Supplemental Figure Legends

Supplemental Figure S1. FISH to determine 8p23 inversion status. The upper panel illustrates the experimental design. The orthodox inversion probes (RP11-399J23 & RP11-589N15), labeled with red and green fluorophores respectively are targeted to the inversion region, adjacent to the flanking Low Copy Repeats (REPD & REPP). The control probe (RP11-73M19) labelled with both red and green fluorophores hybridizes to the centromeric region (*cen*) of chromosome 8, confirming chromosome 8 identity and no accidental shift in the x-y plane during image acquisition. The lower panels represent examples of homozygotes (GM18856 & GM18968) and a heterozygote (GM12156) for the 8p23 inversion; the *N* (Non-inverted) allele denotes the reference orientation as shown in the upper panel (i.e. Order = Telomere / RP11-399J23 / RP11-589N15 / RP11-73M19), whilst the *I* (Inverted) allele denotes the inverted orientation, (i.e. Order = Telomere / RP11-589N15 / RP11-399J23 / RP11-73M19).

Supplemental Figure S2. An overview of the PFIDO algorithm. Each block represents successive stages of the PFIDO algorithm. Specific details can be found in the main text.

Supplemental Figure S3. A subset of HapMap SNPs effectively capture 8p23 genetic substructure. PFIDO was run iteratively on either HapMap Phase II (n=54) or Phase III (n=110) CEU founders in cumulative 25-SNP steps. The proposed clustering solution at each iteration was assessed using 3 internal clustering measures; Connectivity reflects the extent to which observations are grouped with their nearest neighbours, whereas the Dunn Index and Silhouette Width combine metrics of intra-cluster homogeneity and inter-cluster separation. The Connectivity and Dunn Index range from 0 to ∞, and are optimal when minimized and maximized respectively, whilst the average Silhouette Width ranges from -1 to 1, and is optimal when maximized. Regions in which the plot line is interrupted reflect intervals in

which PFIDO mis-classified ≥1 FISH-defined inversion-type. The dashed vertical red line marks SNP numbers (802 & 123) that optimize the three measures.

Supplemental Figure S4. 8p23 inversion-specific SNP content in 3 commercial genotyping arrays. The Venn diagram illustrates the number of overlapping SNPs found between the Affymetrix GeneChip Human Mapping 500K, Affymetrix Genome-Wide Human SNP Array 6 and Illumina HumanHap550 arrays mapping to the inversion interval. Illustration prepared using UCSC genome tables.

Supplemental Figure S5. Genetic substructure in the 8p23 inversion region in the JPT and YRI populations. MDS analyses based on 4595 and 4150 SNP genotypes typed in 41 JPT and 111 YRI founder members respectively are shown, overlaid with FISH data and PFIDO predictions. The inset panels illustrate the frequency distributions of the sample points along the first dimension.

Supplemental Figure S6. Empirical null distributions indicate no strong evidence of selection. **A)** The correlation (Spearman's rho) between distance from Addis Ababa and allele frequency. **B)** Global F_{st} values. **C)** Pairwise F_{st} values. All distributions were constructed using SNPs with global MAF > 0.4 in the HGDP dataset. Dashed and solid vertical black lines indicating the position beyond which 5 and 1% of these SNPs fall respectively. The position of 8p23 Inv in the distribution is included where applicable.

Supplemental Figure S7. Box-plot illustrating the association between *PPP1R3B* expression levels and inversion status in Stranger *et al* (2007). The categorical x-axis refers to inversion status, with sample population labels given beneath.

Supplemental Figure S8. Ancestral crossover events in 8p23 inv heterozygotes. A) HAPMIX crossover detection between segments of inferred ancestry in an *IN* individual. The

inversion interval is flanked by REPD/REPP. Each point represents a SNP. The transfer of *N*-derived alleles into the central section of this individual's *I* chromosome is detected at the point where the inferred number of *N* alleles switches from 1 to 2, representing an ancestral double crossover event. The blue shaded strips mark the 95% confidence interval for the inferred allele number. **B)** The distribution of crossover positions in the inversion interval, derived from the 45 inferred double recombinants. Crossover locations predominantly map to the centre of the inversion (9-10 Mb). **C)** F_{st} values between the 45 *IN* samples with evidence of an ancestral recombination event and the reference samples (*II* or *NN*) were calculated for three 1Mb sections across the inversion. On the whole, the central segment in the *IN* samples showed greater genetic similarity to the *NN* samples.

Supplemental Figure S9. REPP is absent in orangutans. Eleven gorilla BAC end-sequence pairs and five sequenced orangutan BACs and were aligned to the human reference sequence (hg19); the orientation of the reference *8p23-inv* was inverted, to reflect the ancestral orientation and the positions of four genes (*XKR5, CTSB, PRAGMIN & LONRF1*) are provided as points of reference. Alignments are represented as blocks joined by lines, and are annotated with their corresponding GenBank accession number/clone name. The position of syntenic REPP/REPD is marked in pink. Two orangutan BACs span REPP entirely, indicating the absence of REPP in this species.

Supplemental Figure S10. The formation of REPP dates to ~9.5 mya. A) Based on read-depth mapping of whole-genome shotgun sequences against the human REPP, a section appears duplicated in humans and chimpanzees but not in orangutans, and thus may represent the pre-duplication state (chr8:12,592,000-12,597,076 (hg17); red = excess coverage depth of aligned whole-genome shotgun sequences, green=standard coverage depth; data from http://humanparalogy.gs.washington.edu; (Marques-Bonet et al. 2009)). **B)** The section's first duplication occurred ~9.5 mya (assuming the orangutan-human lineage split occurred ~14mya), resulting in separate REPP/REPD lineages. Analyses involved aligning human

