248K13). The second is an 80 kbp collapsed

duplication identified in clones CH17-20H22 and CH17-28E6, which we were unable to completely resolve. These clones map in close proximity to the distal gap at REPD in GRCh37 (Chr8:7401944-7474649). To assemble these clones into a complete tiling path, we lowered our stringency cutoff to include an overlap of 19.7 kbp at 99.93% sequence identity. We allowed an additional position of lowered stringency for two clones, CH17-189F11 and CH17-221K15, that contain a smaller region of overlap (1154 base pair (bp)) at 100% sequence identity, given that these clones mapped completely to a unique segment within the assembly (GRCh37 Chr8:11354264-11458221).

Supplemental Section 1.3: BAC end sequence validation

We validated the quality of our CHM1 alternate reference assembly by mapping publically available BAC end sequences (BES) from the CH17 BAC library (NCBI trace repository (http://www.ncbi.nlm.nih.gov/Traces)[3,4]. We restricted our analysis to high-quality bases in these alignments (Phred quality >30) and identified a total of 249 concordant end sequence pairs (220,534/220,560 bp; 99.99% identity), 6 discordant pairs (6,535/6,535 bp; 100% identity) and 22 single-end mapping clones (9,215/9,216 bp; 99.99% identity) that spanned contiguously across the 6.3 Mbp contig (Supplemental Table 2). We used a combination of SMRT sequencing and publically available fingerprint data (http://www.ncbi.nlm.nih.gov/clone/) to explain the six discordantly mapped clones. Fingerprint data supports that five of the six discordant clones (CH17-189E11, CH17-231D4, CH17-418E9, CH17-91J1 and CH17-17P22) contain smaller than average inserts, ranging from 8-95 kbp. SMRT sequencing of CH17-189E11, CH17-231D4 and CH17-418E9 as well as a working draft assembly of clone CH17-91J1 (AC244426.1) corroborated the insert sizes based on fingerprinting. Of the 22 clones with single-end mappings, 15 mapped within the first or last ~200 kbp of the assembly in a manner consistent with clones that would tile concordantly from the CHM1 assembly into the adjacent regions of the genome. Of the seven remaining singletons, one concordantly aligned to Chromosome X and two were previously identified as transchromosomal pairs.

To further validate our assembly using BES, we lowered the threshold for our mapping quality to include BACs that mapped with >98% but <100% sequence identity. We identified an additional 24 BAC clones that mapped discordantly to our CHM1 assembly and selected three of these clones (CH17-352N3, CH17-349F14 and CH17-257P21) for SMRT sequencing. These three BACs were assembled with uniform read depth and the appropriate BAC ends were observed in

each final assembly. While each insert contains sequence homology with the Chromosome 8p23.1 locus, these clones map with highest sequence identity (>99.9%) to Chromosomes 12, 4 and 2 (GRCh37 Chr12:8316415-8529206, Chr4:48850273-49054096 and Chr2:182845097-183060106). We found that the overall identity for CH17 end sequence mappings was 99.99% across >230 kbp (236,284/236,311 bp) of high-quality bases with a tight insert-size distribution of 209 ± 19 kbp.

Supplemental Section 1.4: Fosmid end sequence validation

In order to detect sequence discordance on a much finer scale[5], we mapped fosmid end sequences (FES) to our CHM1 assembly from four HapMap individuals. We focused on fosmid libraries that are homozygous for the direct (NA18956 and NA18947) and inverted (NA19240 and NA12878) haplotypes of the 4.2 Mbp inversion at the Chromosome 8p23.1 locus[6] (Supplemental Table 2). We found that regardless of inversion status, the sequence identity for >1.3 Mbp of high-quality end sequence mappings exceeded 99.99% for three of the four libraries (NA19240, NA12878 and NA18956). Additionally, end sequence mappings from NA18947 also demonstrated >99.99% sequence identity for >650 kbp of high-quality bases, the discrepancy in total number of bases pairs likely the result of lower coverage for this library. It is expected that inversion signatures identified by FES mapping are likely to cluster directly at the inversion breakpoints, and thus, we observe this effect in the NA18956 and NA18947 libraries. Cumulatively, we find that the individuals homozygous for direct haplotype contained ~38% (98 vs. 61) more discordant FES pairs, ~37% less concordant FES pairs (2917 vs. 1843) and ~43% more singletons (78 vs. 37), in comparison to those homozygous for the inverted haplotype (Supplemental Table 2).

Supplemental Section 1.5: Fosmid sequencing and validation

In order to further assess the validity of the H2 assembly and better characterize structurally variable sites at Chromosome 8p23.1, we targeted a subset of discordantly mapped clones from the NA19240 and NA12878 fosmid libraries for SMRT sequencing (Supplemental Table 1). We selected clones based on observed discordant end sequence pileups mapped against the H2 assembly. In total, we sequenced and assembled 24 fosmids, mapped them to both the H1 and H2 haplotypes using BLAST[7], and visualized the alignment using Miropeats[8]. In cases where we identified multiple, high-identity mapping locations at Chromosome 8p23.1, we performed a

series of MSAs (multiple sequence alignments) and calculated percent sequence identity in 2 kbp sliding windows across the alignment to assign the correct mapping location (Nuttle et al. unpublished). We assigned fosmid sequences to three categories (mismapping of FES, structural variation and misassembly) based on the most parsimonious explanation for fosmid discordancy. In total, we identified 8 clones that represented examples of inter/intrachromosomal FES mismapping, 14 clones exhibiting evidence of structural variation, and 2 clones confirming sites of misassembly in the H2 haplotype. Structurally variant sites included a ~19 kbp deletion encompassing the alpha-defensin cluster, consistent with a single-copy deletion of one of the alpha-defensin repeat arrays (GRCh37 Chr8:6859071-6878169)[9,10], a ~7 kbp deletion that removes the distal copy of the MIR548I3 microRNA (GRCh37 Chr8:7055962-7056110) and a ~24 kbp deletion consistent with the loss of four copies of the ~6 kbp tandem repeat array, distal to the beta-defensin cluster of genes (GRCh37 Chr8:7107657-7171375). Additionally, we identified three clones that mapped to a ~160 kbp region that lacks sequence contiguity between the H1 and H2 assemblies (GRCh37 Chr8:7,881,469-8,038,443). To confirm that this was not a result of sequence collapse within the H2 haplotype, we examined the read-depth profiles for three CH17 clones (CH17-10F5, CH17-268K16 and CH17-251L15) tiling this region. Each clone exhibited uniform sequence coverage, indicating that misassembly was not the cause of the fosmid discordance and was likely a structurally variant site between the two haplotypes.

Our sequence read-depth analysis identified a 15 kbp collapsed triplication containing the *LINC00965* genes in four CH17 BAC clones used in the final H2 Chromosome 8p23.1 assembly. We were able to resolve this collapse using two fosmid clones from the NA19240 library (ABC10_000044638200_G16 and ABC10_000044516000_K3), confirming a misassembly in the H2 haplotype.

Supplemental Section 1.6: Creating a restriction map using BioNano Genomics

To further validate the CHM1 assembly, we obtained lymphoid-cells from CHM1 and NA19240 source material and constructed a BioNano Genomics fingerprint map using the H2 assembly as a reference. CHM1 cells were pelleted and washed with Life Technologies PBS (phosphate buffered saline) at 1X concentration; the final cell pellet was re-suspended in cell suspension buffer using the Bio-Rad CHEF Mammalian Genomic DNA Plug Kit. Cells were then embedded in Bio-Rad CleanCut™ low-melt agarose and spread into a thin layer on a custom support (in development at BioNano Genomics). Cells were lysed using BioNano Genomics IrysPrep®

Lysis Buffer, protease treated with QIAGEN Puregene Proteinase K, followed by brief washing in Tris with 50 mM EDTA and then washing in Tris with 1mM EDTA before RNase treatment with Qiagen Puregene RNase. DNA was then equilibrated in Tris with 50 mM EDTA and incubated overnight at 4°C before extensive washing in Tris with 0.1 mM EDTA followed by equilibration in New England BioLabs (NEB) NEBuffer 3 at 1X concentration. Purified DNA in the thin layer agarose was labeled following the BioNano Genomics IrysPrep® Reagent Kit protocol with adaptations for labeling in agarose. Briefly, 1.25 ug of DNA was digested with 0.7 units of Nt.BspQI nicking endonuclease in NEB NEBuffer 3 for 130 minutes at 37°C, then washed with Affymetrix TE Low EDTA Buffer, pH 8.0, followed by equilibration with New England BioLabs 1x ThermoPol® Reaction Buffer. Nick-digested DNA was then incubated for 70 minutes at 50°C using BioNano Genomics IrysPrep® labeling mix and NEB Taq DNA Polymerase at a final concentration of 0.4 U/µl. Nick-labeled DNA was then incubated for 40 minutes at 37°C using BioNano Genomics IrysPrep® Repair mix and NEB Taq DNA Ligase at a final concentration of 1 U/µl. Labeled-repaired DNA was then recovered from the thin layer agarose by digesting with GELase™ and counterstained with BioNano Genomics IrysPrep® DNA Stain prior to data collection on the Irys system. The H2 assembly was subjected to *in silico* nicking using BspQI to produce a cmap. This cmap was aligned to the CHM1 BioNano map using the comparison function in IrysView 2.3. The CHM1 map produced by the BioNano genome mapping approach was consistent with the predicted *in silico* reference map, with the exception of two regions containing collapsed duplications described above (Supplemental Fig. 1). In combination with the FES and BES mapping data this strongly supported the order and orientation of the CHM1 Chromosome 8p23.1 assembly.

Supplemental Section 1.7: Sequence comparison between CHM1 and GRCh37

To identify structural variation between our CHM1 Chromosome 8p23.1 assembly and the GRCh37 reference, we constructed a visual alignment of the locus using Miropeats[8] (Fig. 1). We differentiated allelic positions and highly identical paralogous duplications through the construction of MSAs using MAFFT[11] and assessing both percent identity and contiguity of sequence. The resulting allelic positions determined through high-confidence alignments were then highlighted using Miropeats (Fig. 1). We identified five large structural differences between our CHM1 assembly and the GRCh37 reference that included three separate inversions and two large deletions relative to H1 (Table 1). We identified the known, cytogenetically visible 4.2

Mbp polymorphic inversion corresponding to Chr8:7920506-12141854 in the GRCh37 reference assembly (hereby referred to as inversion 1). As the human reference is in direct orientation, we refer to this as the H1 haplotype, with the inverted CHM1 assembly referred to as the H2 haplotype. Our analysis also identified two additional inversions: a ~320 kbp inversion corresponding Chr8:7120942-7441280 in GRCh37 and a ~335 kbp inversion completely housed within inversion 1, corresponding to Chr8:7524649-7881468 in GRCh37 (referred to as inversions 2 and 3, respectively). In addition to the three inversion events, we also characterized two large deletions of ~157 and ~136 kbp located in SD regions 2 and 4, respectively (Fig. 1, Table 1).

We called structural variants between the CHM1 assembly and GRCh37 reference with a modified version of the SMRT-SV caller[12]. First, we defined 10 pairs of homologous regions between CHM1 and GRCh37 based on pairwise alignments corresponding to 6.3 Mbp of CHM1 sequence and 6.5 Mbp of GRCh38 sequence. We realigned pairs of sequences from each region with BLASR, using parameters tuned for high-quality sequence alignments (-clipping subread -sam -bestn 1 -affineAlign -affineOpen 8 -affineExtend 0 -maxMatch 30 -sdpTupleSize 13) with CHM1 as the query and GRCh38 as the reference. We collated and sorted alignments from all 10 pairs and identified structural variants in the aligned sequences with the SMRT-SV's final variant calling step. Finally, we repeated the above analysis using GRCh38 as the query and CHM1 as the reference to confirm that we did not miss any variants from the initial analysis. When considering structural variation, we annotated differences in H2 that were relative to the H1 haplotype and identified 59 such events (50 bp – 22 kbp) (Supplemental Table 3).

The same alignments of allelic sequence between H1 and H2 were utilized to consider variation at the single-nucleotide level. We observed 544,609 bp of sequence in H2 that shares 100% identity with H1, the vast majority of which represented sequence contained in unique segments of the locus. We identified 5.3 Mbp (5,285,405 bp) of sequence that shares >98% sequence identity (but less than 100%) between H1 and H2 haplotypes and ~150 kbp that shares between 95-98% sequence identity. Only ~6.9 kbp shared less than 95% but greater than 90% identity, indicating that the majority of allelic positions between H1 and H2, when deletions are not considered, share a relatively high level of sequence identity. Finally, we identified regions of which there is no contiguous sequence match, including ~157 and ~135 kbp of sequence immediately flanking the proximal and distal gaps in the reference assembly (Table 1).

Supplemental Section 1.8: Segmental duplication architecture of the Chromosome 8p23.1 locus

In order to identify the organization of inter/intra chromosomal SDs within the H2 haplotype, we used a genome sequence alignment pipeline (whole-genome assembly comparison or WGAC) to identify all pairwise sequence alignments[13] ≥1 kbp in size and ≥90% sequence identity (Supplemental Table 5). We identified 615 interchromosomal SD pairs across 15 separate chromosomes and 132 intrachromosomal SD pairs localized to the Chromosome 8p23.1 region (Supplemental Table 5). We then filtered SDs in order to focus on those that were capable of mediating non-allelic homologous recombination (NAHR) (interspersed flanking the critical region and direct orientation) (Supplemental Fig. 2). Using these criteria, we identified that despite the GRCh37 H1 assembly containing a greater number of SD pairs (Supplemental Table 11), the alternative H2 assembly contained a greater proportion of large, highly identical SDs in direct orientation that flank the disease-associated critical region (Supplemental Table 13 and Supplemental Fig. 2). The largest SD (SD19), is ~385 kbp of >98% sequence identity and maps in direct orientation in the H2 assembly. This is consistent with previous observations that the REPD beta-defensin cassette is at least ~322 kbp[14]. Analysis of the H1 haplotype in the GRCh37 reference demonstrates that the largest SD (SD91) flanking the critical region is ~182 kbp of 97% sequence identity (Supplemental Fig. 2). Since orientation, length and sequence identity are deemed the most important parameters for NAHR[15], the current organization of the H1 haplotype would unlikely render susceptibility to recurrent microdeletion. However, the lack of contiguous sequence at both the distal (REPD) and proximal (REPP) SD clusters, combined with the misassembly of the proximal beta-defensin cluster 2[16], precludes a more detailed comparison of disease susceptibility risk between the two haplotypes.

Supplemental Section 1.9: Gene annotation of the Chromosome 8p23.1 H2 haplotype

To assess genic differences between the H1 and H2 haplotypes, we used the GMAP (genomic mapping alignment program)[17] gene annotation pipeline to map mRNA/EST RefSeq annotations (wget ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/refGene.txt.gz) to the CHM1 assembly using a minimum sequence identity of 0.98. We compared genes annotated in unique space (>3.3 Mbp) and found that the annotation is identical between the two haplotypes with the exception of the orientation (Supplemental Fig. 3). Comparison of genic content in duplicated space identified six differences between the two haplotypes (Supplemental Fig. 4), including two

separate deletion events that remove three genes in the H2 haplotype (a 157 kbp deletion (GRCh37 Chr8:7881469-8038443) removing *MIR548I3* and a 136 kbp deletion (GRCh37 Chr8:12141855-12277375) removing *DEFB130* and *FAM66A*). We identified an additional gene (*LOC729732*) located in the H2 assembly that maps to the distal gapped sequence present in the H1 haplotype (Supplemental Fig. 4, Supplemental Table 12). Finally, due to a 30 kbp collapsed duplication present in the H2 haplotype, we could only sequence-resolve one of the three tandem copies of the *LINC00965* genes.

Supplemental Section 2: Characterization of 8p23.1 inversions

Supplemental Section 2.1: Structure and breakpoint analysis of inversion 1

We compared sequences between the inverted haplotype (H2) and the directly orientated haplotype (H1) (GRCh37) and identified that the 4.2 Mbp inversion encompasses >40 genes and 3.6 Mbp of unique sequence (Supplemental Fig. 3). To refine the breakpoints associated with inversion 1, we generated a ~400 kbp MSA that represented sequences at least 20 kbp within or outside the inversion event. We aligned four sequences relating to the proximal and distal sides of the inversion event in both the H1 and H2 haplotypes. We annotated each column in the MSA as either (1) outside the inversion when the sequences from both haplotypes were identical at corresponding proximal and distal sides, (2) inside the inversion when proximal and distal sequences from different sides were identical, or (3) uninformative in all other cases. In the case of inversion 1, where the sequence from proximal side of the H1 haplotype included gap bases (Ns), only one pair of sequences was required to be identical and different from the remaining alternate sequence.

It is expected that the putative breakpoint signature within the inversion would include stretches of shared bases between the H2 distal SD3 and the H1 proximal SD3, which transitioned to stretches of shared bases between H2 proximal SD2 and H1 distal SD2. To objectively identify these transition regions, we performed a three-state Viterbi segmentation using HMMSeg[18] on the outside/inside inversion statuses for each base of the alignment with an additional state representing uninformative bases (Fig. 2A). We refined the breakpoint to a 79,704 bp region corresponding to GRCh37 Chr8:7920506-7998357 in the distal and GRCh37 Chr8:12091855-12141854 proximal SDs. The alignment identity between the breakpoint regions in the H2

haplotype was 91.8%. As tracks of perfect sequence identity between paralogs are commonly observed at the site NAHR mediated breakpoints[19], we determined the maximal length of perfect sequence identity in the MSA to be 334 bp with a total of 68,647 identical bases out of 405,566 non-gap bases (17%) across the alignment. Although the proximal breakpoint maps within a segment of missing sequence in the H1 haplotype and thus cannot be refined any further, it is noteworthy that the distal breakpoint maps within a class of repeat units that is present at three additional locations within the H2 assembly. Moreover, we identified that this repeat unit localizes at, or is within close proximity, to the breakpoints of all three inversion events. We call these sequences inversion-associated repeats (IARs).

Supplemental Section 2.2: Structure and breakpoint analysis of inversion 2

We identified a previously uncharacterized inversion (inversion 2) completely contained within the SD19 (Fig. 1). This 320.33 kbp event, inverts 15 genes relative to H1, including the distal β -defensin cluster (SD19) (Supplemental Fig. 4). We resolved the breakpoints of inversion 2, by using the same strategy as described for inversion 1, however, the availability of contiguous sequence in the H1 haplotype allowed us to map the breakpoint to a narrow sequence interval (449 bp), corresponding to GRCh37 Chr8:7120942-7121391 distal and GRCh37 Chr8:7440831-7441280 proximal (Fig. 2B). Sequence analysis identified that the breakpoint interval maps inside a high-identity (98.48%) 70 kbp SD, however we find that within the alignment itself, the breakpoint identity exceeds 99%. The maximal length of perfect sequence identity in the MSA was 521 bp with a total of 40,050 identical bases out of 113,497 bases (35%) in the alignment. We identified two IARs (IAR1 and IAR2) that flank inversion 2 and are located within ~33 kbp of the inversion breakpoint (Fig. 2B). Sequence characterization at the breakpoint interval demonstrates that it lies within a ~6 kbp higher-order tandem repeat unit consisting of a single ancient L1 (~33% divergence) and a more recent ERVK repeat (~8% divergence). We identified multiple copies of this repeat at both the distal and proximal breakpoints in the H1 haplotype (six and five copies, respectively).

Given that the GRCh37 reference assembly represents a mosaic haplotype at the inversion 2 region, we sequenced two additional RP-11 clones (RP11-957L3 and RP11-158L15) and created a contiguous 374 kbp haplotype (using RP11-623J22) that completely spanned the distal breakpoint region of inversion 2 (Supplemental Fig. 5A). Interestingly, comparison with the H2 CHM1 assembly revealed identical structure with respect to orientation, indicating that RP-11

has at least one haplotype representative of the inverted SD19 configuration (Inversion 2). We next reconstructed the clone overlaps of the inversion 2 region in the GRCh37 reference assembly (Supplemental Fig. 5B). We found that while this region contained clones representing both Caltech and RP-11 BAC libraries, the clone overlaps still exceeded 0.9996> sequence identity, suggesting that inversion 2 is polymorphic on both H1 and H2 haplotypes.

Supplemental Section 2.3: Sequence analysis of inversion 3

Alignment of the H1 and H2 haplotype assemblies using Miropeats also appeared to indicate the presence of an "inverted" inversion that maps to a 335 kbp region in SD2 (Fig. 1). Inversion 3 is located on the proximal side of the distal gap in the H1 assembly and spans until the end of the distal β-defensin cluster 2 in the GRCh37 reference (GRCh37 Chr8:7524649-7881468). We find that inversion 3 is in direct orientation for both the H1 and H2 haplotypes, consistent with this sequence being transposed twice; once by inversion 3 and the second time by inversion 1. A previous report demonstrated that the distal beta-defensin cluster 2 was misassembled within the human reference and should be positioned ~5 Mbp apart on the proximal side of the Chromosome 8p23.1 locus[16]. We therefore reasoned that inversion 3 was an assembly artifact in the GRCh37 reference. To investigate this region in more detail we analyzed the assembly quality of the GRCh37 reference by examining the clone overlaps of four RP11 clones extending from the proximal side of the distal gap (GRCh37 Chr8:7524650-8175046). We found that although the sequence identity of the RP11 clone overlaps was high >99.99, two of the three clone overlaps located in duplicated space were below 20 kbp, the length cutoff typically used to differentiate paralogous from allelic sequence matches[20].

Supplemental Section 2.4: Reconstruction of an incomplete RP-11 H1 assembly

Sequence reconstruction of the GRCh37 reference assembly revealed that at in least in one case, ~335 kbp of sequence (inversion 3) was misassembled into REPD at Chromosome 8p23.1. We therefore attempted to reconstruct a new H1 reference restricting clone selection to a single haplotype derived from one known reference genome (RP11). This involved the identification, sequence and assembly of four new RP11 BAC clone inserts and the use of 3 additional RP11 clones not currently present in the GRCh37 or GRCh38 reference assembly. Using this approach, we were able to construct five sequence contigs totaling almost 1.8 Mbp of sequence (Supplemental Fig. 6). While this alternate RP-11 H1 assembly represented an improvement in

contiguity over the current GRCh37 reference, (including the assignment of ~350 kbp of additional sequence to REPP) we were unable to completely close all sequence gaps in the assembly (Supplemental Fig. 6).

Supplemental Section 3: Identification and characterization of an evolutionary instability element

Supplemental Section 3.1: Evolutionary origin of Chromosome 8p23.1 inversions

In order to investigate the evolutionary history of Chromosome 8p23.1 inversion events, we selected chimpanzee, gorilla and orangutan BAC clones (from the CH251, CH277 and CH276 BAC libraries, respectively) based on end sequence mapping against the GRCh37 human reference assembly (NCBI trace repository: http://www.ncbi.nlm.nih.gov/Traces). Nonhuman primate BACs underwent Nextera library preparation and Illumina-based short-read sequencing as a selection filter prior to PacBio SMRT sequencing. We sequenced a total of 211 nonhuman primate BAC clones (54 chimpanzee, 43 orangutan and 114 gorilla) and after alignment to reference, selected 71 (16 chimpanzee, 21 orangutan, 34 gorilla) for high-quality sequencing and de novo assembly using SMRT sequencing (Supplemental Table 1). In order for finished BACs to be assembled into contigs, we required sequence overlaps of >20 kbp and >99.95% sequence identity; the reduced level of stringency was imposed given the potential for allelic variation within diploid clone libraries. We mapped our high-quality nonhuman primate assemblies to identify orthologous positions within the GRCh37 reference, then leveraged genetic distances to estimate the evolutionary time in which these inversions occurred[21].