genome reference sequence representing syntenic REPP/REPD with homologous chimpanzee and orangutan sequence (clone names are provided where applicable; see *8p23 LCR re-assembly* in supplementary note for an explanation of REPP/REPD nomenclature). Evolutionary history was inferred using the Neighbor-Joining method based on Kimura 2-parameter genetic distance. The linearized bootstrap consensus tree (1000 replicates) is given; the percentage of replicate trees in which the associated tips clustered together is annotated next to the branches. Analyses were conducted in MEGA4. **C)** The distribution of pairwise TMRCA estimates between REPP/REPD sub-units. Seven ancestral LCR sub-units (SD4887, SD4890, SD3693, SD3719, SD3720, SD3733 and SD4912; (Jiang et al. 2008)) collectively representing 23,866bp were identified in orangutan BACs (AC205861.3, AC206882.3, AC207519.3 and AC206098). Only duplicons with < 2 non-overlapping alignments in an orangutan BAC were used, to reduce the influence of ancient copy-number variants in the analysis. Each orangutan duplicon was subsequently aligned to RPCI-11 REPP/REPD using MUSCLE, the null hypothesis of equal rates between lineages assessed using the Tajima relative rate test (rejecting alignments with $p < 0.05$), and age estimated as described in Zody *et al.* (2008), assuming the orangutan-human lineage split occurred ~14mya.

Supplemental Figure S11. A shared deletion in African Great Apes in the syntenic REPP. All available trace data was downloaded from the Trace Archive (ftp://ftp.ncbi.nih.gov/pub/TraceDB/; download date 26/10/2011) for HuRef (*Hsa*), *Pan troglodytes (Ptr)*, *Gorilla gorilla (Ggo)*, *Pongo abelii (Pab)* and *Nomascus leucogenys* (*Nle*). These were aligned to AC212986 using MegaBLAST, retaining alignments with PID > 90% and length > 500; this Orangutan BAC maps to single-copy regions in the Human reference assembly either side of REPP, as indicated by arrows (hg19 mapping). The trace coverage (i.e. the amount of times a base is covered) is plotted for each species. In Gorilla, Chimp and Human traces, a ~28 kb segment is not represented, suggesting absence of this segment in these species. Two adjacent duplicons (SD800/SD5601, containing olfactory receptor paralogues (OR7E)) border this segment and appear to have expanded in copy-number in

most species. High-copy repeats (HCRs) with < 10% divergence are indicated by dashes; these regions were masked during alignment. To simplify interpretation, African great ape coverage is plotted above the horizontal and while Orangutan and Gibbon coverage is plotted below the horizontal.

Supplemental Figure S12. A network representing duplicon similarity between RPCI-11 BACs. Each node represents an RPCI-11 BAC, and each edge's colour represents the number of times two BACs share highly similar duplicons (*see key*). Nodes in green have previously been assigned to chromosome 8 by the BAC's submitter, while those in purple have been assigned to other chromosomes.

Supplemental Figure S13. The BAC composition and assembly accuracy of the 8p23.1-p22 LCRs. The first four panels (i - iv) represent LCRs A to D. LCR-A/B and LCR-C/D correspond to REPD and REPP respectively, with each LCR pair separated by an assembly gap in the reference genome build (hg18). In each of these panels, the reference genome build (hg18) is represented by the "Map Contigs" track. The International Human Genome Consortium assembled this from the sequenced clones beneath the track (in orange); red sections represent single-copy sequence. Each clone is annotated with its library name, as well as its GenBank accession code (+version) in the left-hand column of the panel. The track above the reference genome build (in blue) represents re-assembled finished RPCI-11 BACs that map to the LCRs and align to form haplotype-specific assemblies. The alignment accuracy for each RPCI-11 BAC overlap indicated in panels i-iv (overlaps a-i) is given in panel v.

Supplemental Figure S14. Dotplots illustrating the high sequence homology between the RPCI-11 LCR haplotypes A-D. The x-axis indicates position (in Mb) on chromosome 8 (hg18), whilst the y-axis indicates position in the LCR (in Mb). Diagonal lines in the forward orientation ("/") represent homologous segments in direct orientation between the compared

LCRs, whilst diagonals in the reverse orientation ("\") represent homologous segments in indirect orientation that are potential sponsors of inversion formation via NAHR. The level of sequence identity between sequences is colour-coded (*see key*).

Supplemental Figure S15. Sequenced BACs in multiple libraries support the recombinant LCR patterns. See Figure 5B (main text) legend for details.

Supplemental Figure S16. Discordant fosmid ESPs resolved by finished BACs. Only BACs that resolved ≥10 ESPs are plotted. For each BAC, resolved ESPs are plotted as two symbols joined by a line. The symbol denotes the source library's inversion-type: *II* by red diamonds, *IN* by blue squares and *NN* by green triangles (*see legend, bottom left-hand corner*). The adjoining line is coded by the RPCI-11 LCR haplotypes relative to which the ESPs are discordant in orientation: for example, an ESP mapping to the same strand of LCRB is denoted by a solid purple line (*see legend*). The clone name for each ESP is provided, and includes the source library (ABC7-27). The BAC LCR duplicon annotation is given at the top of each plot, with those on the positive strand plotted above the horizontal.

Supplemental Figure S17. Analysis of "recombination event 5". A) Phylogenetic analysis of sequence flanking the proposed breakpoint. Bootstrap consensus neighbour-joining trees were constructed using ~6kb sub-alignments, each flanking the proposed breakpoints in the multiple sequence alignment. Clades are shaded in green, yellow and red, corresponding to the junction, LCR-D or LCR-B related haplotypes respectively. Branches marked with an asterisk (*) are supported by bootstrap values >95% (1000 bootstrap replicates). Genetic distance is caclulated according to the Jukes-Cantor model. **B)** Haplotype-specific PCR sequencing products. Nucleotide variants that distinguish haplotype-groups are shaded in grey. The sample identifiers underlines are coloured according to haplotype-groups as in *a*. See Supplemental Table S7 for haplotype-specific primer pairs.