Supplemental Section 3.2: Timing estimate for inversion 1

For inversion 1, we constructed two separate MSAs (MSA1 Chr8:11547336-11702776 and MSA2 Chr8:8191074-8316115) of 157 and 129 kbp (respectively) using MAFFT. These alignments represented sequences within inversion 1 that were shared between the H1 and H2 haplotypes, as the well as the orthologous region in the chimpanzee (CH251-553E24, CH251-35C5), orangutan (CH276-72N9, CH276-364C1) (Supplemental Fig. 7 and 8). From these MSAs we constructed an unrooted phylogenetic tree (MEGA5) using the neighbor-joining method and complete-deletion option[22]. Genetic distances were computed using the Kimura two-parameter method with standard error estimates and an interior branch test of phylogeny (n = 500 bootstrap

replicates). Using Tajima's relative rate test (MEGA5), we determined that these sequences evolved at the same rate as orthologous counterparts in the chimpanzee and orangutan (MSA1 PPY-H1-H2 p = 0.94, PTR-H1-H2, p = 0.89) (MSA2 PPY-H1-H2 p = 0.10600. PTR-H1-H2 p = 0.24493). Using the average sequence divergence of human versus chimpanzee distance (K), we calculated the average substitution rate using the equation R = K/2T assuming a chimpanzee–human divergence time (T) of 6 million years ago (mya)[23]. We then estimated the coalescence time of the haplotypes using the equation T = K/2R. Using the chimpanzee as an outgroup and the uncertainty in chimpanzee human divergence (6-7 mya), we calculated that the human H1 and H2 haplotypes diverged 0.37-0.43 ± 0.02 mya for MSA1 and 0.45-0.52 ± 0.03 mya for MSA2 (Supplemental Fig. 7 and 8), consistent with previous reports[24].

Supplemental Section 3.3: Timing estimate for inversion 2

To estimate the evolutionary timing of inversion 2, we constructed a ~85 kbp MSA (GRCh37 Chr8:7258093-7344068) using sequence shared between H1 and H2, including the orthologous sequence in gorilla (CH277-98D3) and the orangutan (CH276-501K8) (Supplemental Fig. 9). Using Tajima's relative rate test (MEGA5), we determined that these sequences evolved at the same rate as orthologous counterparts in the gorilla and orangutan (MSA PPY-H1-H2: p = 0.19, GGO-H1-H2: p = 0.14). Using the same strategy as described above, we estimate that the H1 and H2 haplotypes diverged between 0.55-0.69 ± 0.02 mya (Supplemental Fig. 9). It is likely that given the close approximation of timing estimates for inversions 1 and 2 that these inversions occurred in concert or arose over a very narrow evolutionary period (500-700 thousand years ago (kya)).

Supplemental Section 3.4: Inversion-associated repeats (IARs) localize at Chromosome 8p23.1 inversion breakpoints

We aligned the H1 and H2 haplotype assemblies using the program Miropeats and observed that the same paralogous segment of DNA appeared to be in close proximity to the boundaries of all three inversion events identified in the H2 haplotype (Fig. 1). We predicted that this sequence may be important in the context of mediating evolutionary inversions across the Chromosome 8p23.1 locus, and thus, sort to characterize this sequence firstly in the context of the H2 haplotype. Given their proximity to the Chromosome 8p23.1 inversion events, we refer to them as IARs. Using BLAST (http://blast.ncbi.nlm.nih.gov/), we identified four IAR locations in the

H2 haplotype and refer to them as IARs 1, 2, 3 and 4, based on their proximity to the distal end of the Chromosome 8p23.1 locus. We sought to define the greatest extent to which IARs shared sequence homology, so we constructed a ~100 kbp MSA and determined that all four IARs shared between 70-75 kbp of genomic sequence (referred to this as the "intermediate IAR" definition). A subset of the intermediate IARs were then extended, allowing us to define the boundaries of "maximum IAR" sequence shared between IARs 2, 3, and 4. We found that IAR 1 is truncated relative to IARs 2, 3 and 4 and is missing 41 kbp present at the beginning of each repeat element. Using the maximum IAR definition, we mapped our IAR sequences against our refined breakpoint intervals for inversions 1 and 2. We found that the breakpoint interval for inversion 1 is completely overlapped by IARs (IARs 2 and 4), both at the proximal and distal breakpoints (Fig. 2A). Additionally we find that IARs 1 and 2 directly flank the breakpoints of inversion 2, located approximately ~33 kbp from the proximal and distal breakpoints (Fig. 2B).

To confirm that IAR sequences were represented in a multi-copy state across Chromosome 8p23.1 and that this did not represent a misassembly artifact, we used BLAST to map the maximal definition of IAR sequences to the H1 haplotype. We identified four IAR mapping locations in the Chromosome 8p23.1 H1 haplotype, which we refer to as IARs A-D based on their distance from the distal end of the locus. Similar to H2, we found that IAR B maps at the distal breakpoint of inversion 1, IAR A maps in close proximity to the distal breakpoint of inversion 2 and that IAR C maps in close proximity to inversion 3. Given that IARs C and D map within close proximity to each other they were treated as a unit for the purposes of this analysis. In an effort to confirm the orientation and structure of IARs in both the H1 and H2 haplotypes, we used FES mappings from homozygous direct (NA18956 and NA18947) or inverted libraries (NA19240 and NA12878) to investigate mapping concordance at eight IAR locations (IARs 1-4 and IARs A-D). We counted the number of concordant end mappings that mapped to a single location and had a minimum sequence identity of 0.98. We found multiple instances of concordant support at all eight IAR locations including the strongest support in the CHM1 assembly.

Supplemental Section 3.5: Sequence characterization of IAR cores

Given that we identified multiple copies IAR sequences at Chromosome 8p23.1, we expanded our analysis of IARs to consider whether IARs were present within H1 and in alternate loci within the GRCh37 reference assembly. We used BLAST to map the maximum IAR consensus

sequence back to the human reference. Along with identifying three IAR positions within the H1 Chromosome 8p23.1 locus, we identified twelve additional IAR sequences distributed across six different chromosomes (3, 4, 7, 11, 12 and 16). By constructing a ~100 kbp MSA of 20 IAR sequences, (including sequences from CHM1 and the GRCh37 reference), we refined a 30-69 kbp interval representative of the "core" IAR sequence dispersed across seven human chromosomes. We find that all core IARs have a median size of 62 kbp and share a high level of sequence identity (95-97%).

Since core duplicons are frequently sites of gene innovation and are enriched for transcripts and in some case primate-specific genes[25], we sought to assess the sequence composition of the IAR core duplicons. First, we searched for the presence of expressed sequence tags (ESTs) overlapping IAR coordinates in the GRCh37 reference and identified that ~60% of IARs contain partial genic sequences or spliced ESTs. Using NCBI's ORF Finder (http://www.ncbi.nlm.nih.gov/gorf/gorf.html) we identified 23 potential open reading frames (ORFs) and used BLASTprotein to map potential ORFs to known protein sequences but found the majority (19/23) mapped with <50% sequence similarity. We identified four potential ORFs that demonstrated between 60-85% similarity with proteins, including olfactory receptor 4D10, solute carrier family 2-facilitated glucose transporter member 14, and a predicted G-protein coupled receptor. We assessed the repeat content of core IARs using RepeatMasker (http://www.repeatmasker.org/) and identified that these cores are enriched for interspersed repetitive elements including LINEs, SINEs, and LTR elements (Fig. 3A). We find that the repeat content varies between 59-80% in individual IARs, however, each core shares a repeat structure and organization that is shared across all IAR core duplicons. Notably, we identified a 5 kbp mosaic LINE/SINE element that exists in the same order and orientation in all IARs examined.

Supplemental Section 3.6: IAR cores localize at sites of evolutionary inversions

To further establish the role of IARs as potential inversion mediators, we sought to identify whether IAR core duplicons were present at the sites of evolutionary inversions in primate lineages. We mapped BES from the CH17, CH251, CH277 and CH276 BAC libraries to identify a clone-based framework of each genome. This approach identifies potential rearrangements based on discordant end sequence placements and allowed to us identify putative sites of chromosomal inversion[26]. We then overlaid these sites with known locations of IAR core

duplicons. Remarkably, we find that IARs are localized at the sites of four evolutionary inversions that have taken place on Chromosomes 11 and 3. On Chromosome 11, we identified IAR cores at 11q13.4 and 11p15.4, immediately flanking an inversion breakpoint that gave rise to the Chromosome 11 orientation in hominids after divergence from the orangutan (Supplemental Fig. 10). Similarly, on Chromosome 3 we identified three separate inversion events in which we find IAR cores. The first is localized at 3p12.3, the proximal breakpoint of a human-specific inversion and the distal breakpoint of an orangutan specific inversion. The second is located at 3q22.1, the distal breakpoint of the third inversion we also found specific to the human lineage.

We assessed the association between duplication blocks and breakpoints of evolutionary inversions (Antonacci F. and Ventura M. personal communication) by two different permutation tests (1,000,000 iterations per test). In the first test, we randomly distributed duplication blocks across their original chromosomes (excluding gaps) and measured the median distance between the midpoint of each randomly placed duplication block and the midpoint of the closest inversion breakpoint. In the second test, we randomly distributed duplication blocks within SDs on their original chromosomes (allowing at most 50% of each block to overlap non-duplicated space) and measured the same distance between duplication block and inversion breakpoint midpoints. The observed association between duplication blocks and evolutionary inversion breakpoints was significant by permutation across both the whole genome ($p < 0.000001$; Supplemental Fig. 11) and SDs ($p = 0.000078$; Supplemental Fig. 11). Next, we reran the original simulation excluding Chromosome 8p23.1 duplication blocks and inversions to remove any potential bias in the original simulation. The results remained significant (whole-genome significance is now $p = 0.000353$ and the duplications-only significance is $p = 0.00482$; (Supplemental Fig. 11).

We calculated the enrichment of IAR cores duplicons at sites of evolutionary inversion breakpoints by:

1. Calculating the proportion sequence mapped to evolutionary inversion breakpoints = 54 ape evolutionary inversion breakpoints have been narrowed within the limits of comparative FISH refinement to 172 Mbp of the human genome (~5.73% of genomic sequence).

2. Calculating the proportion of sequence represented by the 15 IAR interchromosomal duplication blocks = 5.911886 Mbp of sequence (or 5911886 / 30000000000 *100 = 0.197% of total genomic sequence).
3. Determining the proportion of IAR interchromosomal duplication blocks localized at evolutionary inversion breakpoints = 7/54 breakpoints (13.0%) or 6/27 of all inversions (22.2%).

Since 0.197% of the genome contains an IAR duplication block and breakpoints map to ~5.7% of the human genome, a random expectation would be 0.197% X 172 Mbp or 330 kbp of total breakpoint sequence. Instead, we observed 2.426137 Mbp of IAR duplication block containing sequence associated with breakpoints. We can therefore compute an enrichment of 2.426137/0.330 Mbp or a 7.35-fold enrichment of IAR block sequence associated with evolutionary inversions.

Supplemental Section 3.7: Evolutionary origin of IAR core duplicons

We attempted to reconstruct the evolutionary history of IAR core duplicons by performing a comparative sequence analysis in nonhuman primates that included representatives of Old World monkeys, New World monkeys and great apes. Our initial BLAST analysis suggested that the large scale expansion of these cores onto different chromosomes was specific to the ape lineage, with both the marmoset and macaque demonstrating only a single copy of an IAR core on Chromosome 16. To confirm that this was the case, and not indeed a misassembly artifact resulting in an under representation of IAR cores in these genomes, we performed a series of BAC hybridizations to the macaque (CH250) and gibbon (CH271) library filter sets using probes specific to human, marmoset and macaque core sequences. Using a combination of Illumina short-read sequencing and publically available end sequence mappings, we identified 16 gibbon and 6 macaque BAC clones that mapped within 100 kbp of a known human genomic IAR. Given that the BAC filter sets represent approximately 5X coverage of each genome, we conservatively estimated that the macaque had only a single copy of the IAR core duplicon, which was subsequently duplicated to two additional locations in the gibbon genome.

Supplemental Section 3.8: FISH analysis of IAR core duplicons

In order to confirm the ancestral location of the IAR core duplicon, we combined our sequence-based approach with cytogenetic-based assays. Metaphase spreads were obtained from

lymphoblast cell lines from one human individual (GM12878), one macaque (MMU, Macaca mulatta, 3238 from The Biomedical Primate Research Centre of Rijswijk) and one gibbon (Nomascus leucogenys). They were tested for the presence of the IAR core duplicon using a single-color metaphase FISH assay (Fig. 3B, Supplemental Fig. 12). FISH experiments were performed using gibbon BAC clones directly labeled by nick-translation with Cy3-dUTP (Perkin-Elmer) as described previously[27] with minor modifications. Briefly: 300 ng of labeled probe were used for the FISH experiments; hybridization was performed at 37°C in 2xSSC, 50% (v/v) formamide, 10% (w/v) dextran sulphate, 3μg human C0t-1 DNA, and 3 μg sonicated salmon sperm DNA, in a volume of 10 μL. Posthybridization washing was at high (60°C in 0.1xSSC, three times) or low (37°C in 2X SSC, 50% formamide, three times, and 42°C in 2X SSC, three times) stringency. Nuclei and chromosome metaphases were simultaneously DAPI stained. Digital images were obtained using a Leica epifluorescence microscope equipped with a cooled CCD camera. DAPI and Cy3 fluorescence signals, detected with specific filters, were recorded separately as gray-scale images. Pseudocoloring and merging of images were performed using Adobe Photoshop software.

Consistent with BAC hybridization data, FISH analysis confirmed marked variation in signal intensity, copy number and map location. In the macaque, only a single signal was identified on the distal end of Chromosome 20, a region partially syntenic to Chromosome 16p13.3 in humans (Fig. 3B). These data, in combination with our sequence-based approaches, strongly suggested that the Chromosome 16p13.3 IAR core was indeed the ancestral locus, the segment from which all other cores originated. In contrast to the macaque, FISH analysis in the gibbon identified two signals on Chromosomes 4 and 8, regions that are partially syntenic to Chromosomes 8p21.3-23.3 and 16p13.1-13.3 in humans (Fig. 3B). We compared this data with FISH analysis performed on human and found that the IAR core duplicon had expanded onto seven human chromosomes (3, 4, 7, 8, 11, 12, and 16) with the highest signal intensity observed at the Chromosome 8p23.1 locus (Fig. 3B). Interestingly, we find that a marked expansion of the IAR core duplicon occurred after the divergence of Old World monkeys to the ape lineage (Fig. 3A), which has arisen to particularly high copy among great apes.

Supplemental Section 3.9: Phylogenetic analysis of the IAR core duplicon

To study the evolutionary history of the IAR core duplicon, we constructed a 9 kbp MSA that included 19 human IAR core sequences (Chromosomes 3, 4, 7, 8, 11, 12, and 16) and the

ancestral core sequences from Chromosome 16 represented by the marmoset and the macaque assemblies (marmoset 2009 WUGSC 3.2/caljac3) (macaque 2006 MGSC merged 1.0/rheMac2). Notably, we removed sequences corresponding to one of the Chromosome 3q22.1 IARs that shared only 2 kbp with other interchromosomal IAR cores sequences. We constructed a neighbor-joining phylogenetic tree as described above, and after determining these sequences evolved at a neutral rate, estimated the coalescence of time for the expansion of IAR core duplicons (Fig. 3A, Supplemental Table 10). Based on the constructed phylogeny, the ancestral IAR is located on 16p13.3, recapitulating our FISH data. Using a macaque-human divergence time of 25 mya, we estimated that the ancestral core sequence on Chr16p13.3 is 19 ± 2.0 mya and duplicated to Chr4p16.1, 260 ± 690 thousand years later. We find that core IARs show increased levels of duplication activity at specific intervals in primate evolution (Supplemental Table 10). This initial duplication from Chr16p13.3 to Chr4p16.1, was followed by rapid expansion of IAR cores on to Chromosomes 11, 12 and 3 at a time when smaller apes were diverging to African great apes (Chr11: 12 ± 1.3 mya, Chr12: 12 ± 0.85, Chr3: 11 ± 0.72 mya). Further duplications onto Chromosomes 7 and 4 occurred 8.7 ± 0.57 mya. Interestingly, we find that the majority interchromosomal core events on Chromosomes 4, 11, 12, 3 and 7 share a more ancient coalescence (7-15 mya) (Fig. 3A), while the IAR cores on Chr 8p23.1 appeared to have arisen to in a very short evolutionary window 0.55-1 mya.

Supplemental Section 4: Construction and analysis of an alternative Chromosome 8p23.1 assembly in the orangutan.

Supplemental Section 4.1: Orangutan sequence and assembly of the REPP and REPD clusters at Chromosome 8p23.1

To investigate the ancestral organization of the Chromosome 8p23.1 region, we combined SMRT sequencing of large-insert clones with high-quality finished sequence from the CH276 BAC library (http://www.ncbi.nlm.nih.gov/clone/) to assemble the REPD and REPP regions in orangutan. To identify finished clones in GenBank, we utilized ~1.2 Mbp of sequence corresponding to the REPD and REPP clusters from our complete CHM1 assembly (including ~300 kbp of flanking unique sequence). We next sampled 1 kbp of non-repeatmasked sequence approximately every 50 kbp and searched GenBank by BLAST as previously described[20]. We

recovered a total of 19 finished CH276 BAC clones mapping to the Chromosome 8p23.1 region (Supplemental Table 1), including 12 that mapped to the REPD cluster and 7 mapping to the REPP cluster.

We mapped finished CH276 clones to the REPD CHM1 reference assembly and used end sequence mapping against the high-quality finished sequence to extend contigs. The result was four sequence contigs (416, 537, 255 and 352 kbp) totaling 1.56 Mbp anchored in at least 100 kbp of unique sequence (Fig. 4A). Sequence analysis demonstrated that the orangutan REPD cluster was expanded by ~200 kbp relative to the human CHM1 assembly in part due to the presence of lineage-specific duplications that included defensin genes. For example, the orangutan haplotype contained a ~80 kbp expansion of the alpha-defensin cluster which represented six full length copies of the ~19 kbp repeat array containing the *DEFA1* and *DEFT1P* genes. We extracted the sequences corresponding to the human *DEFA1* transcript (ENSE00001493017) and determined that all six copies maintained an ORF of 94 amino acids (AA), consistent with the canonical *DEFA1* transcript in humans. We performed the same analysis using the *DEFT1P* pseudogene (ENST00000453425) in humans and found that the orangutan *DEFT1* copies possessed an ORF in five out of six cases. Using 12 *DEFT1* sequences, including 3 copies representative from the rhesus macaque (CH250-361C3), 6 copies from the orangutan, and 3 copies from diverse humans (NA19240, NA12156, NA18942), we generated an MSA of the translated protein coding regions corresponding to *DEFT1* (Supplemental Fig. 13). Consistent with previous analysis[9] we found that all macaque and the majority of orangutan *DEFT1* copies were missing a stop codon corresponding to the 17th AA of the DEFT1 peptide. Notably, while all rhesus macaque copies contained an ORF of 75 AA, the orangutan *DEFT1* copies were highly variable ranging from 63 to 95 AA in length. In four out of five functional *DEFT1* copies in the orangutan, the first exon was highly conserved (consisting of the signal peptide and propiece) and the second exon was extremely divergent, which may indicate positive selection[28].

Like the human reference (H1), we determined that inversion 2 was in direct orientation in the orangutan however, the inversion spanned only ~120 kbp and encompassed 7 (beta defensins) of the 15 genes inverted in the human reference (Supplemental Fig. 14, Fig. 4A). On the proximal side of the inversion, the orangutan contained an additional seven copies of the higher-order ~6.1 kbp repeat unit where we had previously defined the human inversion 2 breakpoints. On the

distal side of inversion 2, orangutan is completely missing this higher order repeat structure (Supplemental Fig. 14). Sequence analysis of orangutan contig 4 revealed that it contained large blocks of SD that encompassed beta-defensin genes (Fig. 4A). The largest duplication spanned ~154 kbp of 99.1% sequence identity and contained three beta-defensin genes (*DEFB134*, *DEFB135* and *DEFB136*). On the distal side of this duplication, we identified an additional ~77 kbp duplication of 96.2% sequence identity that included *DEFB130* and *ZNF705D* (Fig. 4A). Despite the fact that *DEFB130* and *ZNF705D* are part of a larger 110 kbp interchromosomal duplication in the human CHM1 assembly (Chr4 and Chr11), the distal 4 orangutan contig contains ~23 kbp of unique sequence located at Chromosome 8p23.1, indicating this represented a true Chromosome 8p23.1 paralog. As the contig containing the *DEFB134-DEFB136* duplication is unanchored within our final CH276 assembly, we performed copy number variation analysis using 56 great ape genomes sequenced as part of the Great Ape Genome Project[29] on the 50 kbp segment containing *DEFB134*, *DEFB135* and *DEFB136*. Consistent with our CH276 assembly, we find that orangutan does indeed contain a duplication of the segment. Moreover the chimpanzee and bonobo also represent higher copy number states in comparison to human and gorilla, which are diploid at this locus (Supplemental Fig. 15).

At REPP, we constructed a ~616 kbp sequence contig consisting of four finished clones (CH276-72N9, CH276-500A7, CH276-327M11, CH276-244L10) completely spanning the REPP cluster at Chromosome 8p23.1 (Fig. 4B). Analysis of sequence overlaps determined that clones CH276-500A7 and CH276-327M11 were derived from the same haplotype (0.9999 sequence overlap), whereas the two flanking clones, CH276-72N9 and CH276-244L10, represented the alternative haplotype (>0.9995 sequence overlap). Sequence comparison using Miropeats revealed that the orangutan REPP cluster was devoid of almost all SDs present on the human CHM1 haplotype (~746.1 kbp) including the proximal beta-defensin cluster, but contained a ~28.8 kbp insertion not present in the CHM1 or human reference genome assemblies (Fig. 4B). To identify whether this insertion was present elsewhere in the orangutan genome, we performed whole-genome shotgun sequence detection (WSSD)[30] using whole-genome shotgun sequencing (WGS) reads generated as part of the orangutan assembly. We found no excess regions read depth, consistent with our BLAST and BLAT searches of the orangutan reference genome (WUGSC 2.0.2/ponAbe2), indicating that this insertion represented unique DNA sequence. Additionally, we queried the gibbon (GGSC Nleu3.0/nomLeu3) and macaque (2006, MGSC Merged

1.0/rheMac2) assemblies and identified a ~25 kbp orthologous region on Chromosomes 4 and 8, respectively. We constructed a ~28 kbp MSA and identified that both the gibbon and macaque sequences lacked a 6 kbp LINE/L1 (L1PA3) repeat sequence within the larger 28 kbp segment however, the flanking sequence showed relatively high levels sequence identity.

Supplemental Section 4.2: DA and Xiao core duplicons—a genomic instability element

To examine the nature of the insertion boundaries before duplicative transposition, we constructed a ~60 kbp MSA containing sequences proximal and distal to the insertion site from our CH276 and CHM1 alternate reference assemblies and used RepeatMasker and DupMasker[31] to characterize the sequence structure at the insertion boundary. At the distal junction of the insertion, we find the presence of an *AluS* repeat mapping precisely at the insertion point (Supplemental Fig. 16, Fig. 4C), consistent with previous reports that younger *Alu* subfamilies are enriched at the SD breakpoints[32]. At the proximal junction a unique pattern of repeat content was discovered in which a series of satellite sequences (SATR1, SATR2) flank a core ERV1 element and a ~1.6 kbp sequence we referred to as the insertion site (Supplemental Fig. 16, Fig. 4C). We assessed all pairwise relationships between the insertion site sequence and the GRCh37 reference assembly (Supplemental Fig. 17) and identified 43 insertion site locations within the GRCh37 reference of which 95% intersected SDs (41/43). We found that this sequence corresponded to a previously described olfactory receptor-like gene family (*OR7E*), associated primarily with euchromatic interchromosomal duplication blocks[33].