Supplemental Figure S18. A median-joining network of phased CEU haplotypes in the breakpoint's vicinity ("recombination event 5") supports the relationship between *I*-carriers and presence of the LCR-B related haplotype. The median-joining network (Bandelt et al. 1999) was constructed using 55 HapMap SNPs (rel. 02/2009, trio-derived haplotypes only) phased in 45 samples predicted to be either *II* or *NN* at *8p23-inv* using SplitsTree4 (Huson et al. 2006); inversion heterozygotes were omitted to avoid errors introduced by an additional phasing step. Although a 200kb window centered around "recombination event 5" (chr8:8000317-8200317) was selected, the majority of the SNPs (83%) map to a 65kb single-copy region flanking the LCR. Haplotype groups are denoted by circles, with the circles' size proportional to the number of chromosomes carrying a particular haplotype. Each circle's color represents the proportion of *II* (blue) or *NN* (red) derived haplotypes at that node. The distance between nodes is proportional to the number of differences distinguishing each haplotype. If a haplotype represents >2 chromosomes, then sample identifiers were grouped and annotated on either side of the network (Groups 1-6). The actual LCR-B haplotype clusters with other *II*-derived haplotypes in Group 3. The sample identifiers contain 4 attributes, delimited by an underscore: inversion-type, positive/negative/untested (+/-/?) for the LCR-B related haplotype in the ARMS assay, Coriell repository sample ID and chromosome (c1/c2). After segregating the network into two clades (dashed horizontal line), 92% of the *II*-derived haplotypes are located in the lower clade, whilst 96% of the *NN*-derived haplotypes are found in the upper clade. Moreover, clade membership correlates perfectly with presence/absence of the LCR-B related haplotype (Fisher's exact test, $p=4.01 \times 10^{-7}$). As such, two out of the three samples not exhibiting a relationship between the haplotype and inversion-type can be accounted for: NA11881 (*II* but negative for the haplotype) has both chromosomes in the upper clade, whilst NA12155 (*NN* but positive for the haplotype) has a chromosome in the lower clade. The third outlier (NA06989) did not qualify for median-joining network analysis as the data had been phased statistically rather than by transmission. This suggests that the imperfect correlation between the LCR-B related haplotype and *I*-carriers is not the result of confounding local gene

conversion, to which LCRs are prone (Chen et al. 2007), but that a rarer alternative breakpoint exists in the ancestrally European samples.

Supplemental Table Legends

Supplemental Table S1. Inversion status determined by FISH in 103 HapMap individuals. "*N*" denotes non-inverted orientation, whereas "*I*" denotes inverted orientation, relative to the reference human genome assembly (NCBI build 36.1). An "f" denotes HapMap phase II CEU founder individuals. YRI and JPT samples exclusively represent sample founders. Samples used to verify Mendelian inheritance patterns are denoted by "T". Results confirmed by other studies are noted ([a] = Antonacci et al., 2009; [b] = Broman et al., 2003; [d] = Deng et al., 2008). All cell lines were obtained from Coriell Cell Repositories.

Supplemental Table S2. Optimal SNP sets for PFIDO use. Analysis performed using CEU (phase II/III) and YRI founders. CEU *8p23-inv* "tagging" SNPs (r^2>0.8) are marked with an asterisk in the first column.

Supplemental Table S3. SNPs genotyped in 1,748 white British individuals. SNP genomic positions were retrieved from BioMart v.0.7, based on the Human ENSMBL 55 variation dataset. The proportion of the cohort with missing genotypes at each locus is included ("Missing"). Pedigree founders were employed to calculate a HWE test statistic (reported as a p-value) for each SNP. SNPs were genotyped on the Illumina GoldenGate and Sequenom MassARRAY iPLEX Gold platforms (primer sequences available on request).

Supplemental Table S4. Global 8p23 inv frequencies. The allele frequency (AF) of the inverted allele in each population sample is provided, along with: a HWE test p-value, the sample's size (n) and estimated distance from Addis Ababa (AA), and the HapMap reference population with which the sample was "seeded".

Supplemental Table S5A. Associations between inversion status and mRNA level. For each of the five studies considered, p-values (corrected for multiple-testing) are reported for

associations between inversion status and a specific probe level or allele-expression (AE) window. The genomic region represented by the probe/AE window is provided. Effect size (beta coefficient) is expressed relative to the *N* allele; note that the magnitude of effect sizes is study-specific, and that the direction of association is comparable between studies. Dataset 1 = Stranger *et al* (2007); 2 = Ge *et al* (2009); 3= Dixon *et al* (2007); 4 = Schadt *et al* (2008); 5 = W.O. Cookson and M. Moffatt, unpublished data**. B)** Associations between inversion status and allele-expression (AE) window conditioned on specific SNPs. For the originally reported associations in Ge *et al* (2009), analyses were repeated to include inversion status as a dependent variable. Associations between inversion status and AE window are reported (p-value corrected for multiple testing; NS denotes $p > 0.05$). **C)** SNPs used to construct median-joining network of CEU haplotypes influencing BLK-FAM167A mRNA levels in Figure 4.

Supplemental Table S6. Estimation of the *8p23-inv* Evolutionary Age. Genomic intervals represented by RPCI-11 BACs (*BAC-1 or -2*), were aligned with either chimpanzee and orangutan (*HSA-PTR-PPY*) or only chimpanzee (*HSA-PTR*) orthologues. The Percentage Identity (*PID*) between the two RPCI-11 BACs is provided, as is the corresponding Kimura 2-parameter estimate (*K*). The time of divergence (in "millions of years ago" (mya)) was calculated using either chimpanzee (PTR) or orangutan (PPY) as an outgroup, assuming species divergence of 6 mya and 14 mya respectively. The standard error (*SE*) is derived from the Kimura 2-parameter SE. Tajima's Relative Rate tests indicated that the genomic sequences are evolving neutrally (p=0.34–1). *N.B.* The chimpanzee sequence is derived from an *NN* sample (Antonacci et al. 2009), whilst the inversion status of the orangutan sequence donor is unknown; to date orangutans have only been found to carry *II* (Antonacci et al. 2009).

Supplemental Table S7. "Amplification Refractory Mutation System" primers. The putative Paralogous Sequence Variants mediating specificity are underlined.

References for Supplemental Legends

Antonacci F, Kidd JM, Marques-Bonet T, Ventura M, Siswara P, Jiang Z, Eichler EE. 2009. Characterization of six human disease-associated inversion polymorphisms. *Hum Mol Genet* 18: 2555-2566.

Bandelt HJ, Forster P, Rohl A. 1999. Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* 16: 37-48.

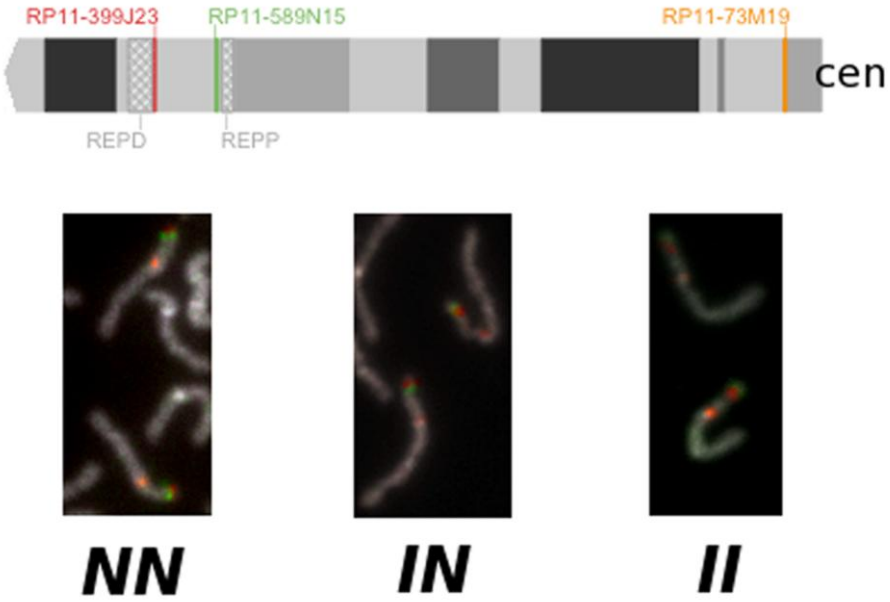
Chen JM, Cooper DN, Chuzhanova N, Ferec C, Patrinos GP. 2007. Gene conversion: mechanisms, evolution and human disease. *Nat Rev Genet* 8: 762-775.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* 23: 254-267.

Jiang Z, Hubley R, Smit A, Eichler EE. 2008. DupMasker: a tool for annotating primate segmental duplications. *Genome Res* 18: 1362-1368.

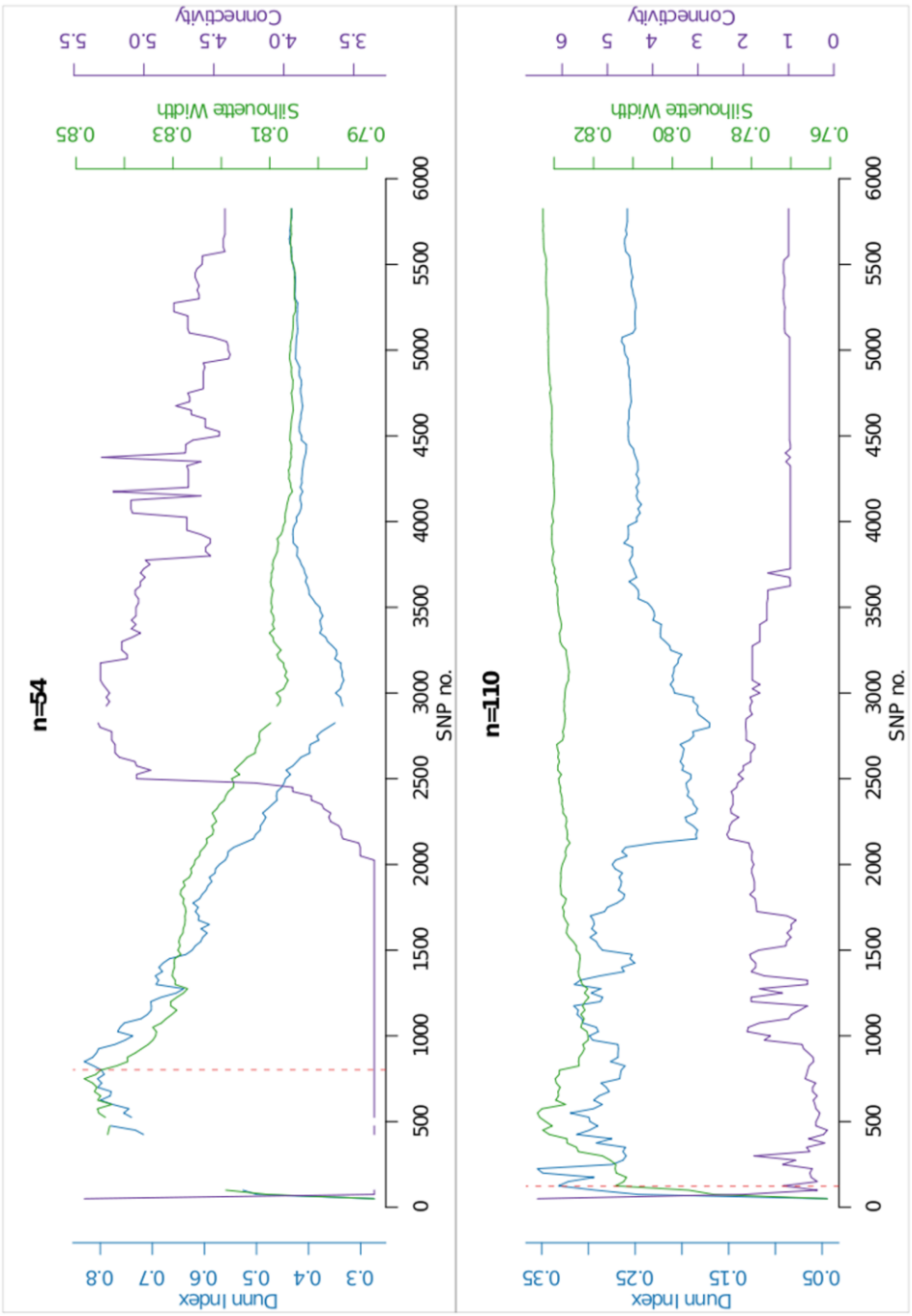
Marques-Bonet T, Kidd JM, Ventura M, Graves TA, Cheng Z, Hillier LW, Jiang Z, Baker C, Malfavon-Borja R, Fulton LA, et al. 2009. A burst of segmental duplications in the genome of the African great ape ancestor. *Nature* 457: 877-881.