Intersection of *OR7E* locations with IAR-containing duplication blocks showed 100% concordance, with the exception of the ancestral 16p13 locus (Supplemental Fig. 18). Using DupMasker, we identified that despite each duplication block showing considerable complexity, a striking higher-order structure existed between the blocks (Supplemental Fig. 18). Sequence analysis determined that the REPP integration point (consisting of SATR1, SATR2, ERV1 and *OR7E*) was a high copy SD previously referred to as the Xiao element (meaning “small” in Chinese)[34,35]. The Xiao element is part of a larger composite, ~200 kbp SD (termed “DA”) that includes the IAR, and comprises two ancestral duplicons located on Chromosomes 16 and 21 (GRCh37 Chr16:5127671-5345785) and (GRCh37 Chr21:33800440-33883647). We determined that DA corresponded to a previously described interchromosomal core duplicon network (M1), one of 24 higher order hierarchical clusters that define the 437 known duplication blocks (Supplemental Fig. 19)[25].

Sequence analysis determined that DA and Xiao are associated with major evolutionary rearrangements occurring at Chromosome 8p23.1. For example, at the orangutan REPD cluster we identified two copies of DA arranged in tandem, localized to the distal boundary of inversion 2 and two copies of Xiao occurring adjacent to lineage-specific duplications encompassing defensin genes (Supplemental Fig. 16). At human REPD, we identified two copies of DA and two copies of Xiao, with the distal copy of DA mapping at the breakpoints of inversion 1 (Supplemental Fig. 16). Comparative sequence analysis determined that a ~18 kbp copy of Xiao localizes at the insertion point of a ~746 kbp duplicative transposition from REPD to REPP (Supplemental Fig. 16). To estimate the evolutionary timing of this duplicative transposition, we constructed a ~87 kbp MSA (GRCh37 Chr8:7258093-7344068) using sequence shared between H1 and H2 REPD and REPP paralogs, including the orthologous sequence in gorilla (CH277-98D3) and orangutan (CH276-501K8) (Supplemental Fig. 20). Using Tajima's relative rate test (MEGA5), we determined that these sequences evolved at the same rate as orthologous counterparts in gorilla and orangutan (MSA PPY-REPP-REPD: $p = 0.78353$, GGO-REPP-REPP: $p = 0.31157$). Using the same strategy as described previously, we estimated that the duplicative transposition occurred 0.84 mya ± 0.99 mya.

Supplemental Section 5: Human sequence diversity analysis

Supplemental Section 5.1: Copy number analysis of β-defensin genes

We specifically sought to address whether individuals carrying the inverted or direct haplotype would show variation in the largest and most highly identical potential deletion mediator, SD19. We note that the particular 96 kbp portion of SD19 that we targeted for genotyping also harbors the seven beta-defensin genes and, as such, this analysis also compares copy-number variation for the beta-defensin cluster. To assess the allelic variation in the general human population, we used WSSD and SUNK analysis[1] to estimate aggregate and paralog-specific copy number. We estimated aggregate and paralog-specific copy number for 236 genomes sequenced as part of the Human Genome Diversity Panel (HGDP)[36] and estimated aggregate copy number for 2504 samples from the 1000 Genomes Project (1KG)[37]. Sequence reads for HGDP were mapped to the H2 assembly, whereas reads from the 1KG were mapped to the GRCh37 reference. We used the subset of samples with inversion status known by FISH validation and/or inferred using PFIDO[24]

to examine copy number differences in the beta-defensin gene cluster among inversion statuses and continental groups. In HGDP, we identified 55 individuals homozygous for H1, 15 homozygous for H2 and 34 harboring the heterozygous haplotype of inversion 1. H1 SUNKs were pulled from the GRCh37 SUNK set and H2 SUNKs were determined by identifying unique 30-mers on H2 that were absent from the rest of the genome. We only used the subset of H2 SUNKs that mapped concordantly to H1 for paralog-specific beta-defensin copy number estimation.

We performed a two-way ANOVA to assess the effect of ancestral group and inversion status on aggregate defensin cluster copy number in the HGDP (Fig. 6). The ancestral group had a significant effect on copy number ($F[1, 99] = 6.602$; $p = 0.012$), and Africans had higher mean copy than non-Africans (5.15 and 4.11, respectively). Individually, non-African continental groups all had lower aggregate beta-defensin copy, as well. When we consider extreme beta-defensin copy number in African and non-African individuals in both the HGDP and 1KG cohorts, we observe that significantly more individuals of African ancestry harbor more than 7 beta-defensin cluster copies relative to their non-African counterparts (Fisher's exact test, $p = 6.3e-6$). Alternatively, inversion status did not have a significant effect on copy number in HGDP ($F[2, 99] = 0.159$; $p = 0.854$) and the interaction between ancestral group and inversion status was not significant ($F[1, 99] = 0.439$; $p = 0.509$).

Considering the SUNK-based, paralog-specific, copy number estimates for the 236 HGDP individuals, we performed a two-way MANOVA with the distal and proximal beta-defensin cluster copy numbers as dependent variables and inversion status and ancestral group as independent variables. In concordance with aggregate copy number results, ancestral group had a significant effect on paralog-specific copy number (Wilks lambda = 0.882; $F[2, 98] = 6.5506$; $p = 0.0021$). Inversion status did not have a significant effect on copy number (MANOVA: Wilks lambda = 0.9859; $F[4, 196] = 0.3484$; $p = 0.8449$). Africans had higher proximal beta-defensin copy than non-Africans ($F[1, 99] = 13.191$; $p = 4.47$ e-4), and all other group differences were not significant (Fig. 6).

While we do not observe differences in overall copy number over this region based on inversion status, when copy number is called in a paralog-specific manner, we observe a lower number of copies of the distal paralog and a higher number of copies of the proximal paralog individuals who are homozygous for the inverted orientation of inversion 1. Because the allelic pairs of beta-

defensins shared between H1 and H2 share greater overall identity than paralogous pairs, our observation of haplotype-specific paralogous copy number differences is not likely a result of gene conversion.

Supplemental Section 5.2: Unique patterns of sequence diversity at Chromosome 8p23.1

To understand patterns of nucleotide diversity between the H1 and H2 haplotypes involving the Chromosome 8p23.1 critical region, we generated a ~3.6 Mbp MSA using BLASR[38] and calculated percent identity in 2 kbp sliding windows (100 bp step) using the program align slider (Nuttle et al. Unpublished). Interestingly, our analysis identified several unique patterns of sequence diversity, including six regions of near-perfect identity segregating between the two haplotypes (Supplemental Fig. 21, Supplemental Table 14). We next calculated Pi using DnaSP[39] and identified five regions showing elevated levels nucleotide diversity between the H1 and H2 haplotypes (Supplemental Table 14). We cross checked these regions against our align slider annotation and found that they did indeed show elevated levels of sequence diversity. To sub sample these sites among diverse humans, we used FES to select and sequence clones from seven fosmid libraries using SMRT sequencing from each of the five regions identified by DnaSP (including two additional regions outside the critical region as a control, region A and G). We generated MSAs based on the largest reciprocal overlap between each of the 7 fosmid haplotypes (~25-30 kbp), including the orthologous region in the chimpanzee, gorilla and orangutan reference assemblies and constructed a series of unrooted phylogenetic trees using the neighbor-joining method. In most cases, we found that subsampled regions from the H1 haplotype did not group separately from the H2 haplotype (regions A, B, C, D, F and G; Supplemental Fig. 21). However a single ~25 kbp region (region E) completely contained within *XKR6* (GRCh37 Chr8:10815626-10839946) demonstrated a tree topology that completely separates the H1 and H2 clades (99% bootstrap support) (Supplemental Fig. 22). Sequence analysis of individuals confirmed for inversion 1 status (direct: NA18956, GRCh37, inverted: CHM1, NA19240, NA12878) identified 36 single-nucleotide variants (SNVs) that directly segregated with inversion 1 status. To estimate the age of the phylogeny, we calculated the number of mutations per branch by multiplying the branch length by the number of total sites (22,930). We next calculated genetic distances and standard errors using MEGA5 (as described in evolutionary origin of 8p23.1 inversions). Assuming the chimpanzee and human lineages

diverged 6 mya, we estimate that the H1 and H2 haplotypes diverged 0.63 mya ± 0.170 consistent with our initial timing estimates for inversion 1.

Given that the Chromosome 8p23.1 locus has been proposed as a target of natural selection[40], we sought to investigate the 3.6 Mbp critical region for evidence of a recent selective sweep based on extended haplotype homozygosity (eHH)[41]. We used SNV/indel calls from the HGDP and 1KG (ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/ALL.Chr8.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz). HGPD variants were phased using BEAGLE 4.0 (r1399). We calculated observed heterozygosity (oHET), EHH, integrated haplotype score (iHS) and F_{ST} using the genotype-phenotype association toolkit (GPAT), within VCFLIB (https://github.com/ekg/vcflib). We began our analysis by calculating eHH in HGDP and 1KG individuals stratified for inversion 1 status based on either FISH/PFIDO results (HGDP DIR = 41, INV = 15, 1KG DIR = 123, INV = 118). Our analysis identified a strikingly polar pattern of eHH between the direct H1 and inverted H2 haplotypes (Fig. 5). The large block of eHH, ~75 kbp (eHH^D; GRCh37 Chr8:10878357-10953092; eHH: 0.93) detected on the direct haplotype (between clones D and F, Supplemental Fig. 21 and 23) is centered over an H3K27Ac mark in the gene *XKR6* that has been previously tagged by the iHS in Asian populations[40]. Consistent with the frequency of the H1 haplotype in Asian populations, we see high iHS amongst homozygous H1 1KG individuals. The eHH^D block is ~38 kbp downstream of a region (region E, Supplemental Fig. 21) we previously identified as showing elevated sequence diversity between the H1 and H2 haplotypes. *XKR6* (including two other nearby genes on Chromosome 8p23.1 *C8orf12* and *BLK*) was previously implicated as a susceptibility locus for systemic lupus erythematosus (SLE)[42,43] and this gene also shows apparent expression differences between the H1 and H2 haplotypes[24]. While little is known about the function of *XKR6*, it has previously been implicated in regulating the timing of cell apoptosis[44]. We identified a second larger stretch of eHH, (~115 kbp) (eHH^I; GRCh37 Chr8:9484432-9599827, eHH: 0.75) on the inverted haplotype, centered on gene *TNKS* (*Tankyrase*). The eHH^I signal is visible at a genome-wide scale (Supplemental Fig. 24). The identification of patients containing atypical Chromosome 8p23.1 deletions and duplications implicated *TNKS* as being the gene responsible for behavior anomalies as phenotypic features of Chromosome 8p23.1 rearrangements[45,46]. The TNKS protein has previously been shown to bind to telomeric repeat binding factor 1, a negative regulator of

telomere length maintenance[47]. Recent work has also indicated that TNKS interacts with axin, the concentration limiting component of the multi-subunit destruction complex, a regulator of the Wnt pathway transcription factor β-catenin[48]. As patterns of eHH are also consistent with loss of heterozygosity, we calculated oHET in smoothed 50 kbp windows (1 kbp step) over a distance of ~4 Mbp using phased 1KG individuals stratified by population and inversion 1 status. We identified a clear drop in heterozygosity corresponding to locations of eHH identified on the direct and inverted haplotypes (Supplemental Fig. 25). This effect was not observed in populations independent of inversion 1 status, with the exception of individuals out of Asia (CHB, CHS and JPT) whereby a drop in oHET was observed at the block of eHH identified on the direct haplotype (Supplemental Fig. 25).

To observe the breakdown of linkage disequilibrium on core haplotypes, we created bifurcation diagrams indicative of long range haplotypes. We identified a predominant thick branch extending proximally and distally from the core single-nucleotide polymorphism (SNP) (rs4841222) at GRCh37 Chr8:9656302 on the H2 haplotype. In combination with the correspondingly high eHH value (0.75), we surmised that this was indicative of long range linkage disequilibrium (Fig. 5). We next assessed whether this eHH pattern was population specific by measuring eHH for several populations in the 1KG, independent of inversion status. We found that the eHH^D signal can be identified in the Chinese population (CHB), but is absent from populations of African (YRI, GWD, LWK) or European (CEU) origin (Supplemental Fig. 26). This finding is consistent with our oHET data, but is also representative of the global frequency for the direct haplotype. Of note, the direct haplotype has been shown to have a frequency of 70-80% in Asian populations[6,24]. The eHH^I pattern was found in CEU, GWD, LWK and YRI, populations in which the frequency of the inverted haplotype is high (Yoruba ~60%, European 20-50%) (Supplemental Fig. 26).

Using the ~115 kbp eHH^I haplotype block extending from core SNP rs4841222, we performed principle component analysis (PCA) on 1KG individuals classified for inversion 1 status using PFIDO[24]. Our PCA analysis shows that even when the genetic substructure is limited to an eHH^I haplotype block, inversion status can be robustly called (Supplemental Fig. 27).

Supplemental Section 6: Breakpoint refinement in patients with 8p23.1 microdeletion

Supplemental Section 6.1: Breakpoint assessment of Chromosome 8p23.1 microdeletions

We analyzed 13 DNA samples from individuals with congenital heart defects and developmental delay by array comparative genomic hybridization (CGH) (two previously identified as carrying Chromosome 8p23.1 microdeletions (https://catalog.coriell.org) (Supplemental Table 6). We also included 23 HapMap individuals from European ancestry that likely represented a population of ethnically matched control individuals. We utilized a customized microarray, which included dense probe coverage of the Chromosome 8p23.1 locus (500 bp spacing). We identified 13 cases of Chromosome 8p23.1 microdeletion, including 7 typical cases (SD mediated) and 6 atypical cases (non-SD mediated) (Fig. 7, Supplemental Fig. 28).

In order to determine the likely breakpoints within the SD blocks we further processed the single-channel Cy3 data from each array and normalized the signal to generate copy number estimates for each array probe (Supplemental Fig. 29). A visualization of smoothed copy number signals (11 probe media) were generated for all samples and clearly demonstrate the expected signal depletion of the unique inter-SD region in all cases with expected deletions as well as signal depletion in the proximal beta-defensin region suggesting multiple breakpoints within the SD cluster. In order to specifically ascertain the probable breakpoints and their mediating SD in the H2 assembly, we generated model array profiles representing the signal depletion expected for each unique and duplicated array probe across the H2 contig under recombination between each of the direct SD pairs. To increase the specificity of our analysis, this was restricted to probes with a median array predicted copy number under 10 in the 20 European HapMap samples profiled. We then compared each model to determine the strength of the signal that would be associated with each specific model (Supplemental Tables 7 and 8).

Examination of the ideal signal responses suggests that we have good performance in discriminating SD19- and SD41-mediated events, from SD20/21/25-mediated events. No probes under 10 copies (European median) differentiate between SD20- and SD21-mediated events, thus, we classify these as a single prediction. There is a borderline significant difference (Supplemental Table 7) between SD25 and SD20/21 events; however, the magnitude of the signal (Supplemental Table 8) is only a fraction of that separating other states. Subsequently we

have the lowest confidence in confidently separating these SD blocks as mediators of the deletion. We next compared each samples per-probe copy number difference from the European median copy number (sample – European median) to the predicted differences under each model (Supplemental Table 9).

Examination of the European controls shows the expected best fit to the no deletion model, as do the two non-deleted Signature samples (GC36743 and GC25092). Among the Signature deletions that appear to be SD mediated we observed two distinct classes. Three samples are most similar to the SD19 model (GC11200, GC28058 and GC42961), while the remaining samples most closely fit a predicted proximal breakpoint mediated by SD20/21/25. While the residuals fitting the SD25-mediated event are slightly superior to those for SD20/21, the minimal differences of the observed data and models combined with the larger homology region afforded by the SD20/21 blocks is most likely supportive of SD20/21 events. It is also possible that another structure in the H1 haplotype could generate this signal; however, the lack of structural knowledge of the proximal SD block in H2 prohibits us from directly testing this signal. Even in the presence of these potential biases, this signal strongly suggests a breakpoint more proximal than SD19 in a subset of cases, and that this breakpoint localizes to the DA core in the context of H2.

Supplemental Section 7: References

1. Sudmant, P.H. *et al.* Diversity of Human Copy Number Variation and Multicopy Genes. *Science* **330**, 641-646 (2010).
2. Hach, F. *et al.* mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nat Meth* **7**, 576-577 (2010).
3. Antonacci, F. *et al.* A large and complex structural polymorphism at 16p12.1 underlies microdeletion disease risk. *Nat Genet* **42**, 745-750 (2010).
4. Huddleston, J. *et al.* Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Research* **24**, 688-696 (2014).
5. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nat Genet* **37**, 727-732 (2005).
6. Antonacci, F. *et al.* Characterization of six human disease-associated inversion polymorphisms. *Human Molecular Genetics* **18**, 2555-2566 (2009).
7. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403-410 (1990).
8. Parsons, J.D. Miropeats: graphical DNA sequence comparisons. *Computer applications in the biosciences : CABIOS* **11**, 615-619 (1995).
9. Aldred, P.M.R., Hollox, E.J. & Armour, J.A.L. Copy number polymorphism and expression level variation of the human α-defensin genes DEFA1 and DEFA3. *Human Molecular Genetics* **14**, 2045-2052 (2005).
10. Black, H.A., Khan, F.F., Tyson, J. & Armour, J.A. Inferring mechanisms of copy number change from haplotype structures at the human DEFA1A3 locus. *BMC Genomics* **15**, 1-11 (2014).
11. Katoh, K., Misawa, K., Kuma, K.i. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* **30**, 3059-3066 (2002).
12. Chaisson, M.J.P. *et al.* Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608-611 (2015).
13. Bailey, J.A., Yavor, A.M., Massa, H.F., Trask, B.J. & Eichler, E.E. Segmental Duplications: Organization and Impact Within the Current Human Genome Project Assembly. *Genome Research* **11**, 1005-1017 (2001).
14. Ottolini, B. *et al.* Evidence of Convergent Evolution in Humans and Macaques Supports an Adaptive Role for Copy Number Variation of the β-Defensin-2 Gene. *Genome Biology and Evolution* **6**, 3025-3038 (2014).
15. Cooper, G.M. *et al.* A copy number variation morbidity map of developmental delay. *Nat Genet* **43**, 838-846 (2011).
16. Bakar, S.A., Hollox, E.J. & Armour, J.A.L. Allelic recombination between distinct genomic locations generates copy number diversity in human β-defensins. *Proceedings of the National Academy of Sciences* **106**, 853-858 (2009).
17. Wu, T.D. & Watanabe, C.K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859-1875 (2005).
18. Day, N., Hemmaplardh, A., Thurman, R.E., Stamatoyannopoulos, J.A. & Noble, W.S. Unsupervised segmentation of continuous genomic data. *Bioinformatics* **23**, 1424-1426 (2007).

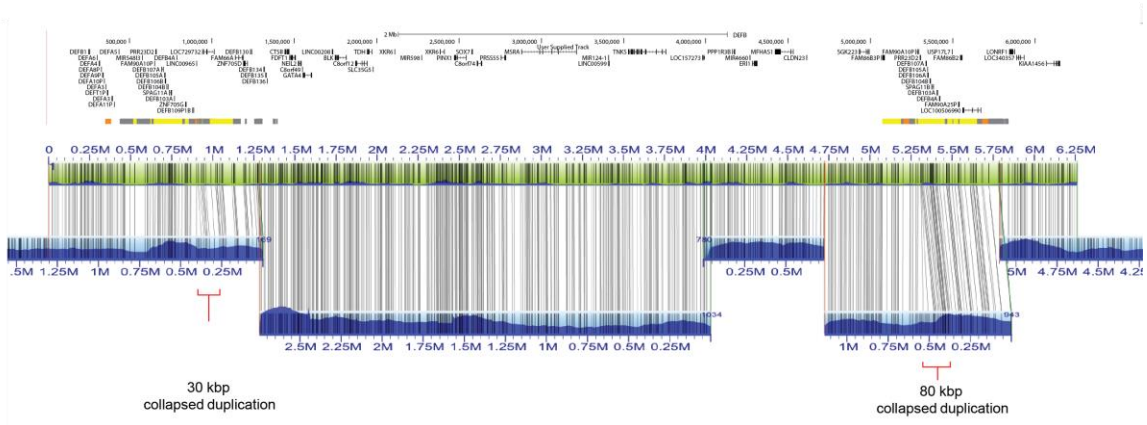
19. Liu, P., Carvalho, C.M.B., Hastings, P.J. & Lupski, J.R. Mechanisms for recurrent and complex human genomic rearrangements. *Current Opinion in Genetics & Development* **22**, 211-220 (2012).
20. Zody, M.C. *et al.* Evolutionary toggling of the MAPT 17q21.31 inversion region. *Nat Genet* **40**, 1076-1083 (2008).
21. Dennis, Megan Y. *et al.* Evolution of Human-Specific Neural SRGAP2 Genes by Incomplete Segmental Duplication. *Cell* **149**, 912-922 (2012).
22. Tamura, K. *et al.* MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Molecular Biology and Evolution* **28**, 2731-2739 (2011).
23. Patterson, N., Richter, D.J., Gnerre, S., Lander, E.S. & Reich, D. Genetic evidence for complex speciation of humans and chimpanzees. *Nature* **441**, 1103-1108 (2006).
24. Salm, M.P.A. *et al.* The origin, global distribution, and functional impact of the human 8p23 inversion polymorphism. *Genome Research* **22**, 1144-1153 (2012).
25. Jiang, Z. *et al.* Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. *Nat Genet* **39**, 1361-1368 (2007).
26. Ventura, M. *et al.* Gorilla genome structural variation reveals evolutionary parallelisms with chimpanzee. *Genome Research* **21**, 1640-1649 (2011).
27. Lichter, P. *et al.* High-resolution mapping of human chromosome 11 by in situ hybridization with cosmid clones. *Science* **247**, 64-69 (1990).
28. Nguyen, T.X., Cole, A.M. & Lehrer, R.I. Evolution of primate θ-defensins: a serpentine path to a sweet tooth. *Peptides* **24**, 1647-1654 (2003).
29. Prado-Martinez, J. *et al.* Great ape genetic diversity and population history. *Nature* **499**, 471-475 (2013).
30. Bailey, J.A. *et al.* Recent Segmental Duplications in the Human Genome. *Science* **297**, 1003-1007 (2002).
31. Jiang, Z., Hubley, R., Smit, A. & Eichler, E.E. DupMasker: A tool for annotating primate segmental duplications. *Genome Research* **18**, 1362-1368 (2008).
32. Johnson, M.E. *et al.* Recurrent duplication-driven transposition of DNA during hominoid evolution. *Proceedings of the National Academy of Sciences* **103**, 17626-17631 (2006).
33. Newman, T. & Trask, B.J. Complex Evolution of 7E Olfactory Receptor Genes in Segmental Duplications. *Genome Research* **13**, 781-793 (2003).
34. Ji, X. & Zhao, S. DA and Xiao-two giant and composite LTR-retrotransposon-like elements identified in the human genome. *Genomics* **91**(2008).
35. Li, X., Slife, J., Patel, N. & Zhao, S. Stepwise evolution of two giant composite LTR-retrotransposon-like elements DA and Xiao. *BMC Evolutionary Biology* **9**, 1-10 (2009).
36. Sudmant, P.H. *et al.* Global diversity, population stratification, and selection of human copy-number variation. *Science* **349**(2015).
37. Sudmant, P.H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75-81 (2015).
38. Chaisson, M.J. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13**, 1-18 (2012).
39. Rozas, J., Sánchez-DelBarrio, J.C., Messeguer, X. & Rozas, R. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19**, 2496-2497 (2003).