Supplemental Figure S1

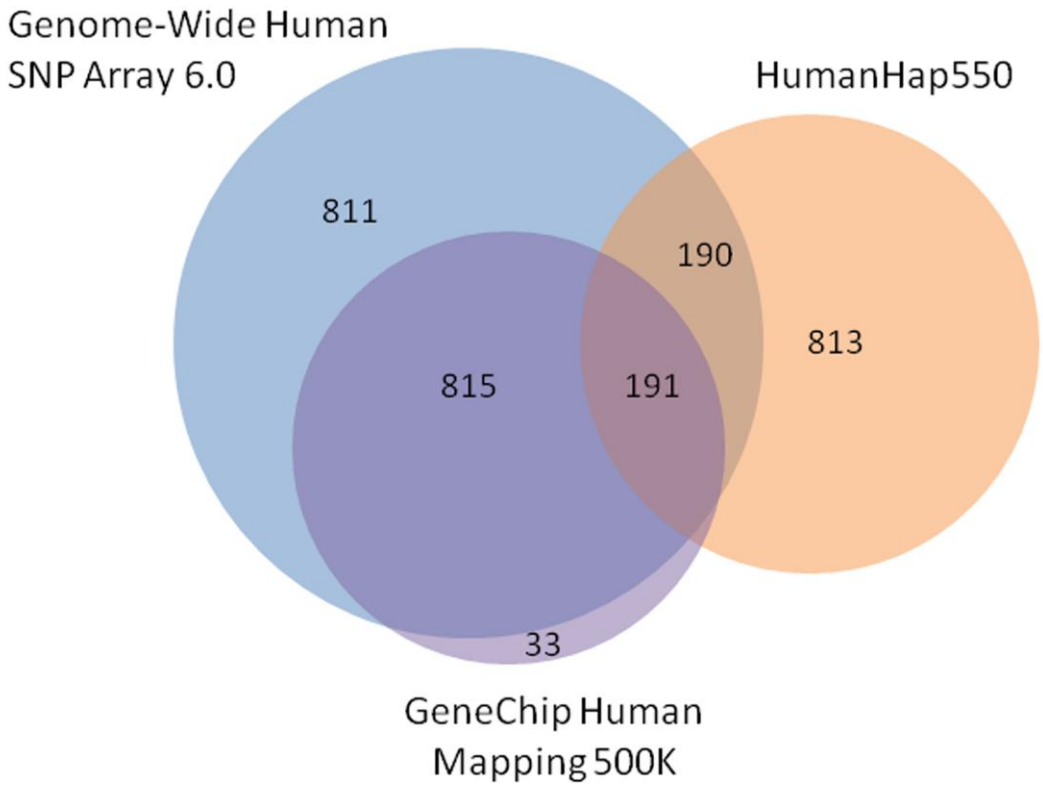


Supplemental Figure S2

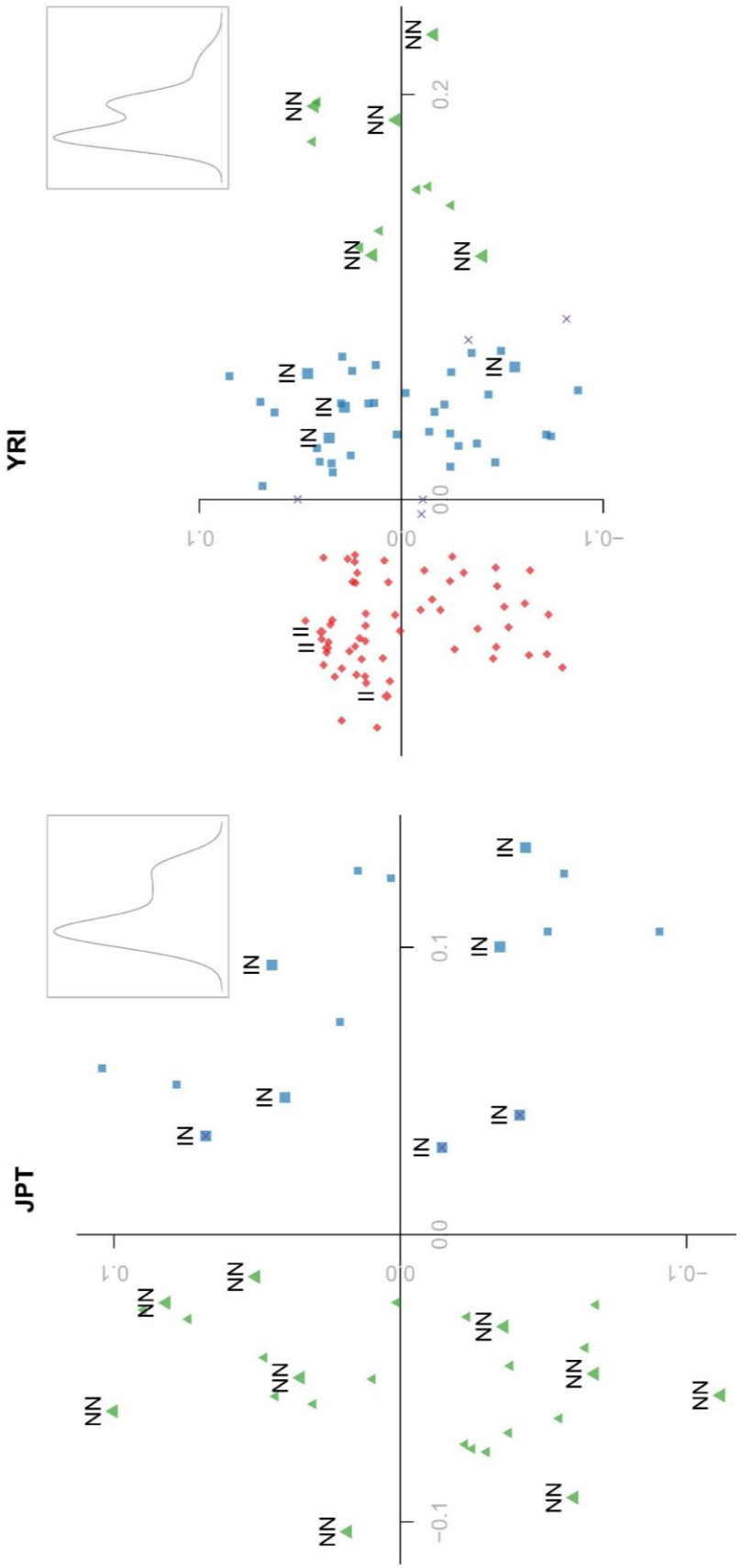