40. Pickrell, J.K. *et al.* Signals of recent positive selection in a worldwide sample of human populations. *Genome Research* **19**, 826-837 (2009).
41. Sabeti, P.C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832-837 (2002).
42. Deng, L. *et al.* An unusual haplotype structure on human chromosome 8p23 derived from the inversion polymorphism. *Human Mutation* **29**, 1209-1216 (2008).
43. Harley, I.T.W., Kaufman, K.M., Langefeld, C.D., Harley, J.B. & Kelly, J.A. Genetic susceptibility to SLE: new insights from fine mapping and genome-wide association studies. *Nat Rev Genet* **10**, 285-290 (2009).
44. Giallourakis, C. *et al.* A molecular-properties-based approach to understanding PDZ domain proteins and PDZ ligands. *Genome Research* **16**, 1056-1072 (2006).
45. Barber, J.C.K. *et al.* 8p23.1 duplication syndrome; common, confirmed, and novel features in six further patients. *American Journal of Medical Genetics Part A* **161**, 487-500 (2013).
46. Páez, M.T. *et al.* Two patients with atypical interstitial deletions of 8p23.1: Mapping of phenotypical traits. *American Journal of Medical Genetics Part A* **146A**, 1158-1165 (2008).
47. Smith, S., Giriat, I., Schmitt, A. & de Lange, T. Tankyrase, a Poly(ADP-Ribose) Polymerase at Human Telomeres. *Science* **282**, 1484-1487 (1998).
48. Huang, S.-M.A. *et al.* Tankyrase inhibition stabilizes axin and antagonizes Wnt signalling. *Nature* **461**, 614-620 (2009).

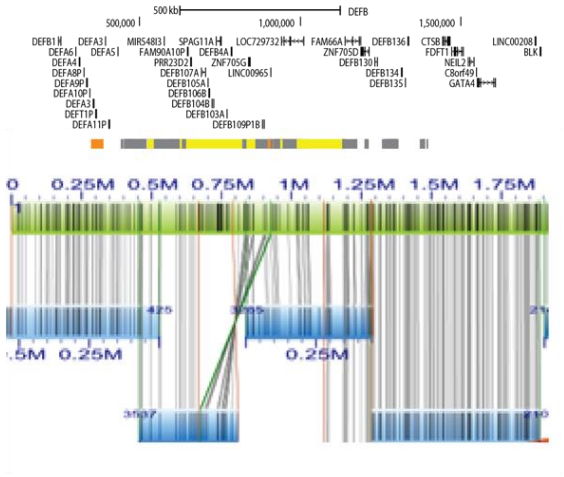
Section 8: Supplemental Figures

Supplemental Fig. 1

a

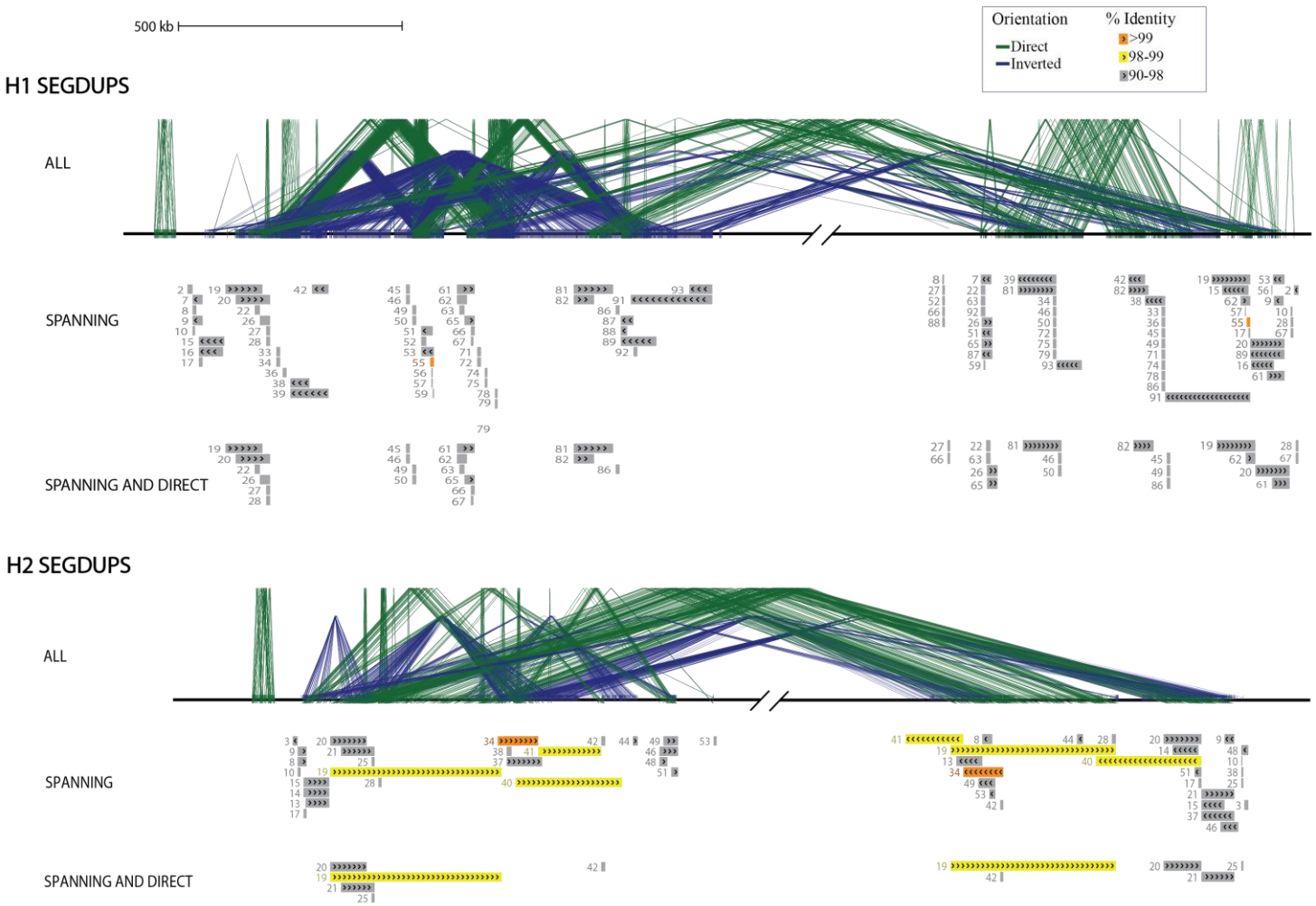


b



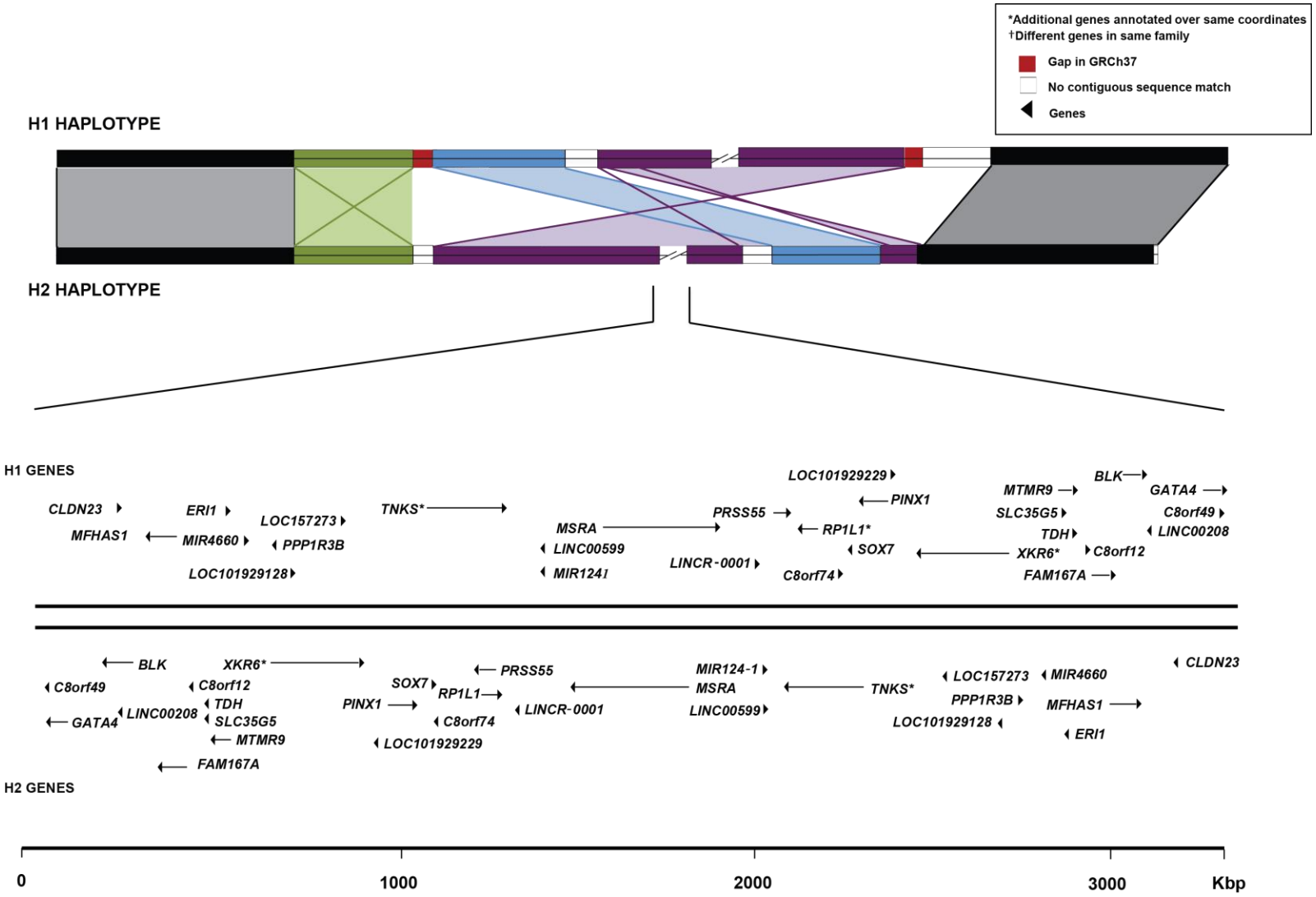
Supplemental Fig. 1: BioNano Genomics fingerprint maps of the CHM1 H2 assembly and the REPD locus of NA19240. a) The fingerprint map of the CHM1 H2 assembly is shown. The consensus genome map (blue with molecule map coverage show in dark blue) is aligned to the *in silico* map (green) created from the CHM1 assembly. The consensus map agrees with the genetic map with the exception of two positions of sequence collapse indicated by the red bars. Overall the order and orientation of the assembly is supported by the genetic mapping data. b) Fingerprint map of the REPD locus within NA19240 is shown. This map clearly displays inversion 2 to be in direct orientation with respect to our CHM1 H2 assembly, similar to the human GRCh37 reference (H1). Based on this map, we estimate the inversion to be ~250 kbp in size and polymorphic.

Supplemental Fig. 2



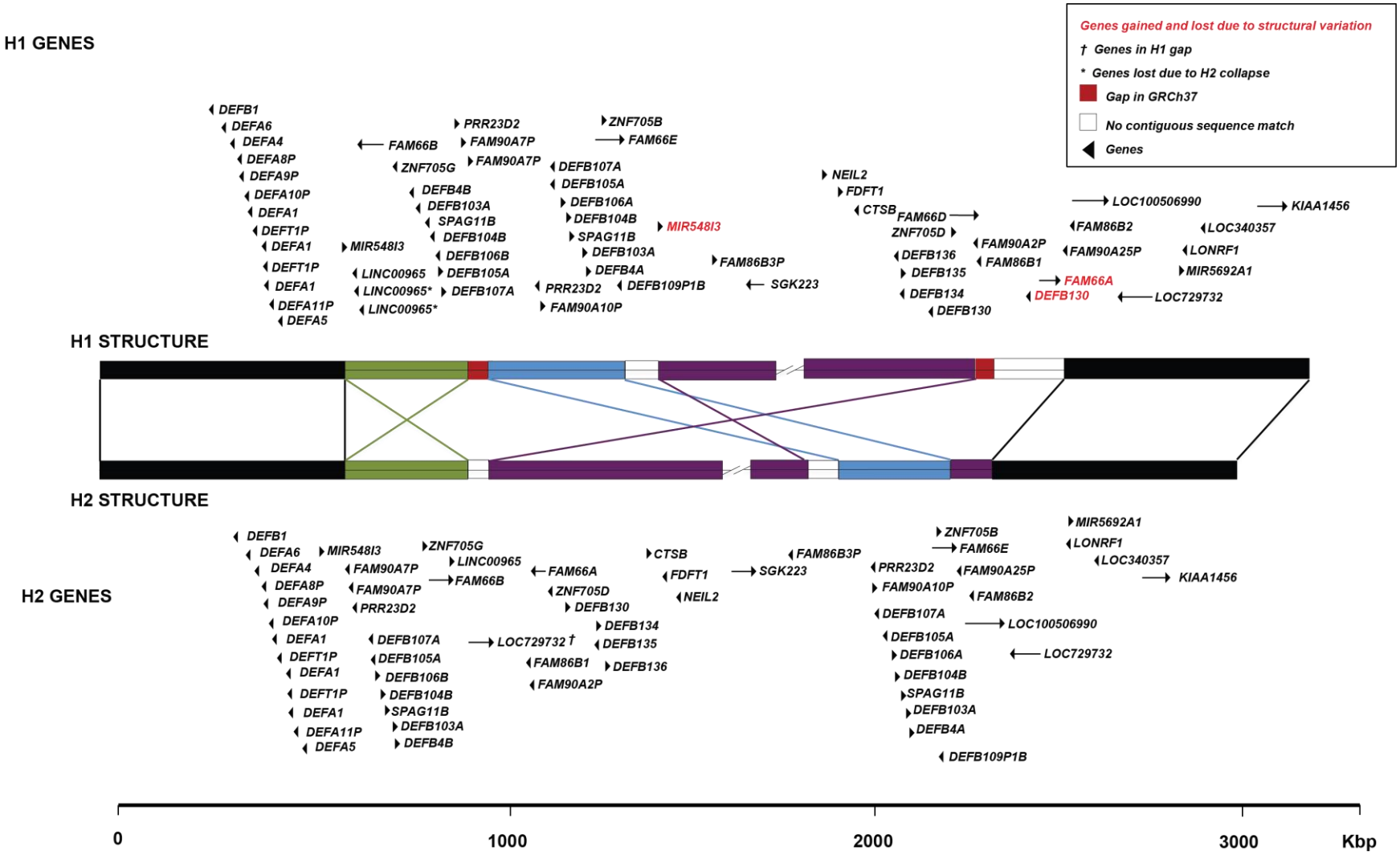
Supplemental Fig. 2: Comparison of segmental duplications at Chromosome 8p23.1 between the GRCh37 reference and H2 assembly. A visual representation of the repeat content using Miropeats and Parasight is performed on the H1 (GRCh37) and H2 (CHM1) assemblies (3.25 Mbp of unique sequence was removed for the purpose of this analysis). Blue lines connect matching segments within a contig that represents inverted orientation, while green lines connect matching segments in direct orientation. SDs are annotated (WGAC). Intrachromosomal duplications that flank the Chromosome 8p23.1 critical region (~3.6 Mbp) are compared between the two haplotypes. Duplication blocks with >99% sequence identity are annotated in orange; blocks demonstrating 98-99% sequence identity are annotated in yellow, and blocks showing 90-98% are annotated gray. The H2 haplotype contains four SDs that >100 kbp directly flanking the Chromosome 8p23.1 critical region. These duplications are missing from the annotated GRCh37 H1 assembly.

Supplemental Fig. 3



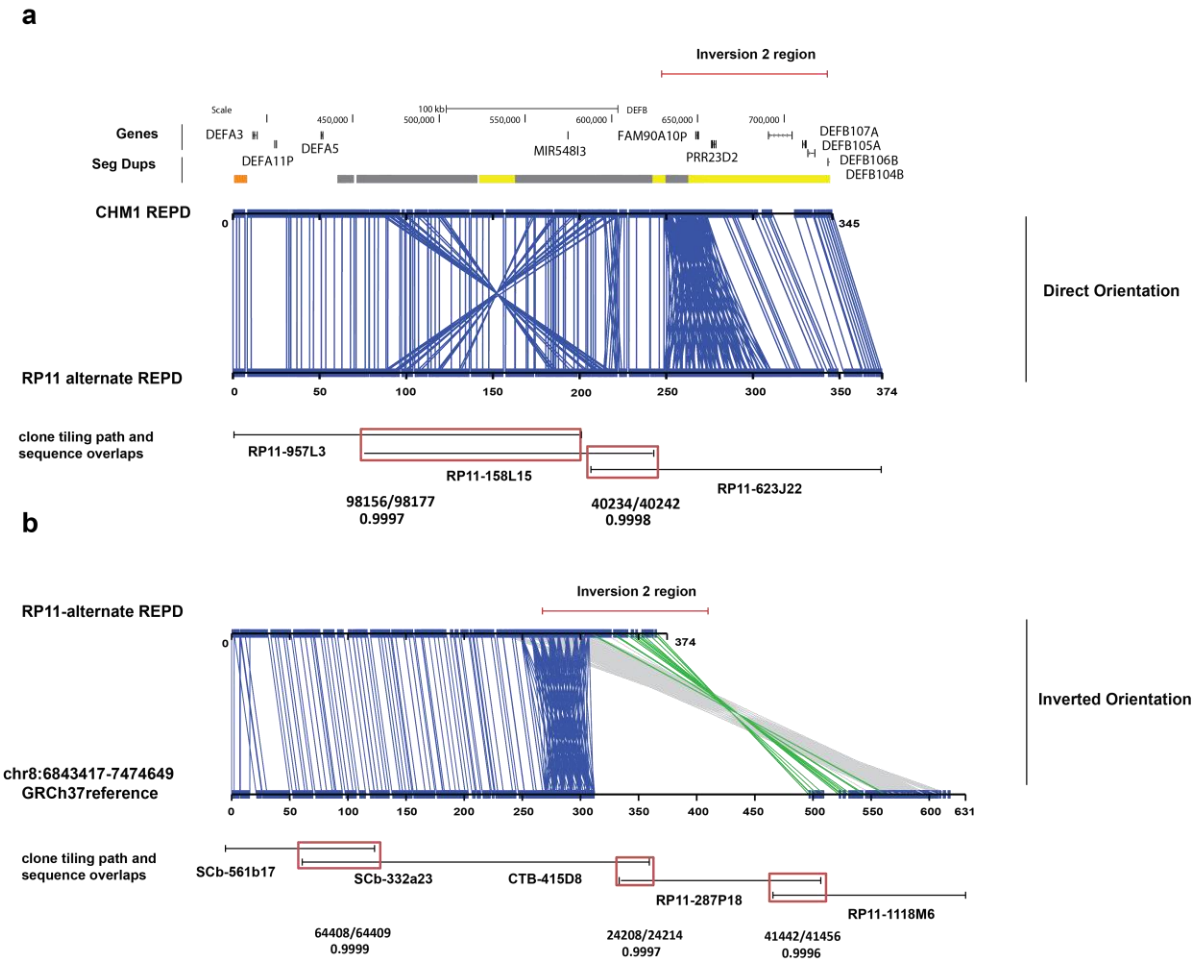
Supplemental Fig. 3: Gene annotations of the Chromosome 8p23.1 critical region compared between the H1 and H2 haplotypes. RefSeq annotations (GRCh37 Chr8:8422607-11667166) are plotted against a diagrammatic structure of the Chromosome 8p23.1 region. The shaded colored boxes depict the three Chromosome 8p23.1 inversion events (green: inversion 2, blue: inversion 3 and purple: inversion 1) and the black shading denotes positions of unique space. The H2 haplotype is annotated using the software GMAP implementing all available RefSeq gene annotations from the GRCh37 reference. Comparison between the H1 and H2 haplotypes shows no differences to genic structure with the exception of orientation.

Supplemental Fig. 4



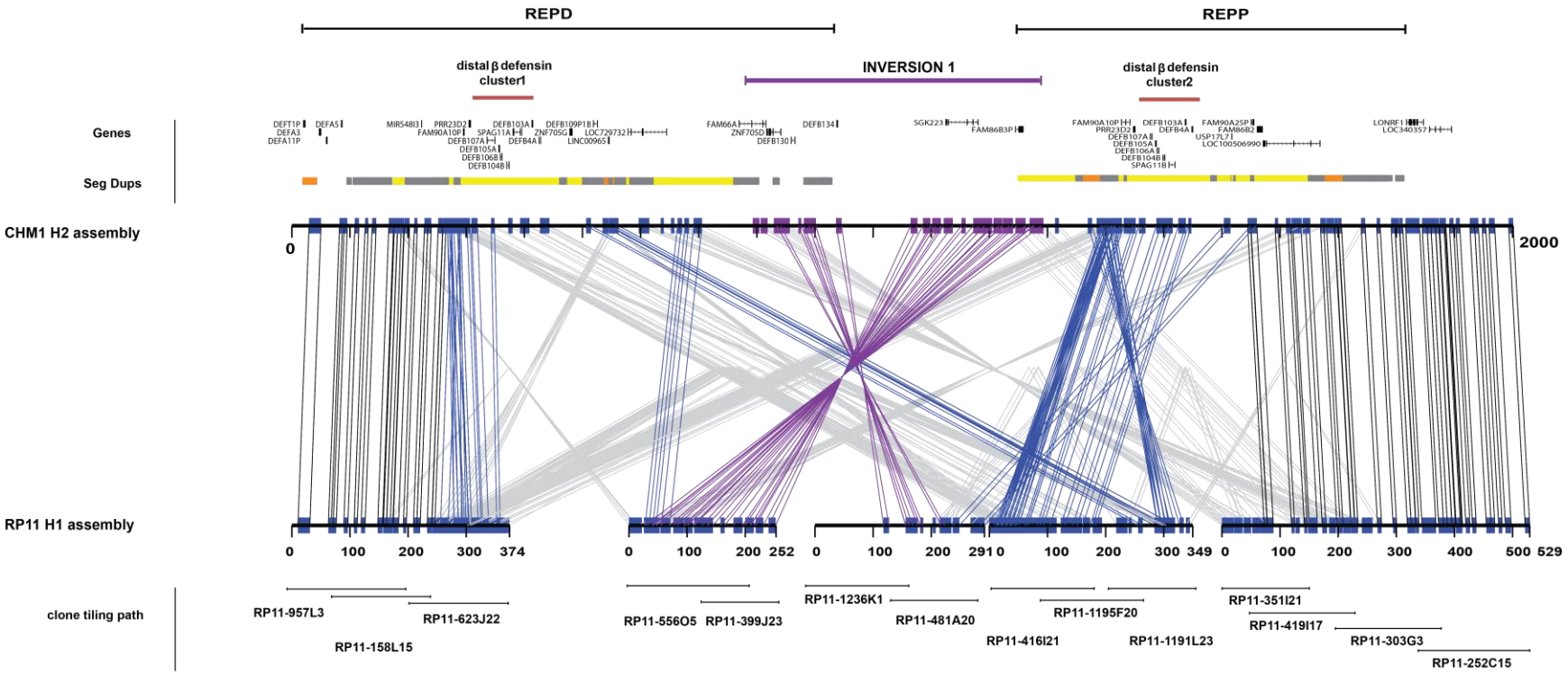
Supplemental Fig. 4: Gene annotations mapping to Chromosome 8p23.1 segmental duplications (REPD and REPP). RefSeq annotations (GRCh37 Chr8:8422607-11667166) are plotted against a diagrammatic structure of the Chromosome 8p23.1 region. The shaded colored boxes depict the three Chromosome 8p23.1 inversion events (green: inversion 2, blue: inversion 3 and purple: inversion 1) and the black shading denotes positions of unique space. The H2 haplotype is annotated using the software GMAP implementing all available RefSeq gene annotations from the GRCh37 reference. Comparison between the H1 and H2 haplotypes revealed six genic differences. Genes annotated in red are missing from the H1 haplotype due to structural variation between the two haplotypes. A ~157 kbp deletion corresponding to position GRCh37: Chr8:7881469-8038443 removes the *MIR548I3* on the H2 haplotype. In addition, a ~136 kbp deletion corresponding to position GRCh37 Chr8:12141855-12277375 removes *DEFB130* and *FAM66A* on the H2 haplotype. Genes annotated with a dagger represent genes that are missing from H1 haplotype (*LOC729732*) due to sequence gaps within the GRCh37 reference assembly. Genes annotated using an asterisk correspond to regions of collapsed duplication within the H2 assembly (*LINC00965*) and are therefore underrepresented on the H2 haplotype.