Supplemental Figure S3



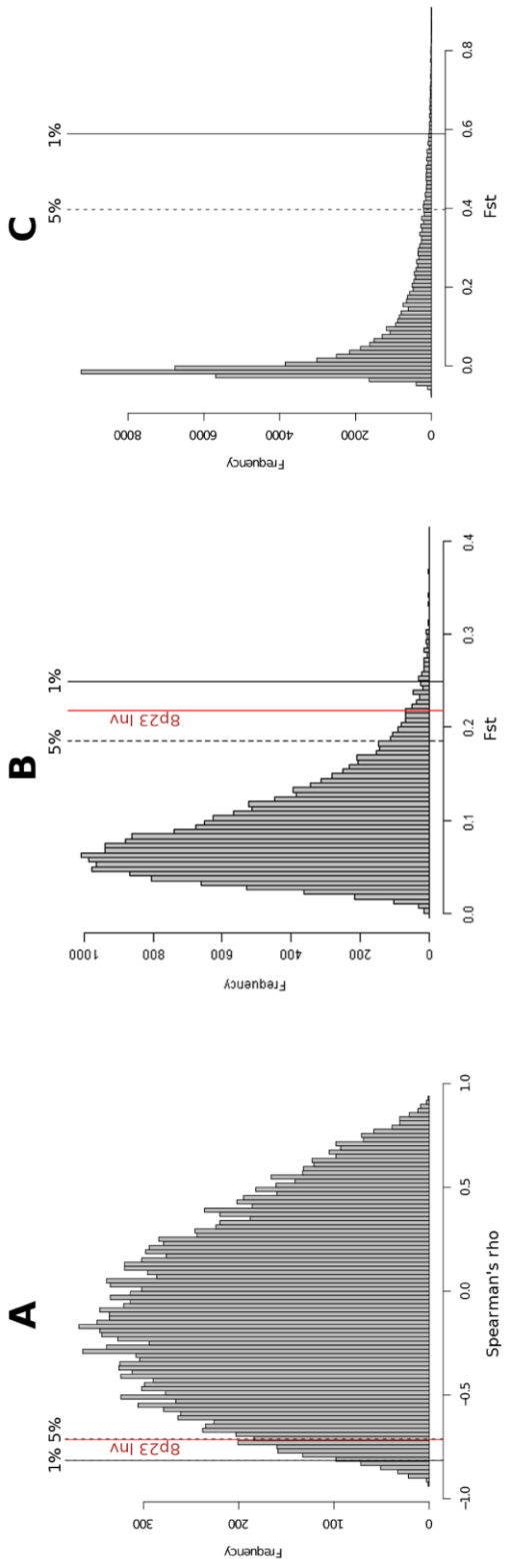
Supplemental Figure S4



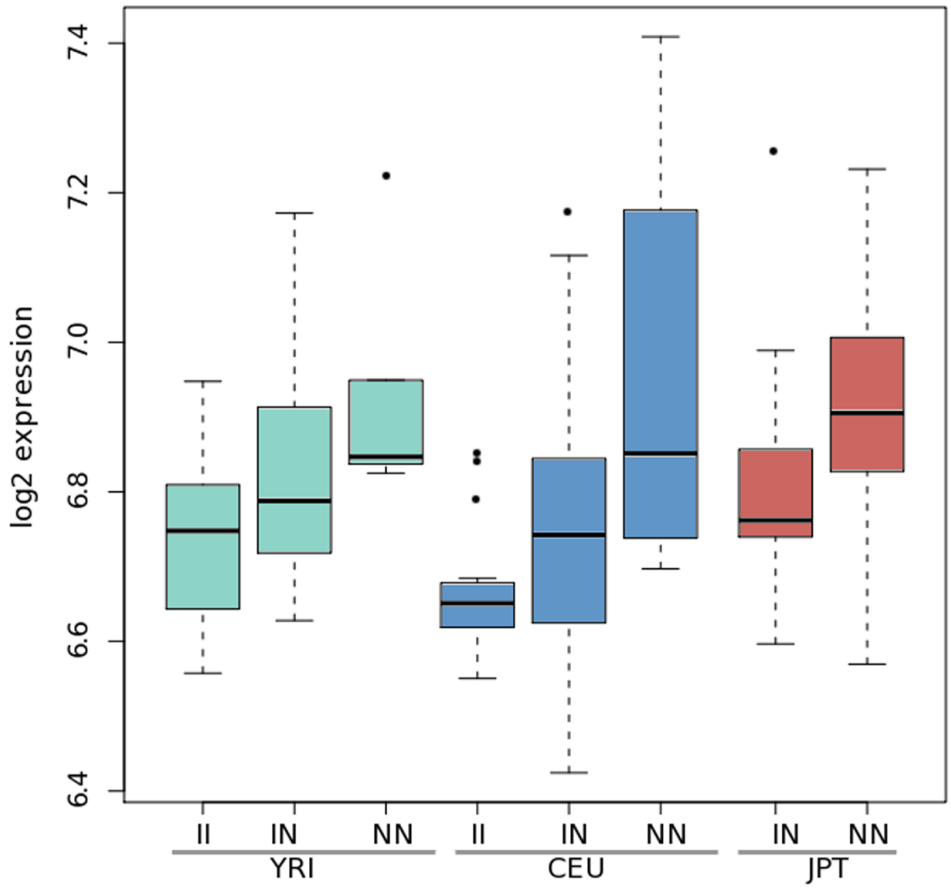