Supplemental Fig. 5



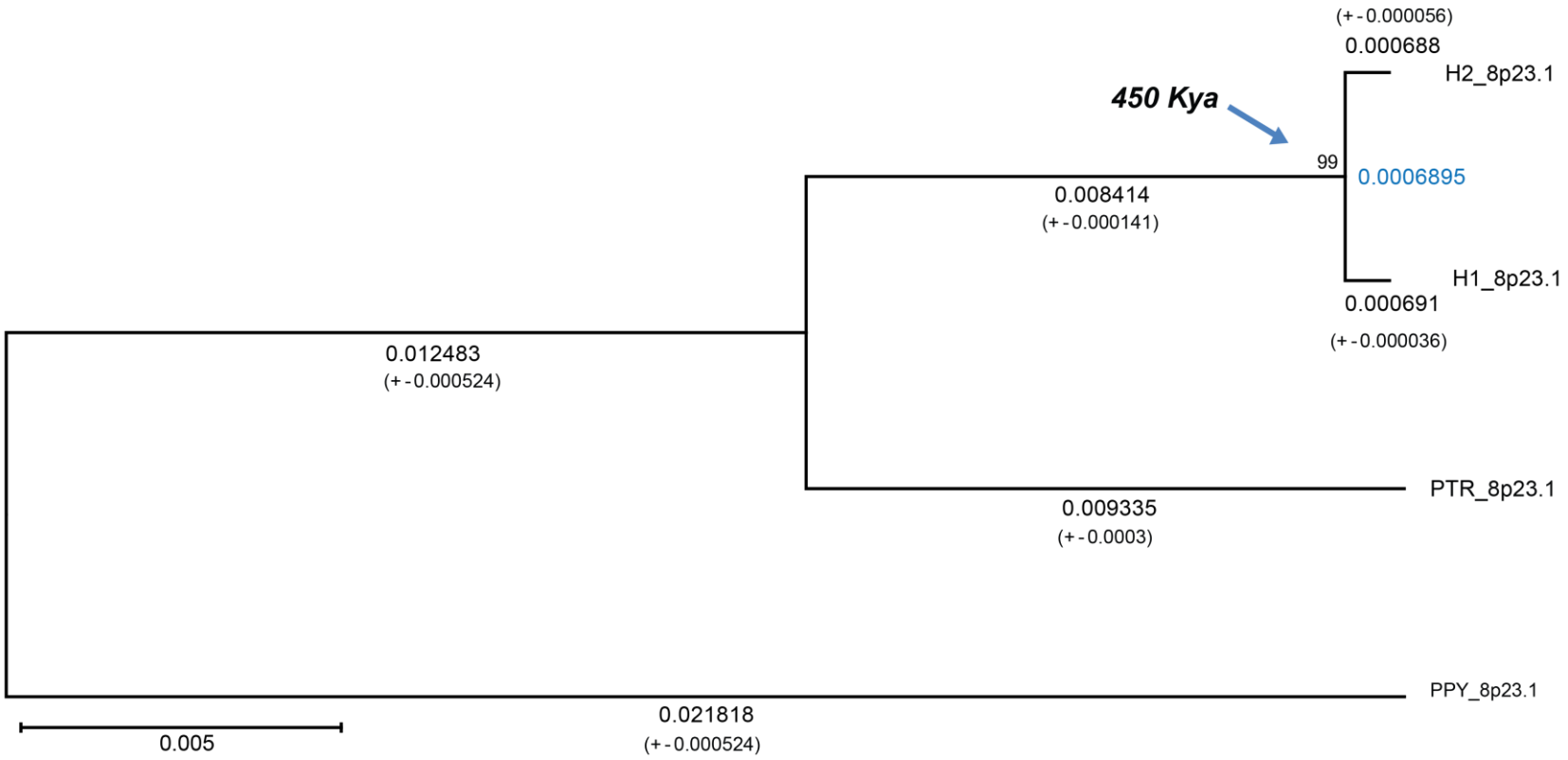
Supplemental Fig. 5: Sequence reconstruction of the RP11 H1 8p23.1 inversion 2 region. a) Sequence comparison between a ~380 kbp RP11 contig spanning the distal breakpoint of inversion 2 and the corresponding region in CHM1 is performed using Miropeats. The two contigs are identical with respect to structure and orientation with the exception of an RP11 expansion in the ~6 kbp tandem repeat array distal to the beta-defensin genes. The analysis demonstrates that RP11 is in the inverted H2 orientation for SD19. b) Sequence comparison between a ~380 kbp RP11 contig spanning the distal breakpoint of inversion 2 and the corresponding region in the GRCh37 reference assembly, comprised of a mosaic haplotype between RP11 and Caltech clone inserts. The tiling path depicts the high-identity clone overlaps in this region despite the use of a mosaic haplotype. Miropeats shows that the GRCh37 reference is in direct orientation relative to the RP11 haplotype.

Supplemental Fig. 6



Supplemental Fig. 6: Sequence reconstruction of the RP11 H1 8p23.1 alternative haplotype. Five sequence contigs are compared with the complete CHM1 H2 haplotype using Miropeats. Joining lines indicate homologous sequence between the two haplotypes with the extent and orientation of segmental duplications shown in the context of REPD and REPP. Black colored lines represents contigs anchored in unique sequence and purple colored lines indicate the presence of the large 4.2 Mbp inversion 1. The locations beta-defensin copies as well as the tiling path of clones sequenced and assembled are depicted. The schematic focuses on SD blocks contained within the flanking sequence of the H2 assembly. Due to the absence of high-quality clone-based sequence in the reference, contigs could not be further extended on the H1 haplotype.

Supplemental Fig. 7



157649 alignment positions
branch length
average branch length for all sequences in a clade

Timing Estimate

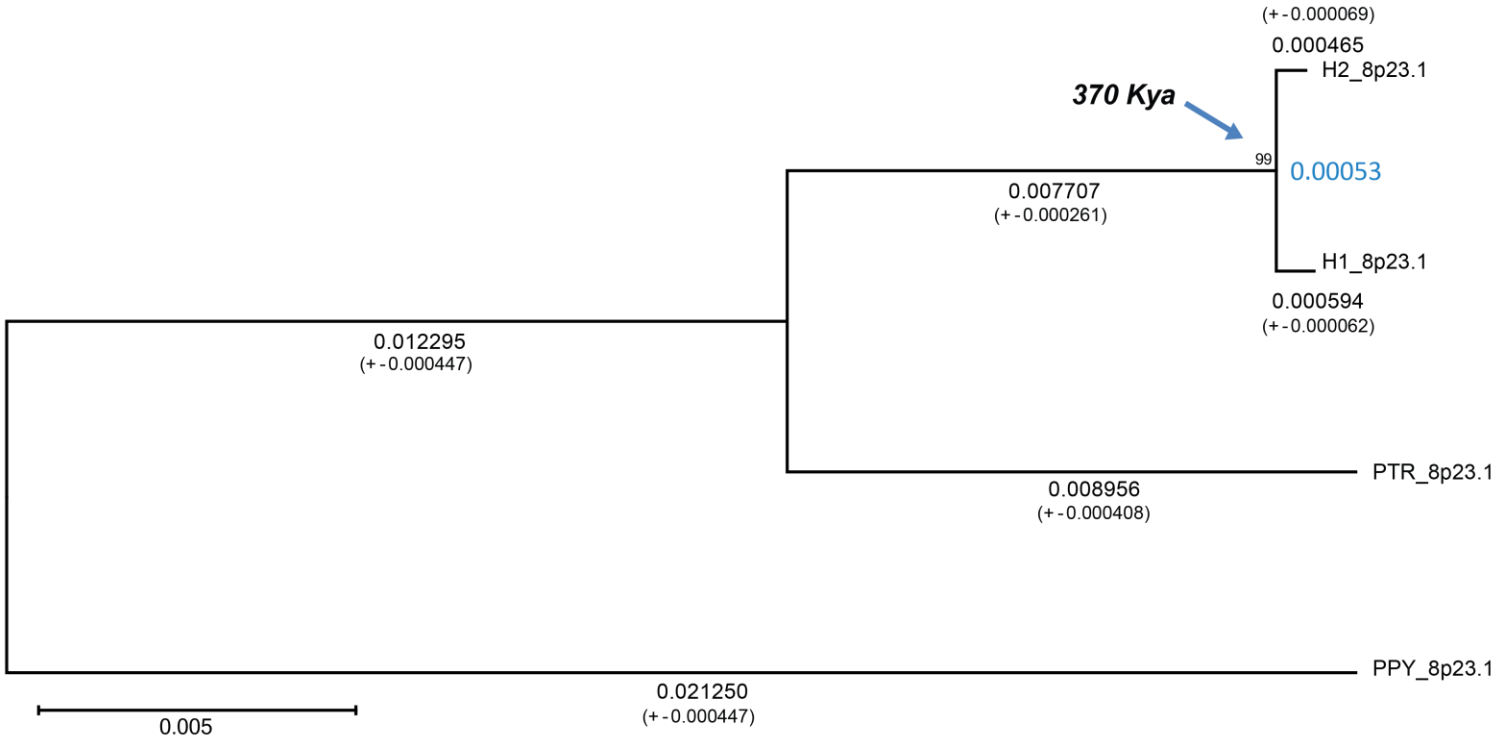
T=((0.0006895 subs/site)/(0.0006895+0.008414+0.009335))*(12 MYA) = 0.45 MYA

T=((0.0006895 subs/site)/(0.0006895+0.008414+0.009335))*(13 MYA) = 0.49 MYA

T=((0.0006895 subs/site)/(0.0006895+0.008414+0.009335))*(14 MYA) = 0.52 MYA

Supplemental Fig. 7: Evolutionary analysis of proximal segment of Chromosome 8p23.1 inversion 1. An unrooted neighbor-joining tree was constructed using the MEGA5 complete deletion option based on ~158 kbp of aligned sequence within the inversion 1 region (Chr8:11547336-11702776). We estimate that the H1 and H2 haplotypes diverged 450 kya.

Supplemental Fig. 8



129062 alignment positions
branch length
average branch length for all sequences in a clade

Timing Estimate

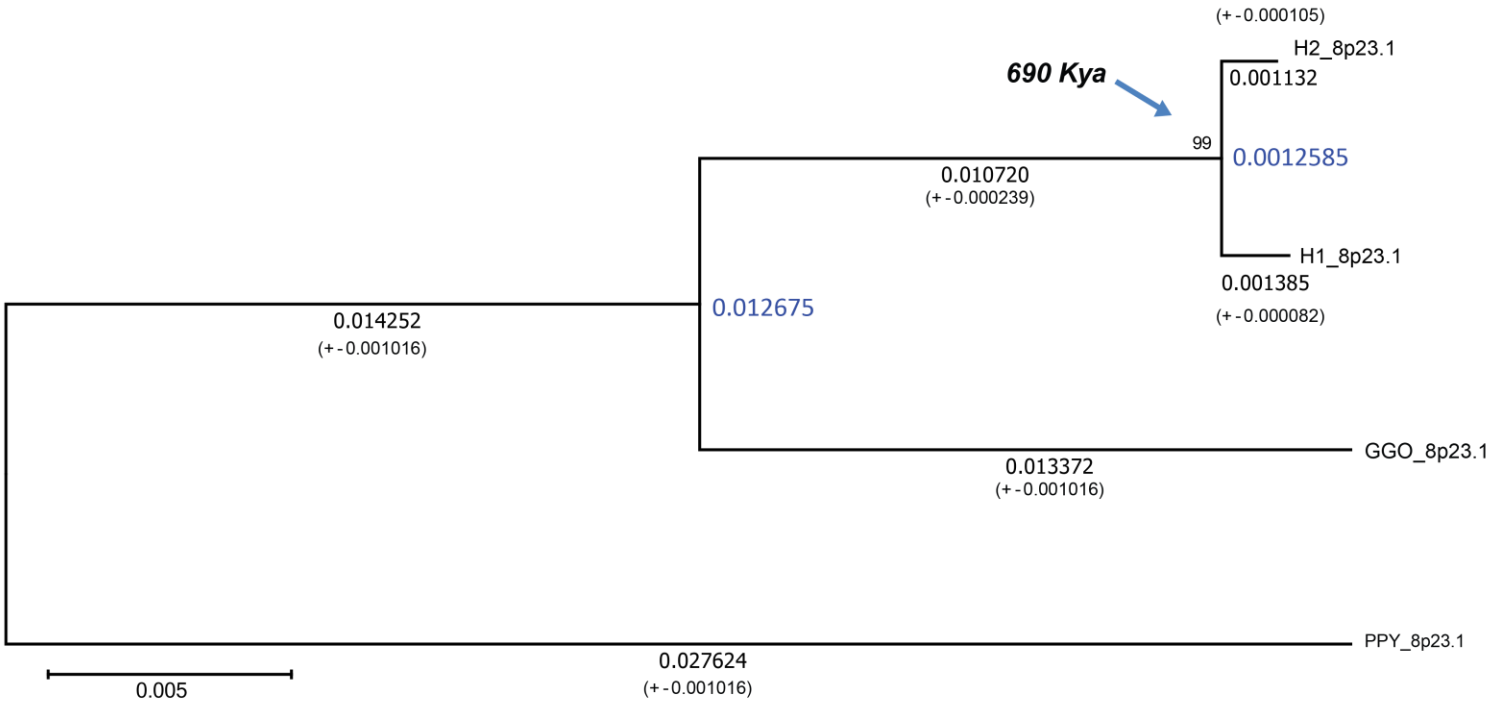
T=((0.00053 subs/site)/(0.00053+0.007707+0.008956))*(12 MYA) = 0.37 MYA

T=((0.00053 subs/site)/(0.00053+0.007707+0.008956))*(13 MYA) = 0.40 MYA

T=((0.00053 subs/site)/(0.00053+0.007707+0.008956))*(14 MYA) = 0.43 MYA

Supplemental Fig. 8: Evolutionary analysis of distal segment of Chromosome 8p23.1 inversion 1.
An unrooted neighbor-joining tree was constructed using the MEGA5 complete deletion option based on ~129 kbp of aligned sequence within the inversion 1 region (Chr8:8191074-8316115). We estimate that the H1 and H2 haplotypes diverged 370 kya.

Supplemental Fig. 9



84703 alignment positions
branch length
average branch length for all sequences in a clade

Timing Estimate

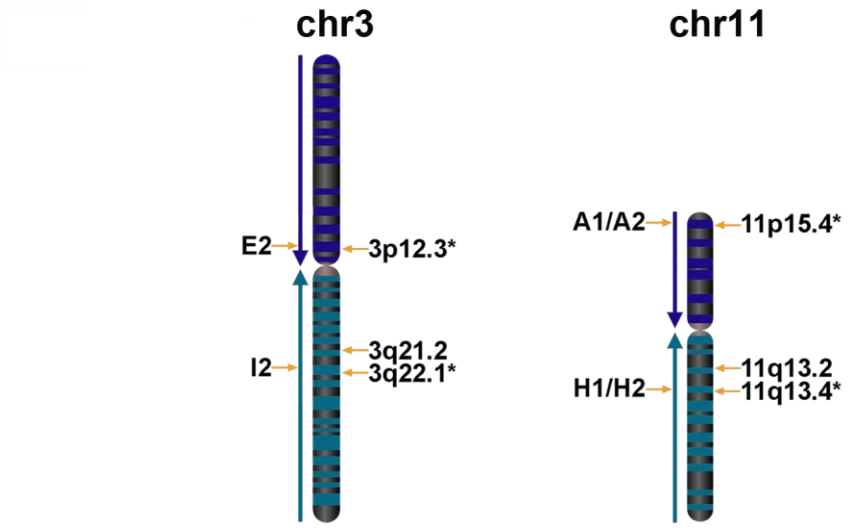
T=((0.0012585 subs/site)/(0.01267525+0.014252+0.027624))*(24 MYA) = 0.55 MYA

T=((0.0012585 subs/site)/(0.01267525+0.014252+0.027624))*(28 MYA) = 0.64 MYA

T=((0.0012585 subs/site)/(0.01267525+0.014252+0.027624))*(30 MYA) = 0.69 MYA

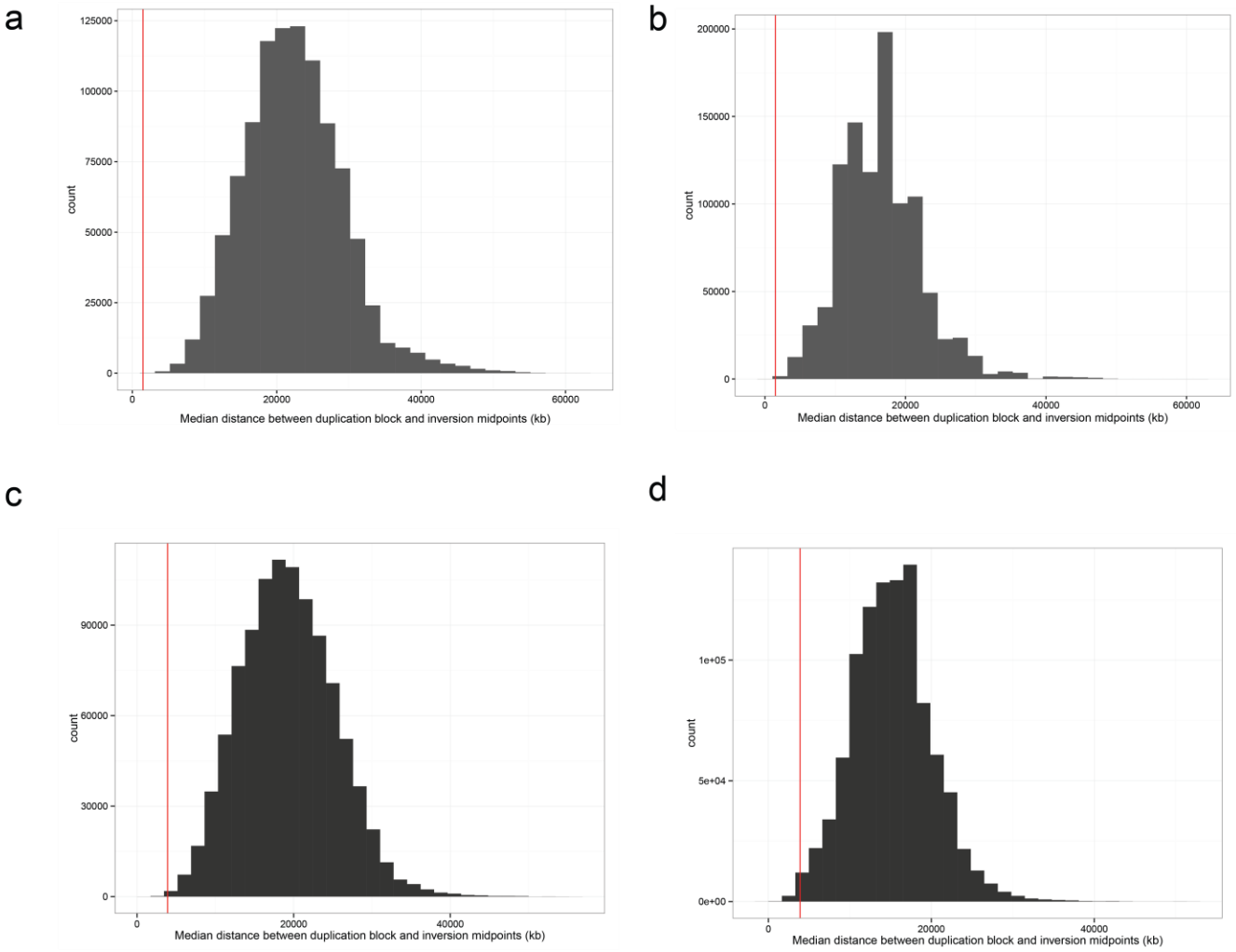
Supplemental Fig. 9: Evolutionary analysis of Chromosome 8p23.1 inversion 2. An unrooted neighbor-joining tree was constructed using the MEGA5 complete deletion option based on ~87 kbp of aligned sequence within the inversion 2 region (Chr8:7258093-7344068). We estimate that the H1 and H2 haplotypes diverged 690 kya.

Supplemental Fig. 10



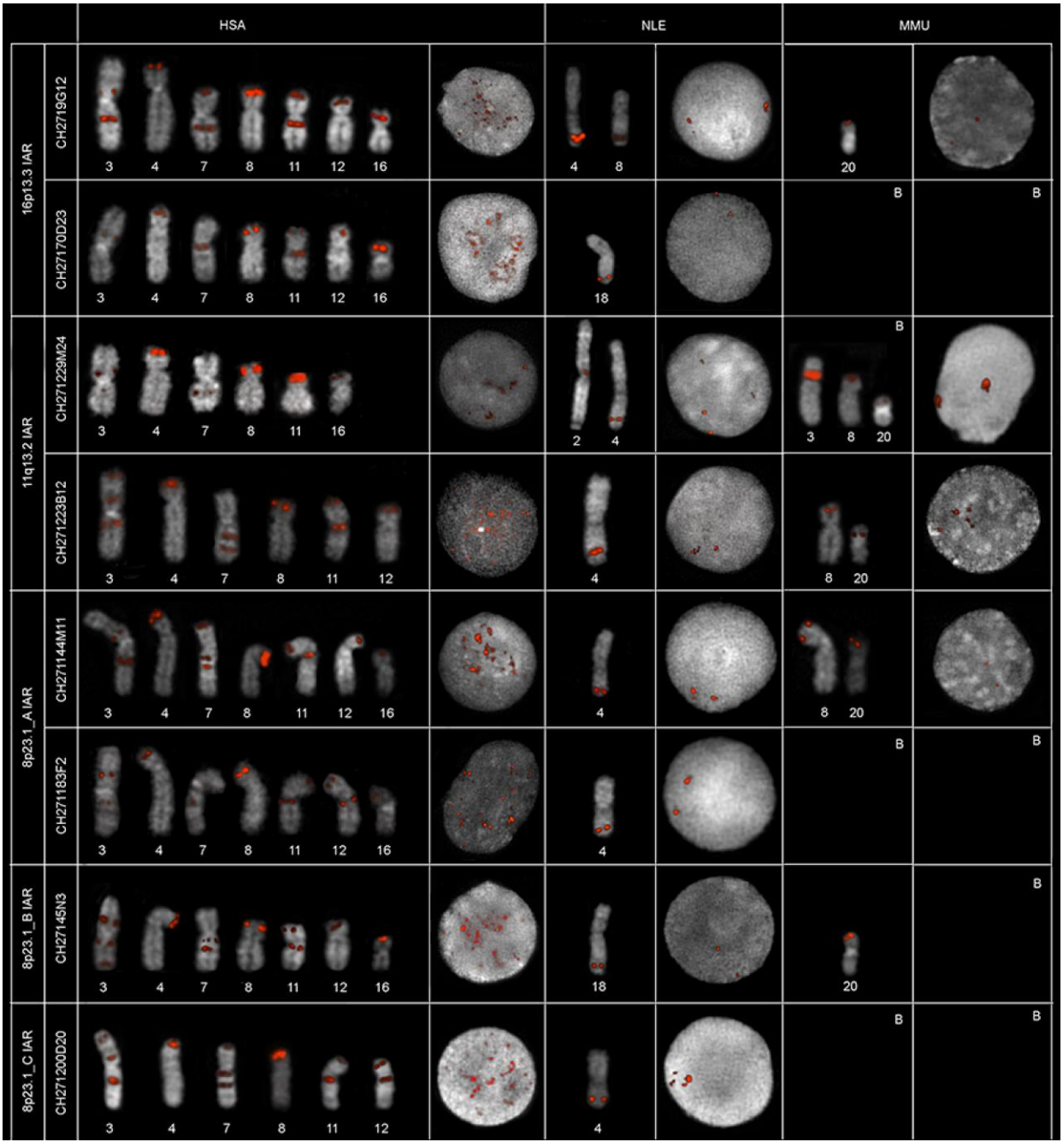
Supplemental Fig. 10: IAR core duplicons mapping at evolutionary inversion breakpoints. IARs are localized at the sites of four evolutionary inversions that have taken place on 11q13.4, 11p15.4, 3p12.3 and 3q22.1.