Supplemental Figure S5



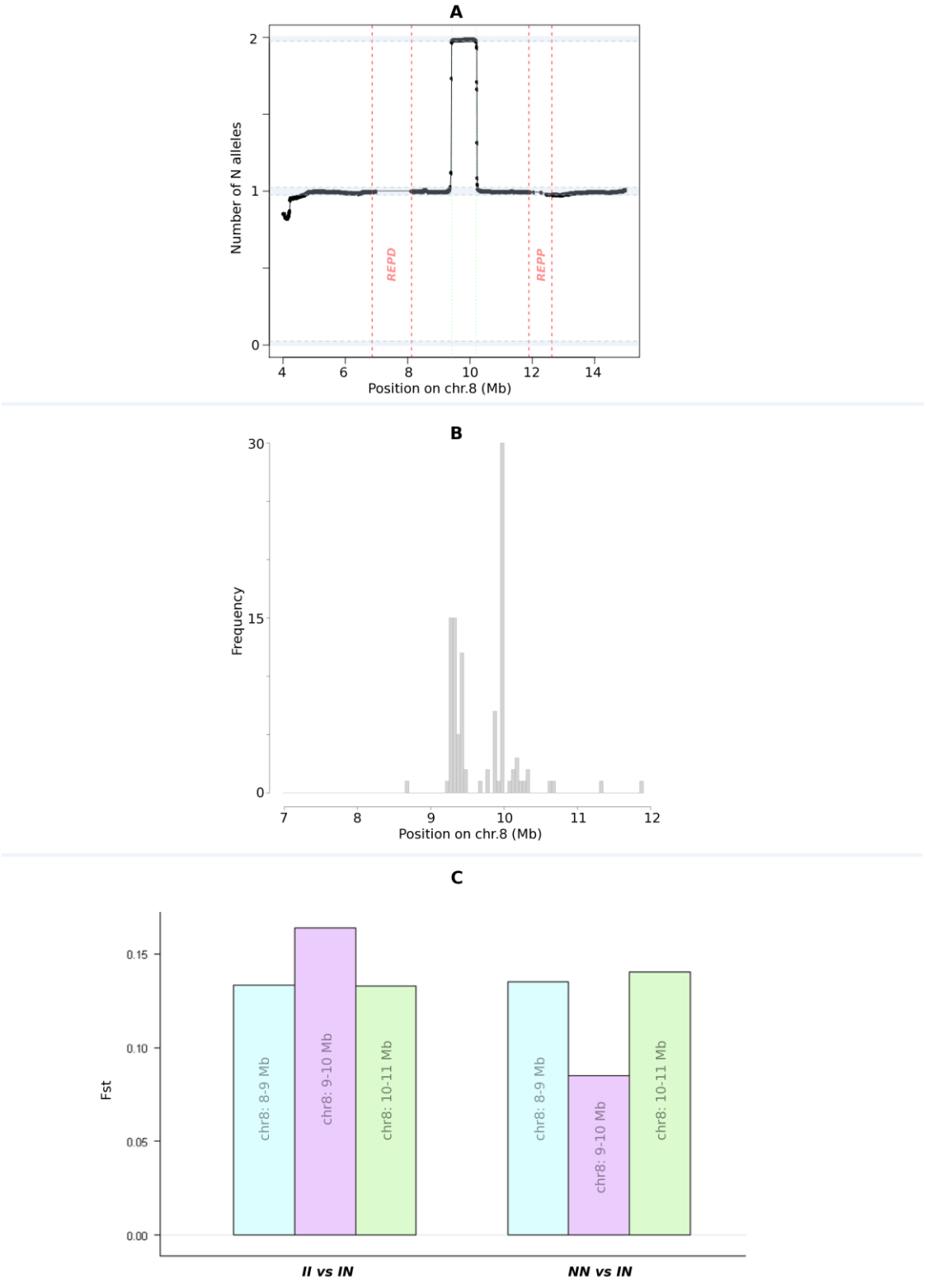
Supplemental Figure S6