Supplemental Fig. 11



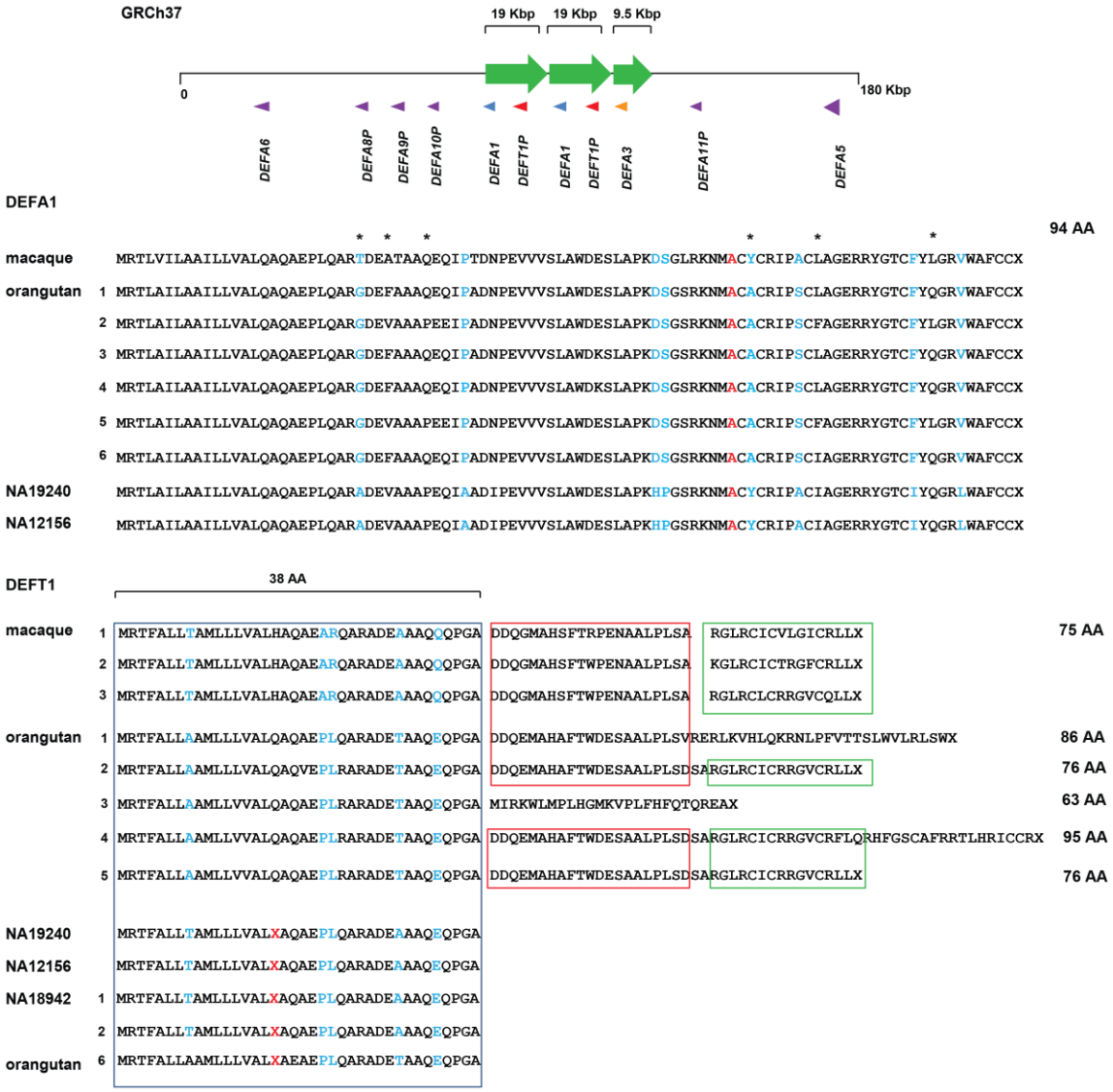
Supplemental Fig. 11: Permutation tests for association between evolutionary inversions and DA containing duplication blocks. Null distribution of median distance between midpoints of evolutionary inversions and randomly placed duplication blocks on a) the same chromosome as the original blocks ($p < 0.000001$) and b) SDs in the same chromosome as the original blocks ($p = 0.000078$). The observed median distance for the original duplication blocks is shown by the red vertical line. Null distribution of median distance between midpoints of evolutionary inversions without Chromosome 8p23.1 and randomly placed duplication blocks on c) the same chromosome as the original blocks ($p = 0.000353$) and d) SDs in the same chromosome as the original blocks ($p = 0.00482$).

Supplemental Fig. 12



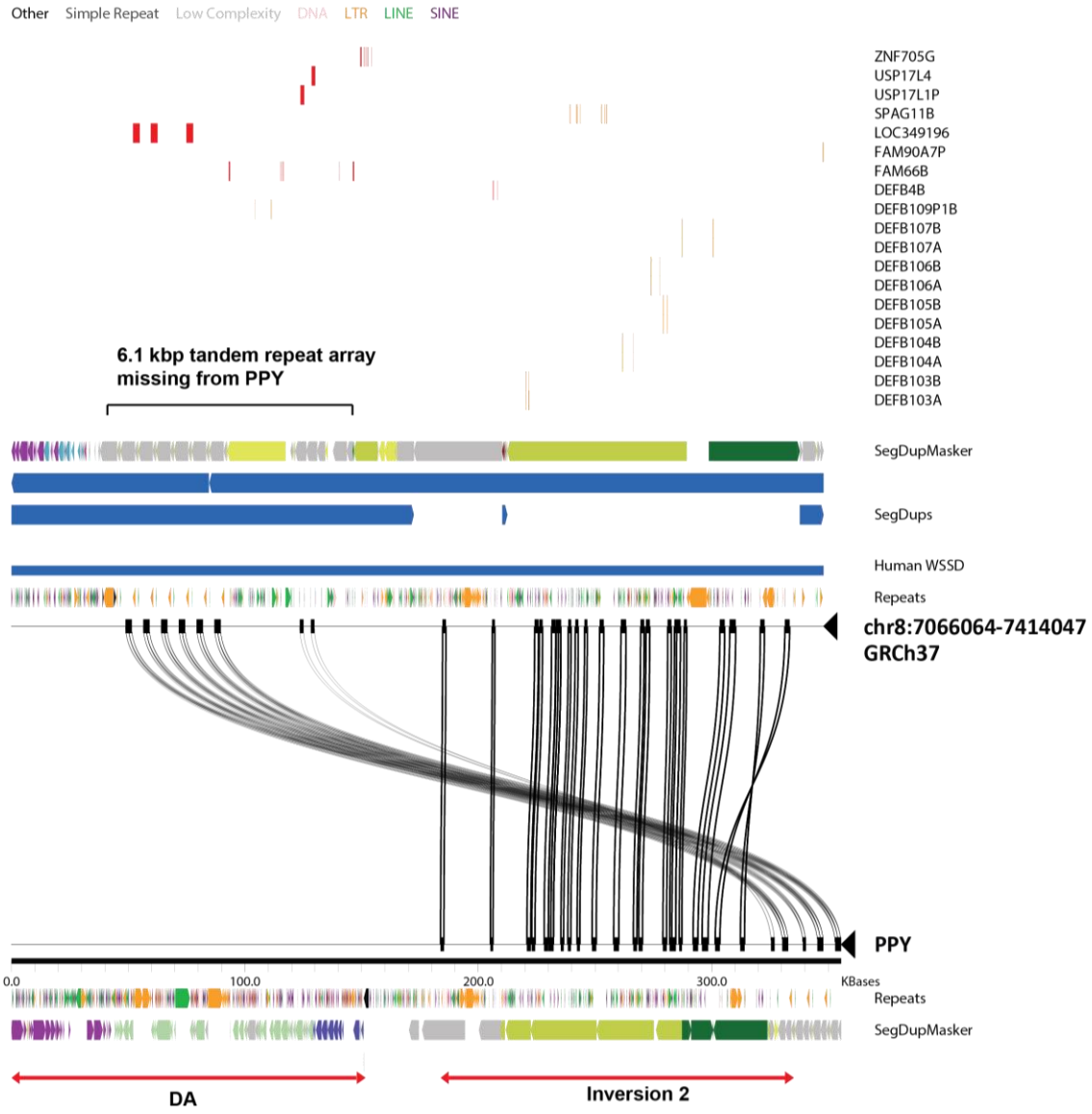
Supplemental Fig. 12: FISH analysis using a probe containing IAR core sequences from the gibbon BAC library. Clones used as probes placed over various genomic IARs based on end mapping (first column). Signals on both extracted chromosomes and interphase nuclei are displayed to show interspersed and tandem duplications. Chromosomes are numbered based on each species karyotype. “B” indicates high background.

Supplemental Fig. 13



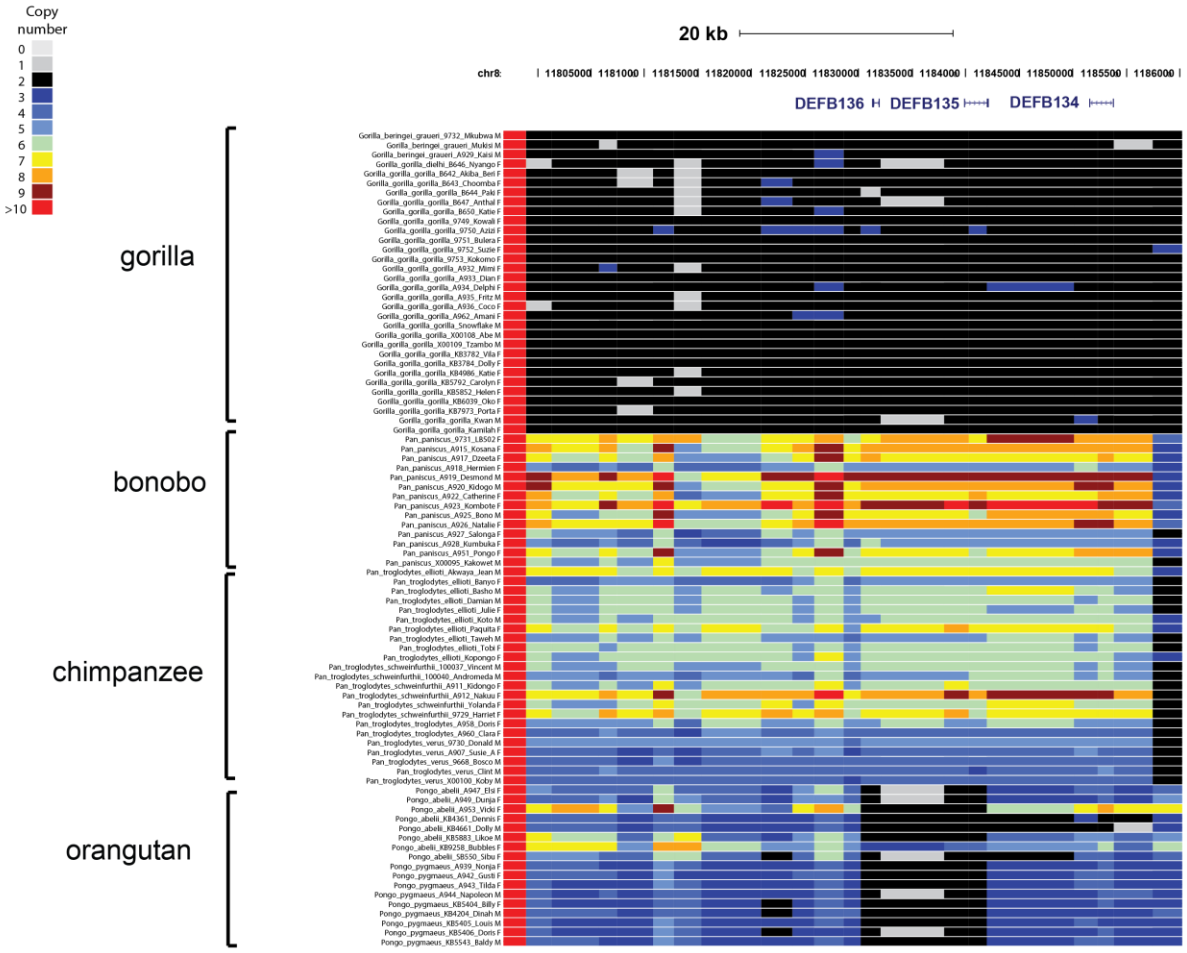
Supplemental Fig. 13: An MSA of translated protein-coding ORFs in the alpha defensin repeat array. The orangutan contains six copies of the *DEFA1* arranged in tandem that encompass a 19 kbp SD. Sequence analysis shows that all copies possess an ORF of 94 amino acids. Comparison of *DEFT1* ORFs shows that the orangutan possesses an ORF in five out of six *DEFT1* copies. In all human haplotypes sampled *DEFT1* is pseudogenized, showing a stop codon in the 17th amino acid.

Supplemental Fig. 14



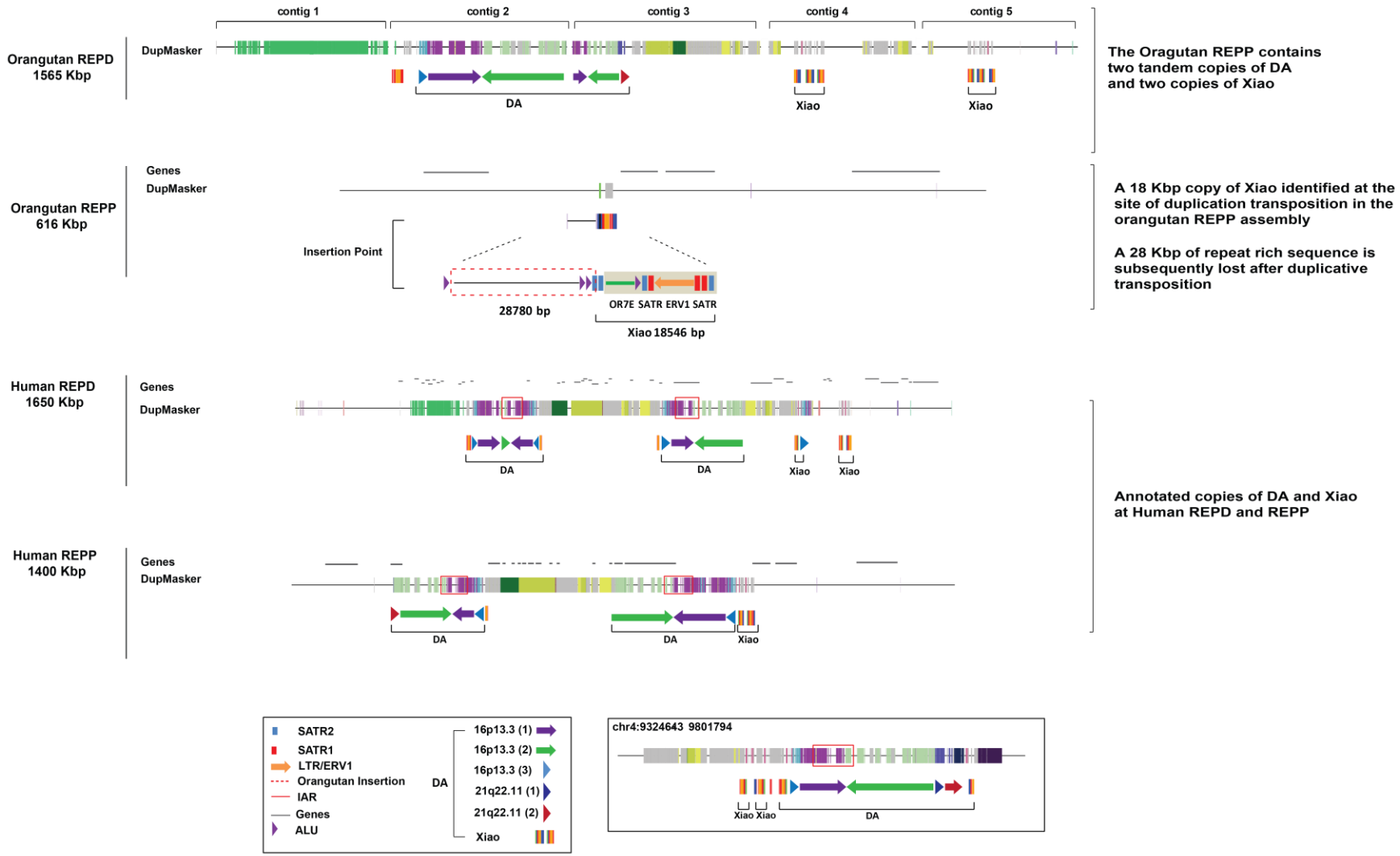
Supplemental Fig. 14: Nonhuman primate clone analysis of Chromosome 8p23.1 inversion 2. Miropeats analysis of a ~350 kbp region encompassing inversion 2 showing that orangutan is in direct orientation at inversion 2 but lacks the 6.1 kbp tandem repeat array on the distal side of the inversion. At the distal boundary of the inversion the orangutan contains a DA interchromosomal core duplicon.

Supplemental Fig. 15



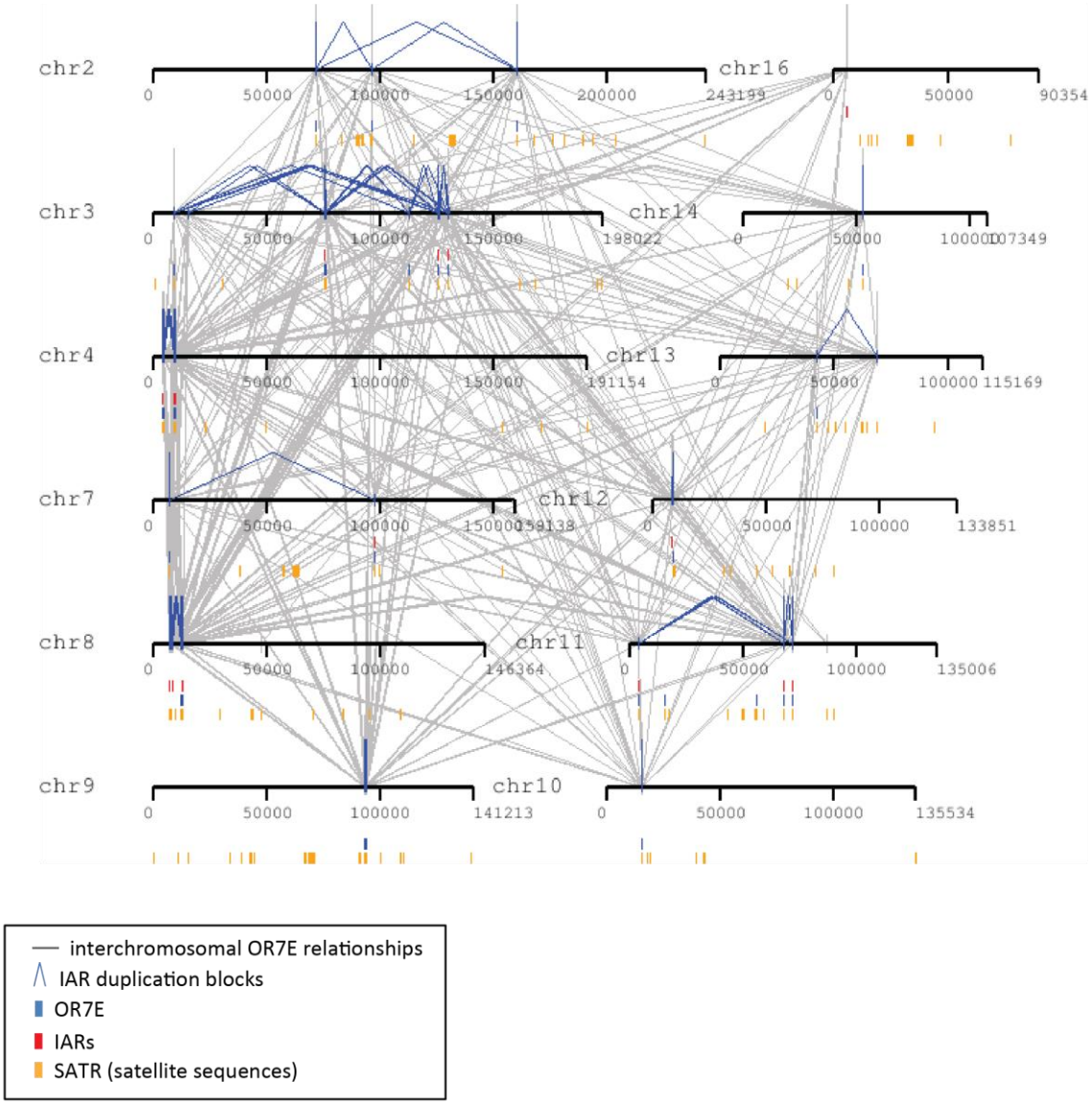
Supplemental Fig. 15: Identification of lineage-specific expansions of *DEFB134*, *DEFB135* and *DEFB136* defensin genes. Duplicated regions were identified after calculating read depth of Illumina reads mapped to GRCh37 human reference genome from nonhuman primates including gorillas (N = 32), chimpanzees (N = 23), bonobos (N = 14), and orangutans (N = 17). These individuals were sequenced at high coverage (~25X) as part of the Great Ape Genome Project. The different boundaries of the duplicated segments in different ape lineages and the high degree of sequence identity of the duplicate copies in the orangutan sequence assembly (~99%) suggest independent expansions among great apes.

Supplemental Fig. 16



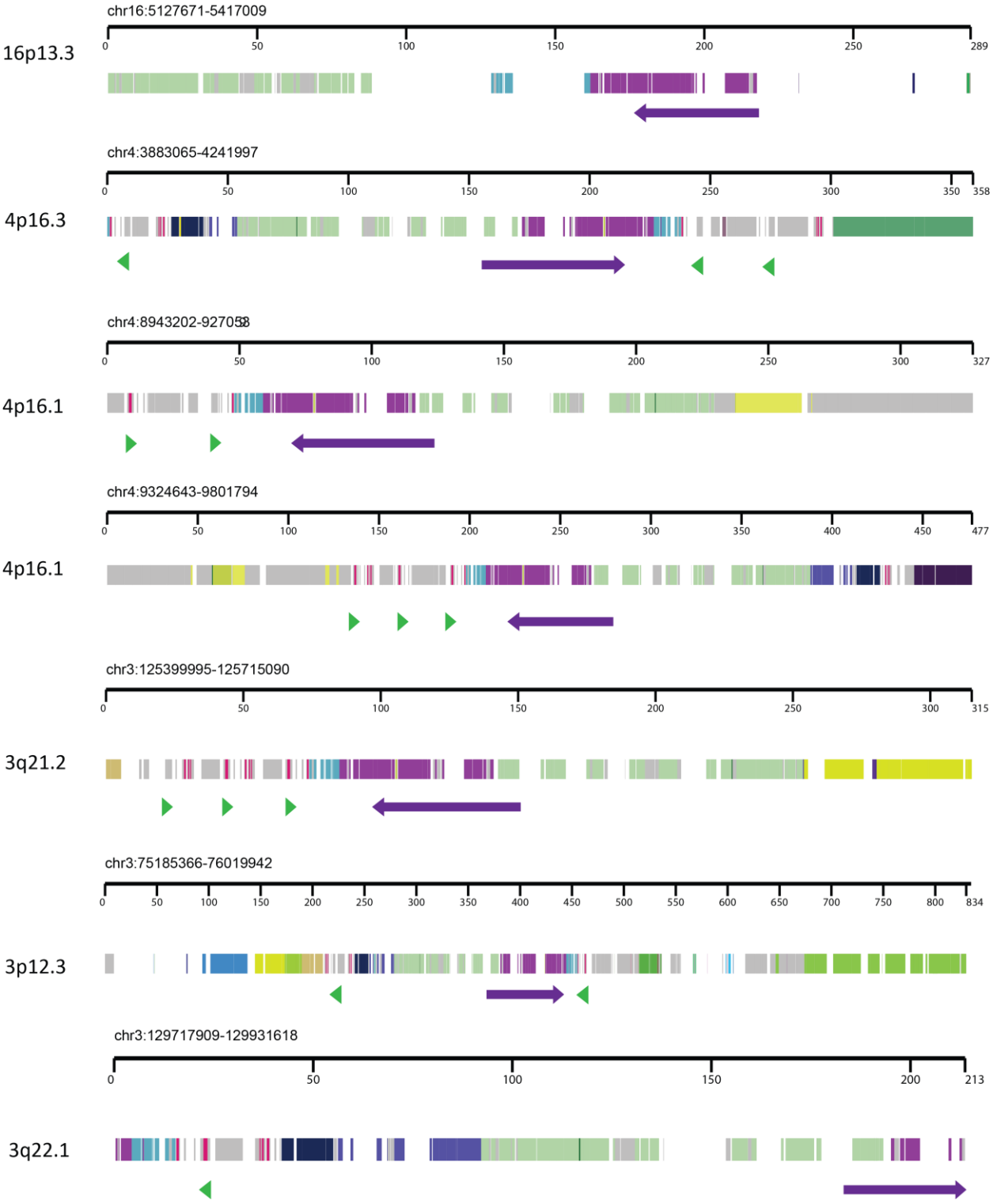
Supplemental Fig. 16: Characterization of core duplicons at Chromosome 8p23.1 in the orangutan and human assemblies. DupMasker annotation of the orangutan REPD locus shows that DA and Xiao localize at positions of genomic instability, including inversions and lineage-specific expansions of defensin genes. In contrast, the orangutan REPP locus is devoid of duplication, with the exception of a Xiao element, the integration point for a ~746 kbp duplicative transposition from REPD. Both human REPD and REPP contain large blocks of SD including four complete copies of DA and three copies Xiao.

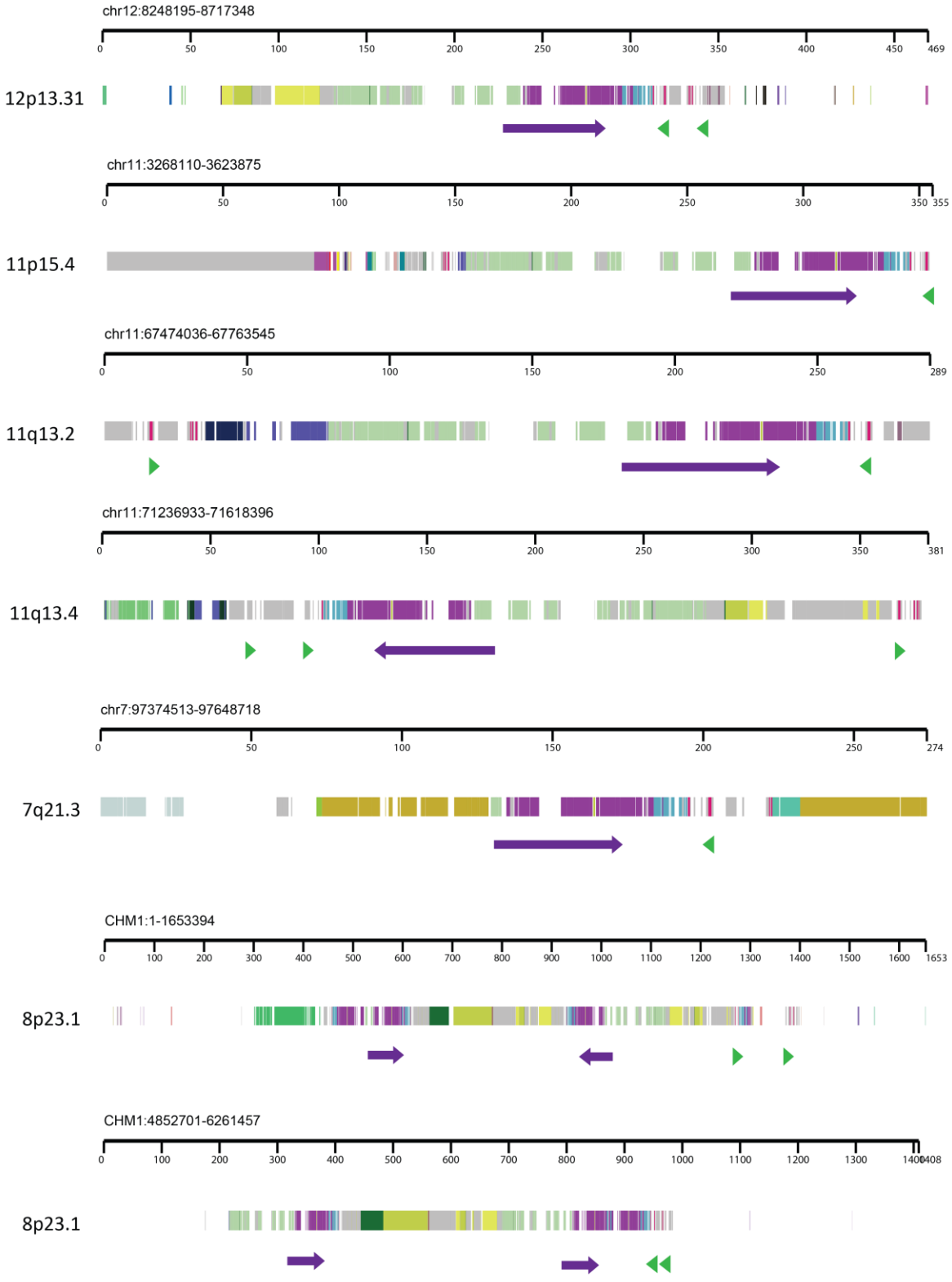
Supplemental Fig. 17



Supplemental Fig. 17: Distribution of Xiao and DA core duplicons in the human genome. The graphic shows the distribution of Xiao (gray lines) across 12 human chromosomes (GRCh37). Xiao maps to 43 locations including 15 that intersect IAR coordinates. IAR duplication blocks (DA) are represented by blue connecting lines and are identified on 7 human chromosomes. Three regions on Chromosomes 3, 4, and 8 contain DA duplication blocks that are interspersed by ~4-5 Mbp and show evidence of inversion.

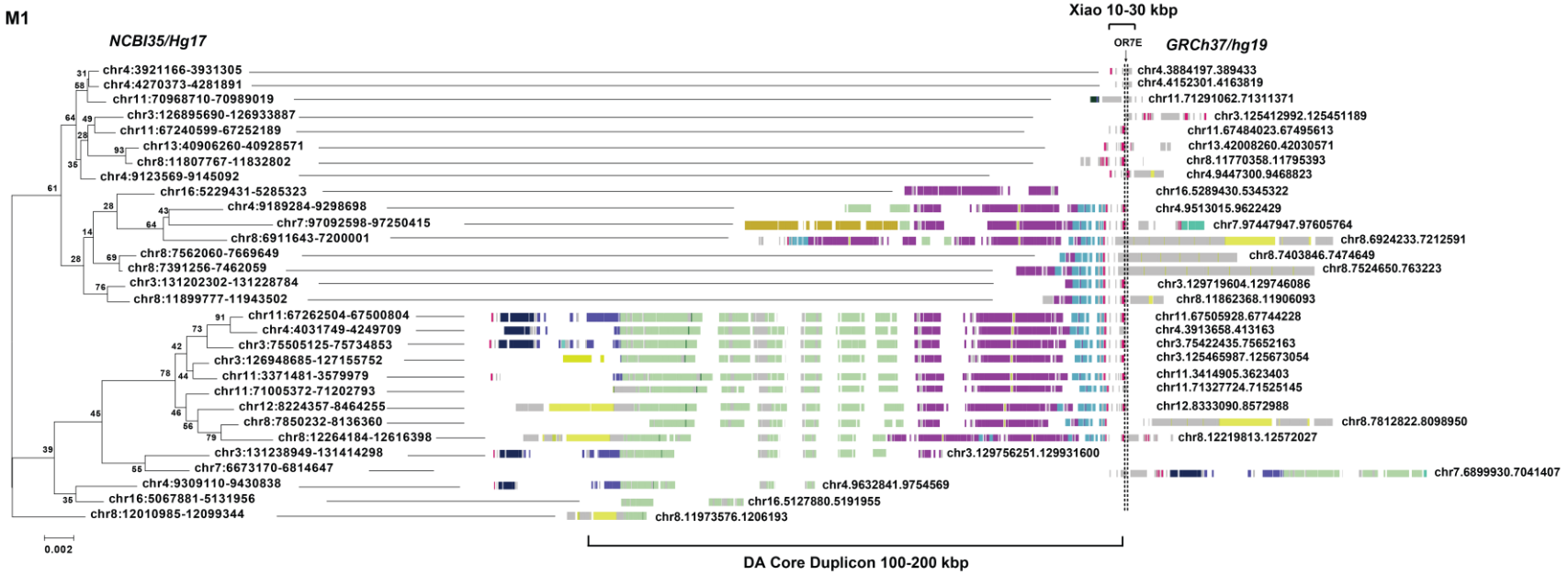
Supplemental Fig. 18




chr12:8248195-8717348
0 50 100 150 200 250 300 350 400 450 469
12p13.31
chr11:3268110-3623875
0 50 100 150 200 250 300 350 355
11p15.4
chr11:67474036-67763545
0 50 100 150 200 250 289
11q13.2
chr11:71236933-71618396
0 50 100 150 200 250 300 350 381
11q13.4
chr7:97374513-97648718
0 50 100 150 200 250 274
7q21.3
CHM1:1-1653394
0 100 200 300 400 500 600 700 800 900 1000 1100 1200 1300 1400 1500 1600 1653
8p23.1
CHM1:4852701-6261457
0 100 200 300 400 500 600 700 800 900 1000 1100 1200 1300 1400 1408
8p23.1

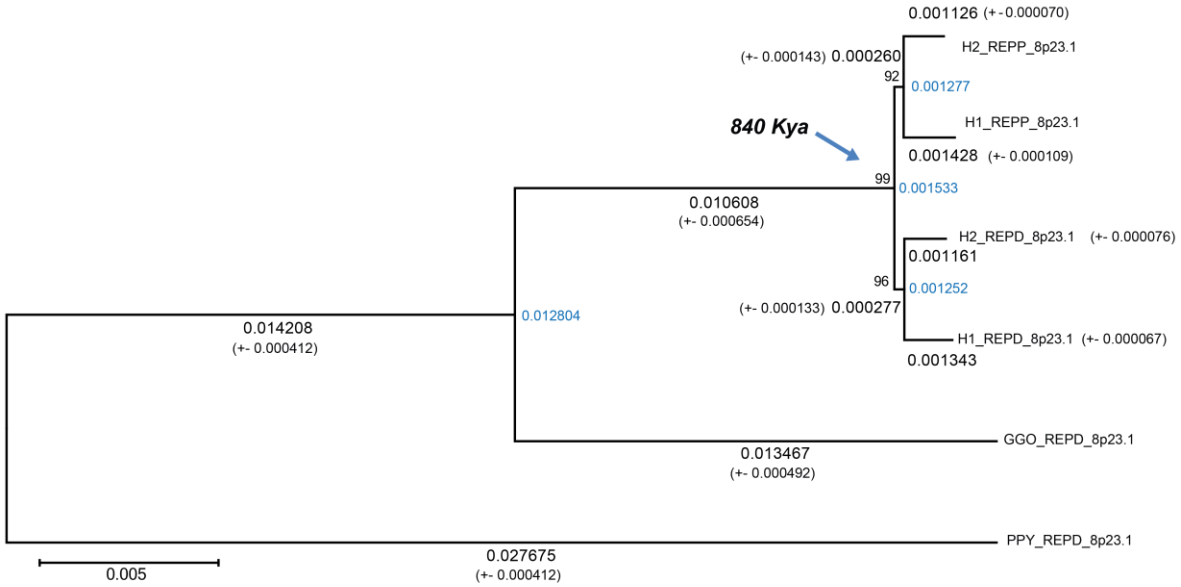
Supplemental Fig. 18: Annotation of interchromosomal duplication blocks containing IARs. The substructure of 15 complex duplication blocks is defined using DupMasker. The organization consists of ancestral duplicons mapping to 16p13 (green and purple) and 21q13 (dark blue). These ancestral duplicons define the composite DA core duplicon that maps at regions of genomic instability including inversions. The positions of IARs (purple arrow) and Xiao (green triangle) are annotated within the larger DA element.

Supplemental Fig. 19



Supplemental Fig. 19: Hierarchical clustering of interchromosomal duplication blocks containing DA and Xiao elements. Complex duplication blocks were clustered together according to their phylogenetic profile. The phylogeny represents the M1 clade, one of the few interchromosomal core duplicons to be associated predominantly with euchromatic regions of the genome.

Supplemental Fig. 20



87913 alignment positions
branch length
average branch length for all sequences in a clade

Timing Estimate

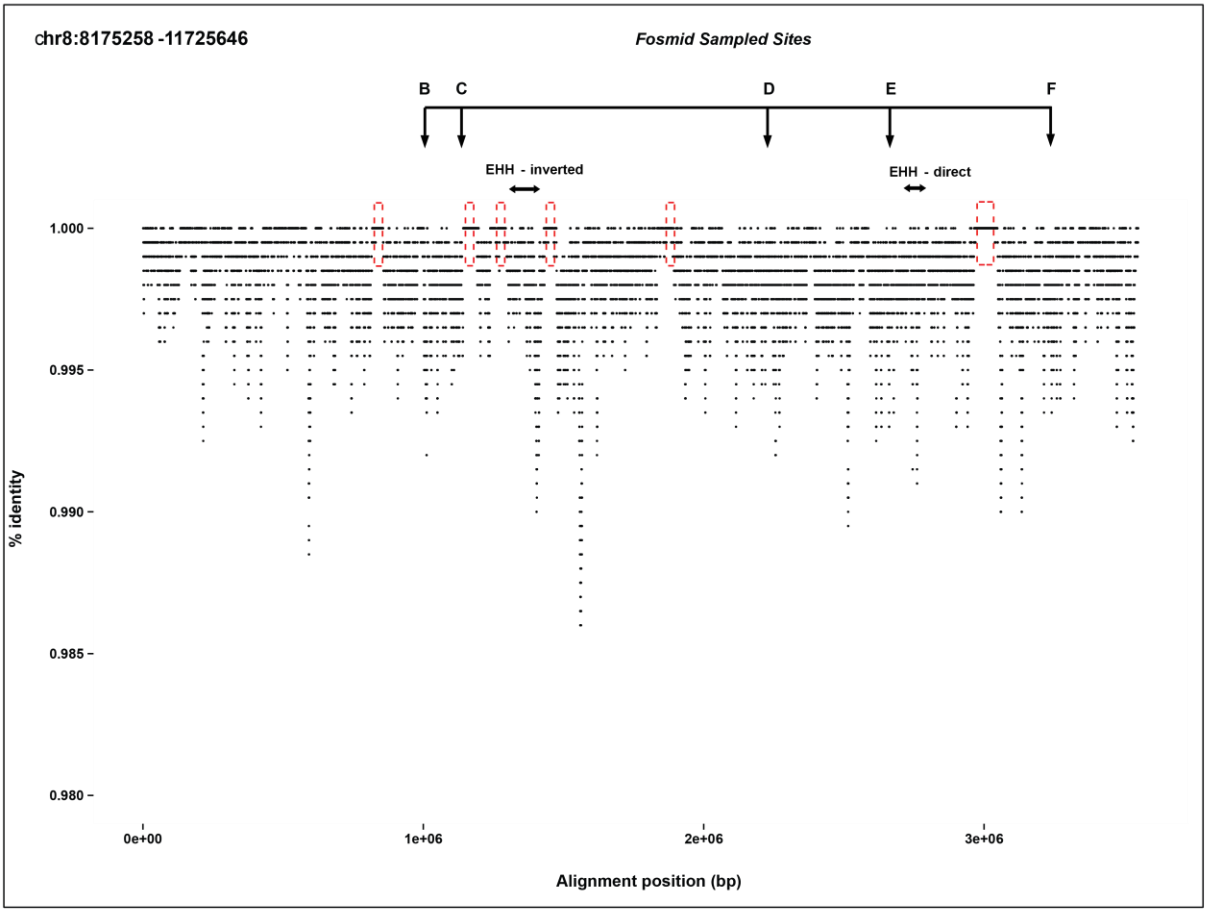
T=((0.001533 subs/site)/(0.012804+0.014208+0.027675))*(30 MYA) = 0.84 MYA

T=((0.001533 subs/site)/(0.012804+0.014208+0.027675))*(28 MYA) = 0.78 MYA

T=((0.001533 subs/site)/(0.012804+0.014208+0.027675))*(24 MYA) = 0.67 MYA

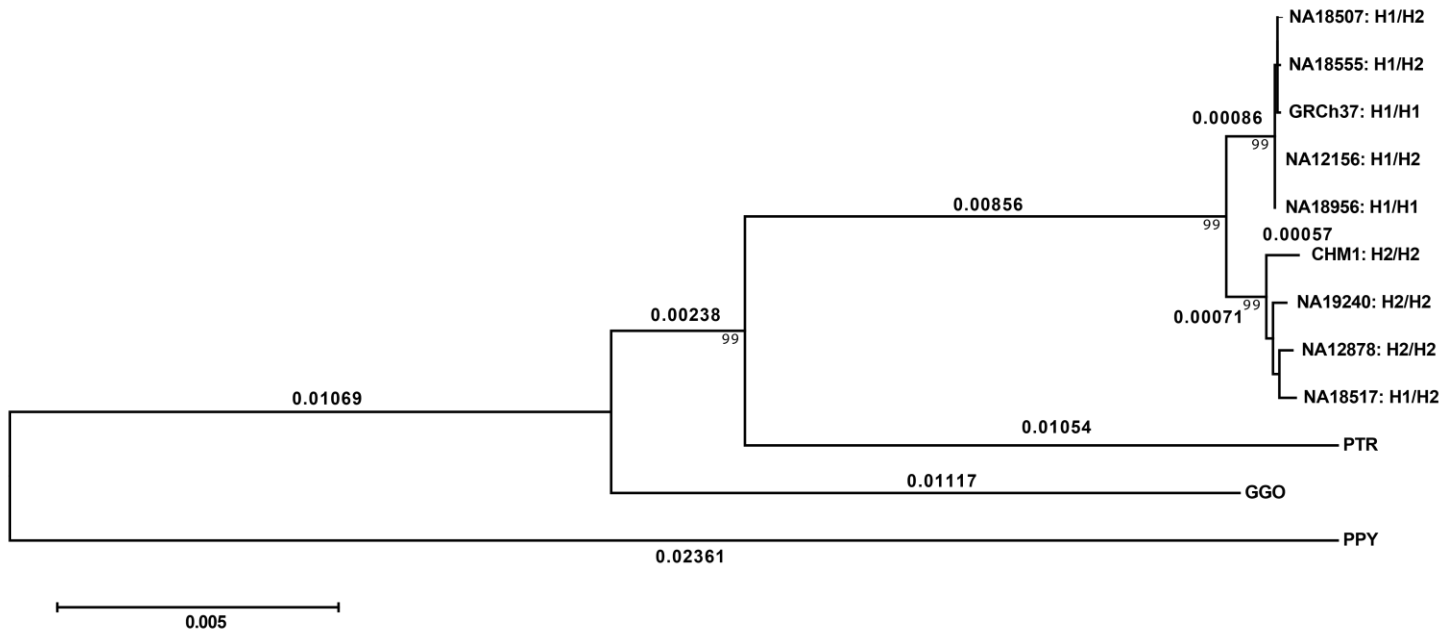
Supplemental Fig. 20: Evolutionary analysis of the ~746 kbp duplicative transposition from REPD to REPP. An unrooted neighbor-joining tree was constructed using the MEGA5 complete deletion option based on ~87 kbp of aligned sequence from sequences representing REPD and REPP, including the orthologous region in gorilla and orangutan. We estimate that ~746 kbp duplicated from REPD to REPP ~840 kya.

Supplemental Fig. 21



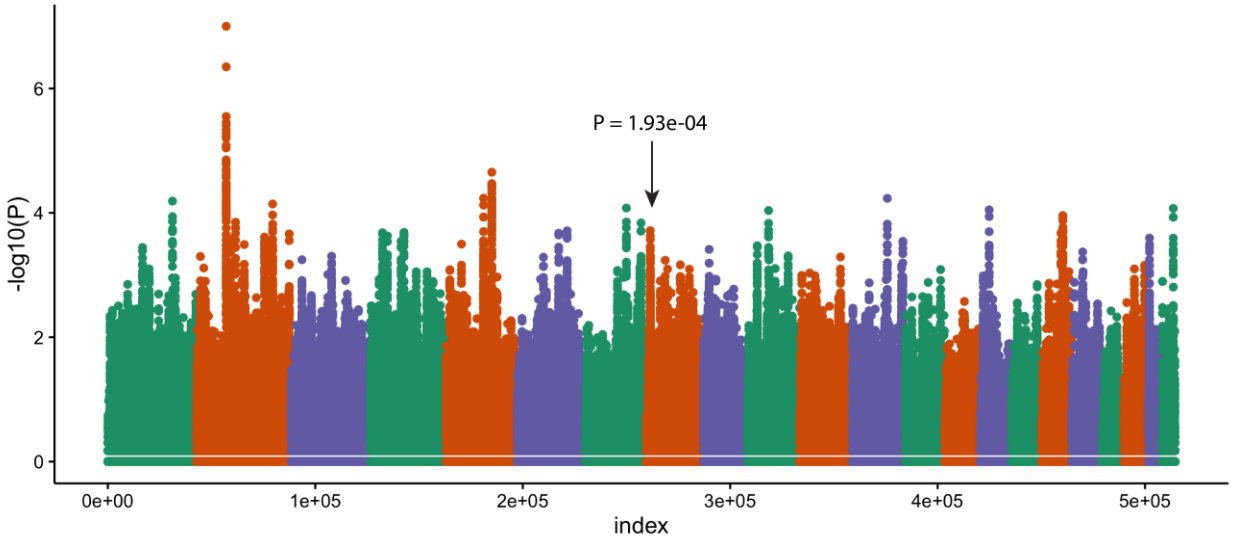
Supplemental Fig. 21: Percent identify calculated in sliding windows across the Chromosome 8p23.1 critical region. Percent identity is calculated in 2 kbp sliding windows (100 bp step) using the program Align Slider based on a 3.6 Mbp pairwise alignment of the Chromosome 8p23.1 critical region. Percent identity is plotted against aligned bases between the H1 and H2 haplotypes. Six regions of perfect sequence identity are found to be shared between H1 and H2 (red dashed boxes) and regions exhibiting excess sequence diversity are annotated by sites of fosmid haplotype sampling.

Supplemental Fig. 22



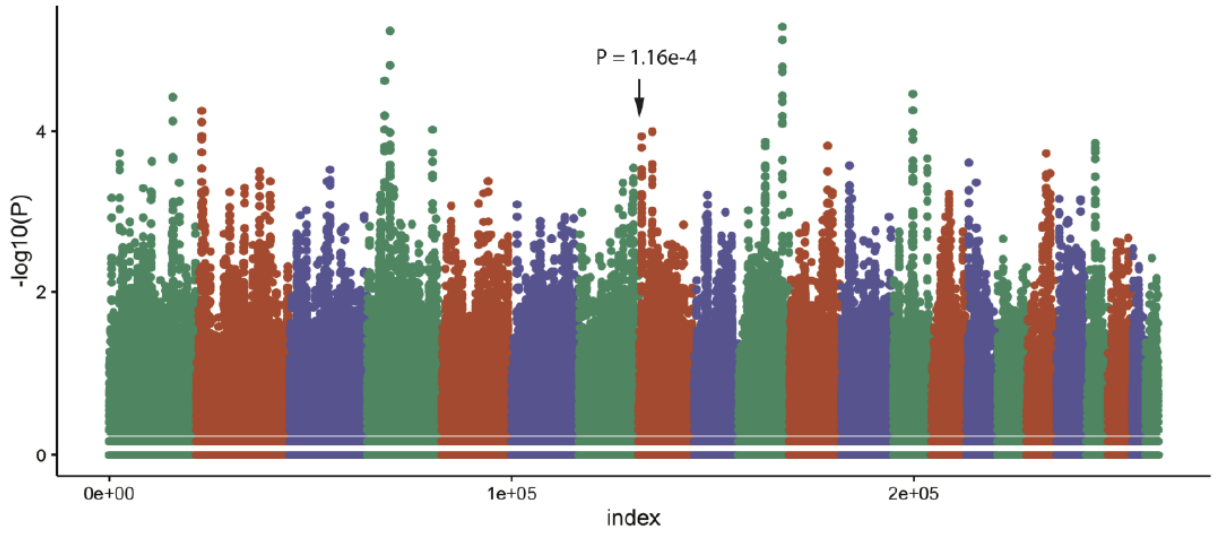
Supplemental Fig. 22: Evolutionary and phylogenetic analysis of the H1-associated EHH region. An unrooted neighbor-joining tree was constructed using the MEGA5 complete deletion option based on ~26 kbp of aligned sequence (Chr8:10,815,626-10,839,946) within a ~75 kbp region of eHH detected on the H1 haplotype. Haplotypes homozygous for inverted and non-inverted haplotypes (H1: GRCh37, NA18956 and H2: CHM1, NA12878 and NA19240) demonstrate a tree topology that completely separates the H1 and H2 clades (99% bootstrap support). Individuals heterozygous for inversion status (NA18507, NA18555, NA12156, and NA18517) cluster with the corresponding H1 and H2 clades relative to the haplotype that was sequenced.

Supplemental Fig. 23



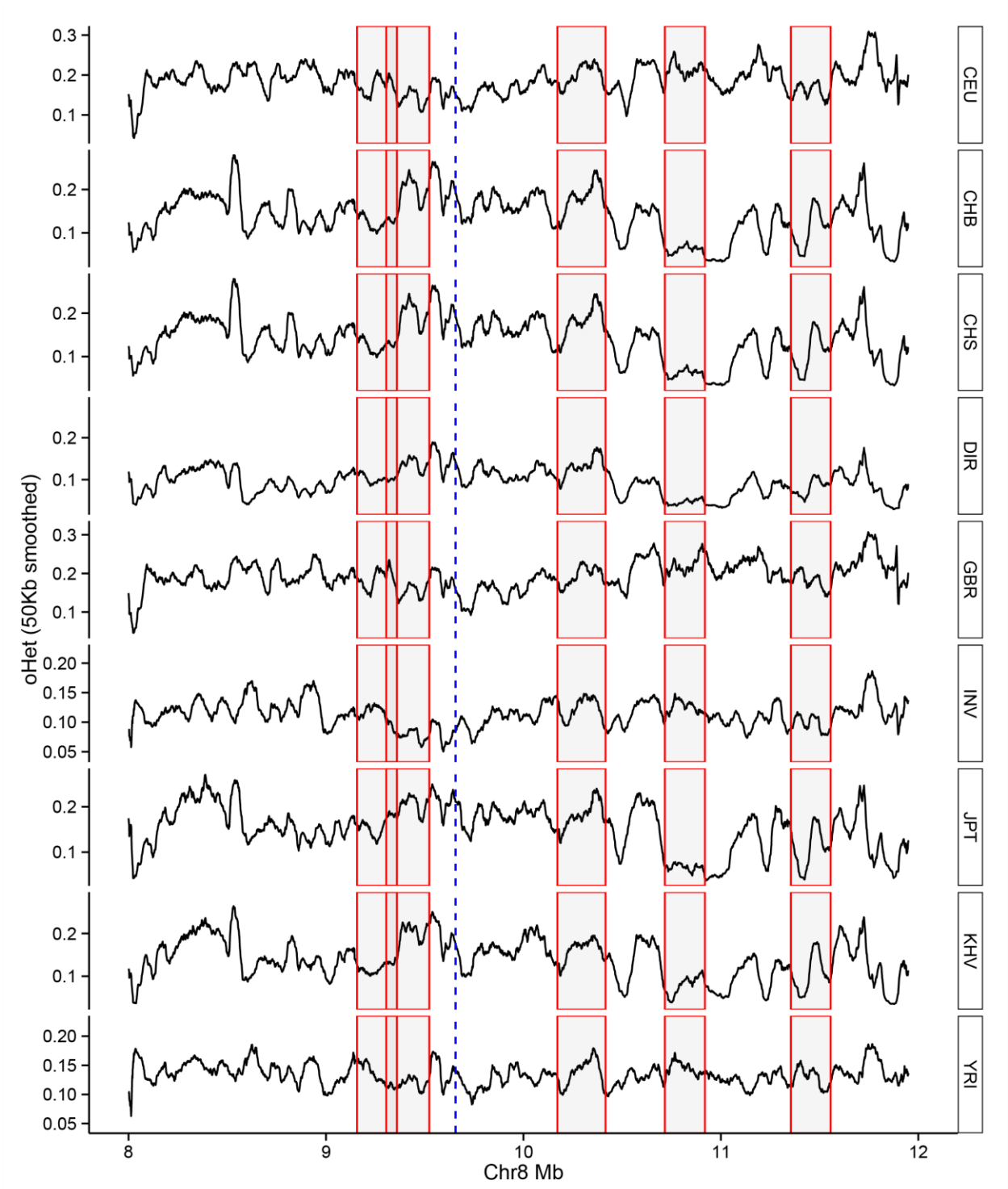
Supplemental Fig. 23: Genome-wide eHH for HGDP individuals homozygous for the H1 (direct) haplotype. The x-axis is an index for each 100 kbp window and the y-axis is the empirical probability for the window. The empirical probability is the number of times a window, with the same number of eHH observations has a higher average score. The H1 region is denoted with the arrow and reported probability.

Supplemental Fig. 24



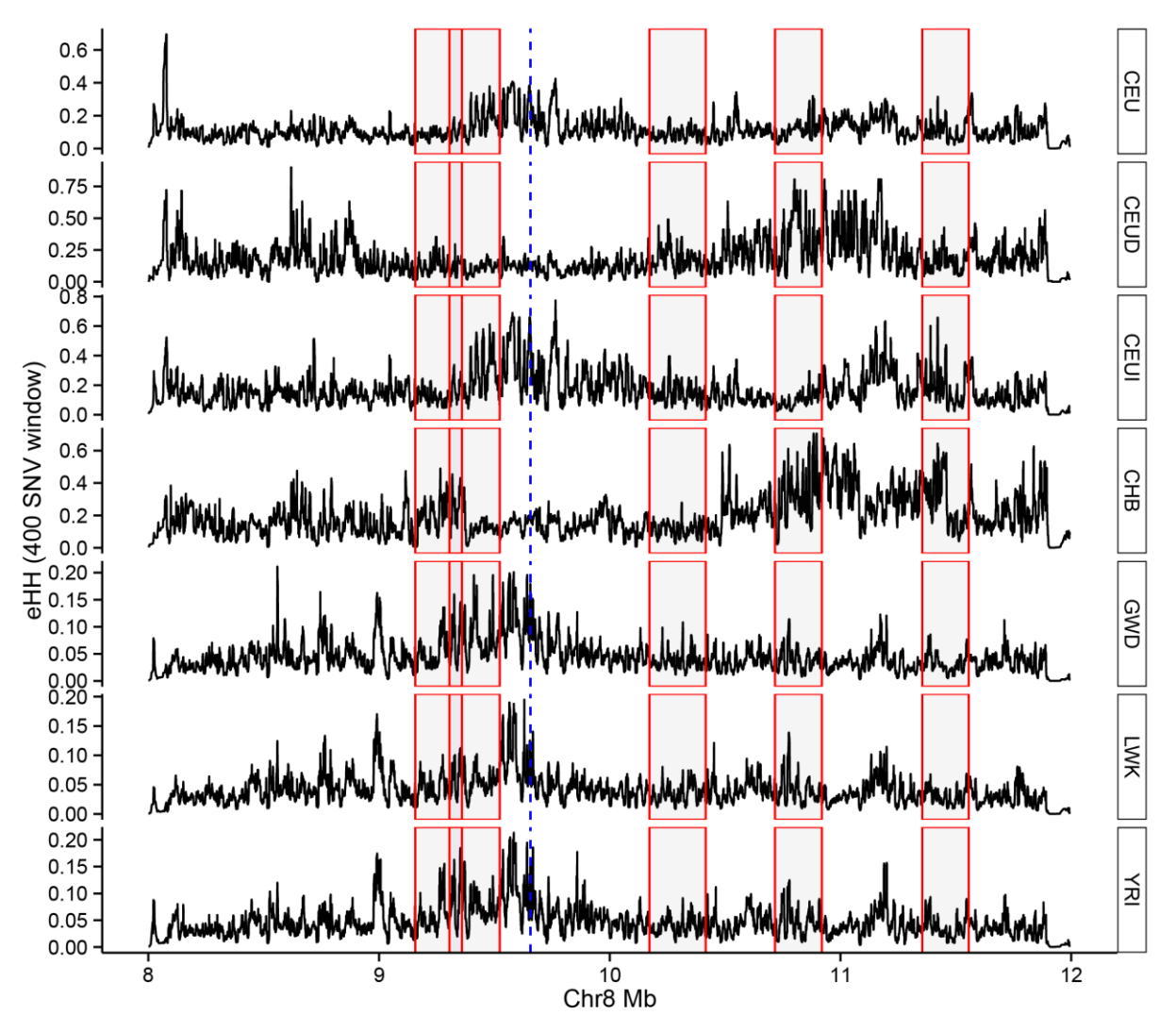
Supplemental Fig. 24: Genome-wide eHH for HGDP individuals homozygous for the H2 (inverted) haplotype. The x-axis is an index for each 100 kbp window and the y-axis is the empirical probability for the window. The empirical probability is the number of times a window, with the same number of eHH observations has a higher average score. The H2 region is denoted with the arrow and reported probability.

Supplemental Fig. 25



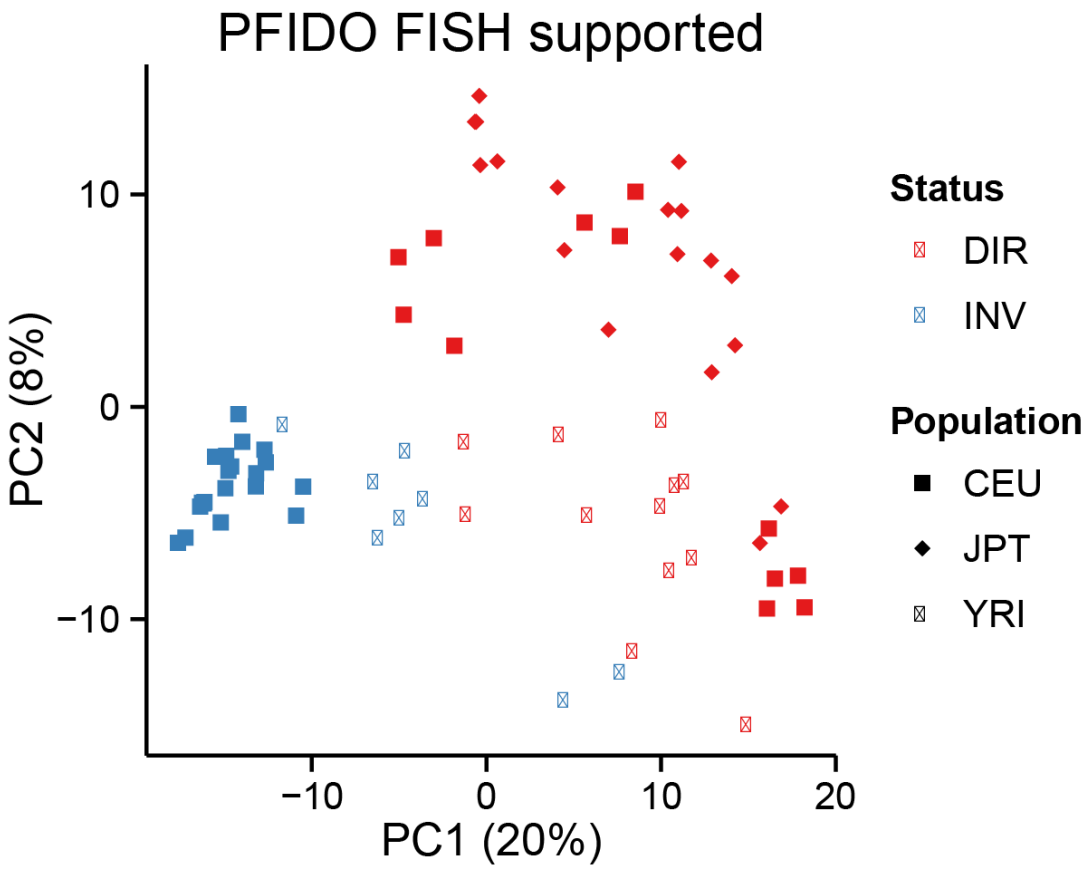
Supplemental Fig. 25: The observed heterozygosity for the phased 1KG individuals at the Chromosome 8p23 inversion. INV and DIR stand for individuals homozygous for either the inverted or direct haplotype, respectively. The red boxes denote the position of the sequenced fosmids. The blue dashed line corresponds to rs4841222 (see Figure 5). Heterozygosity was smoothed in a 50 kbp window with a 1 kbp step. There is also a drop in heterozygosity on the inverted haplotype near rs4841222. oHET: observed heterozygosity.

Supplemental Fig. 26



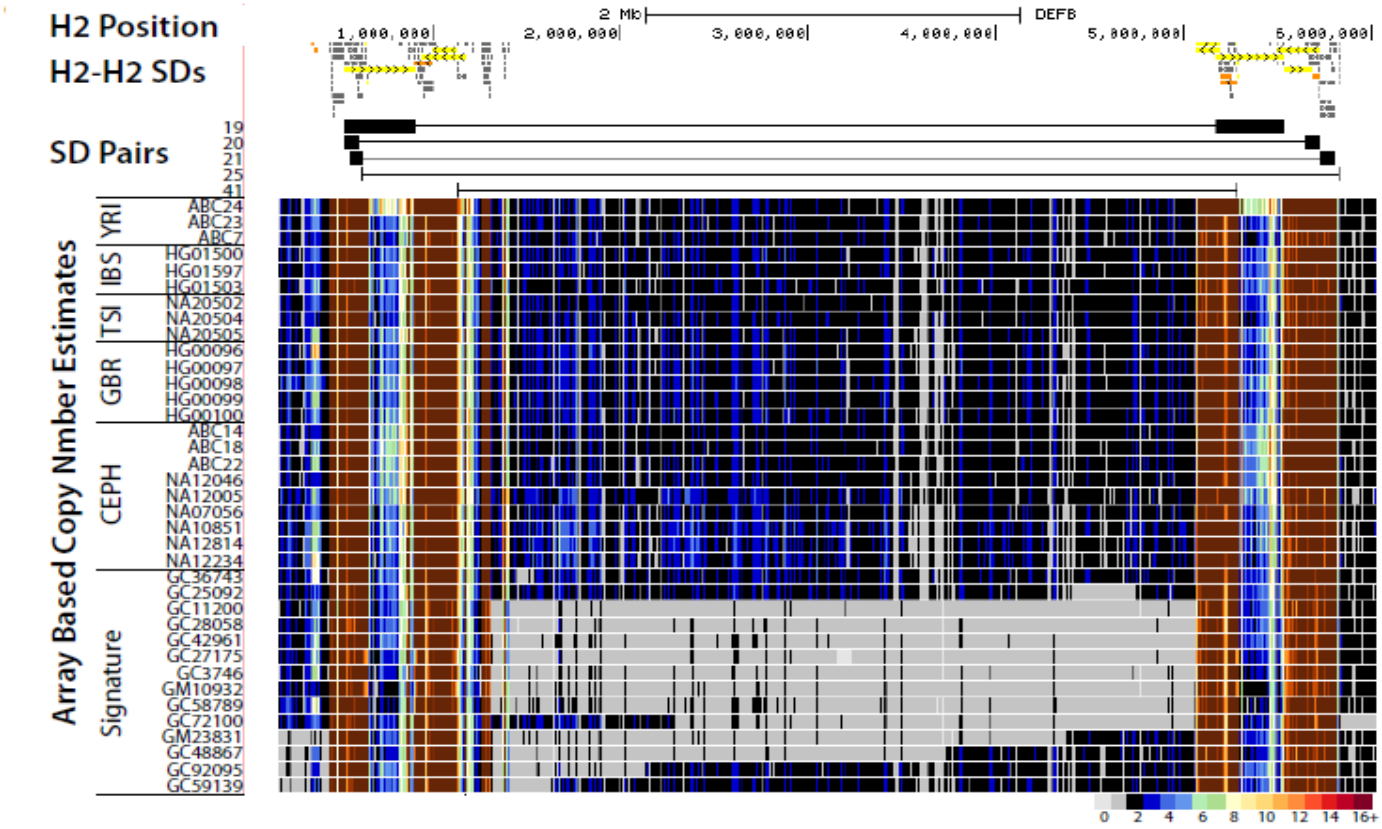
Supplemental Fig. 26: eHH for the 1KG within the Chromosome 8p23.1 inversion. The red boxes represent the sequenced clones and the blue line corresponds to rs4841222. CEUI are CEU individuals homozygous for the inverted haplotype and similarly CEUD are homozygous for the direct haplotype.

Supplemental Fig. 27



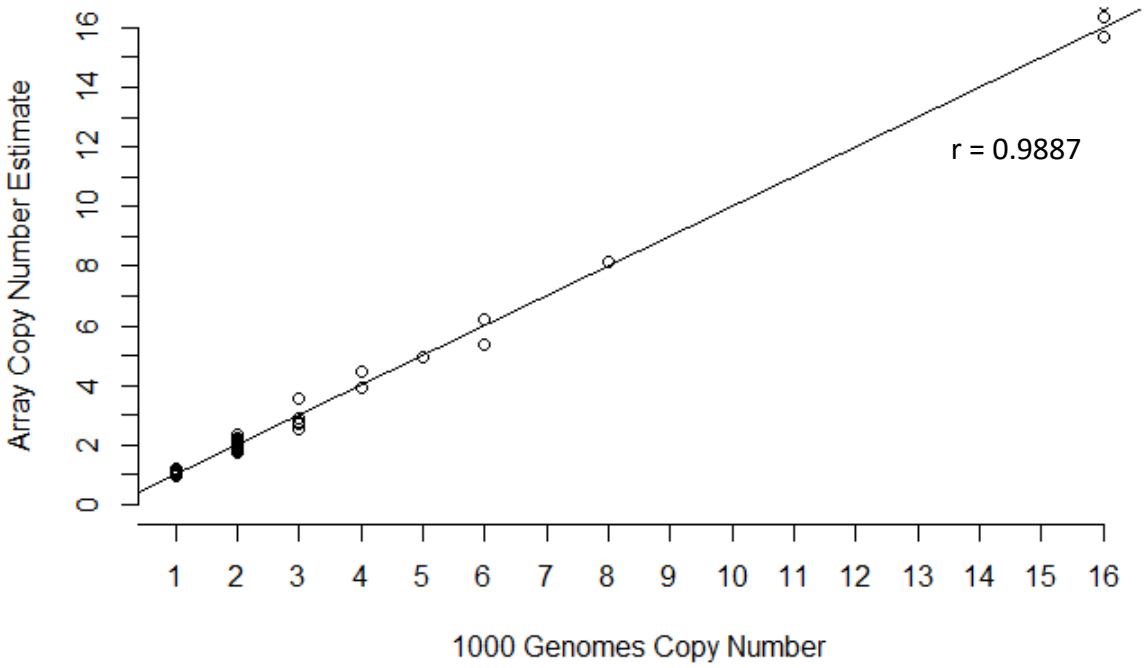
Supplemental Fig. 27: PCA for FISH validated inverted/direct haplotypes within the 1KG dataset.

Supplemental Fig. 28



Supplemental Fig 28: Array CGH copy number analysis. Array CGH-based copy number heatmaps for HapMap control individuals and cases with deletions encompassing the Chromosome 8p23.1 critical region are shown in the context of the H2 haplotype and the local SD structure. Large directly oriented SD pairs that potentially mediate deletions are shown as connected black bars. Profiles for GC27175 GC3746, GM10932 and GC58789 demonstrate significant depletion of signal in the proximal region between SD pairs 19 and SD pairs 20/21, indicating that a subset of these cases have breakpoints in the DA associated with SD20 or SD21 rather than the larger SD19.

Supplemental Fig. 29



Supplemental Fig. 29: Calibration of array signal intensity data to read-depth-based copy number estimates targeting Chromosome 8p23.1 segmental duplications. Robust read-depth copy number estimates from 23 HapMap individuals are strongly correlated (r = 0.988) with estimates obtained using array CGH signal intensity data.