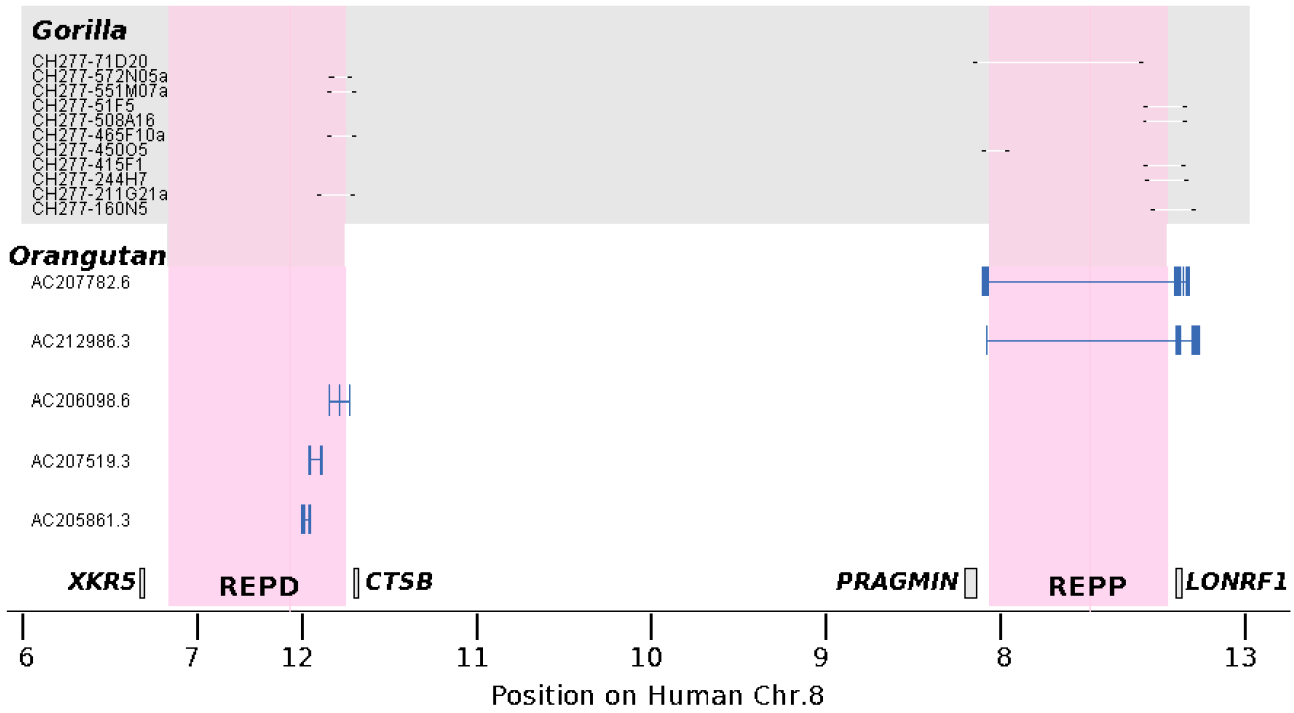
Supplemental Figure S7



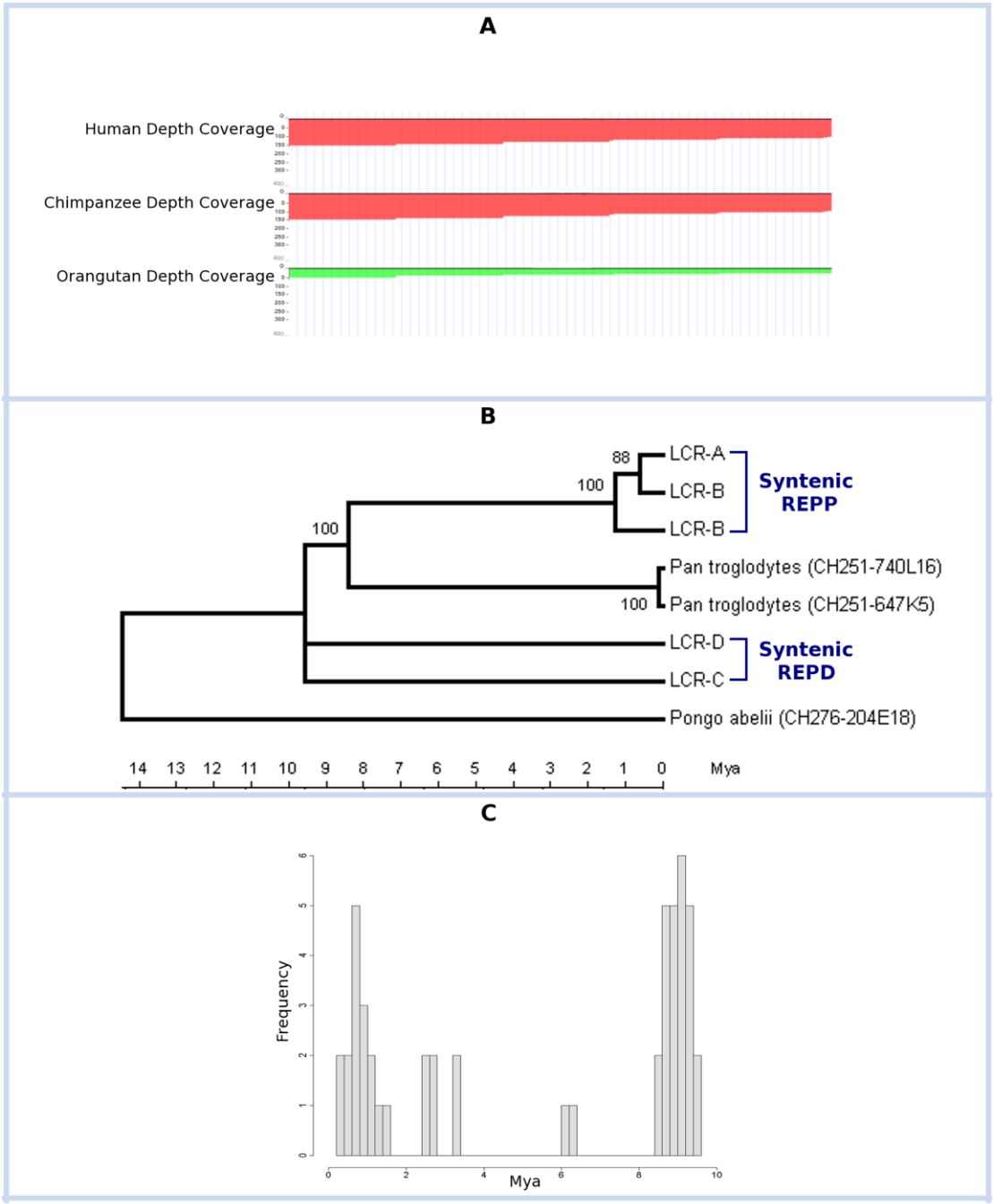
Supplemental Figure S8



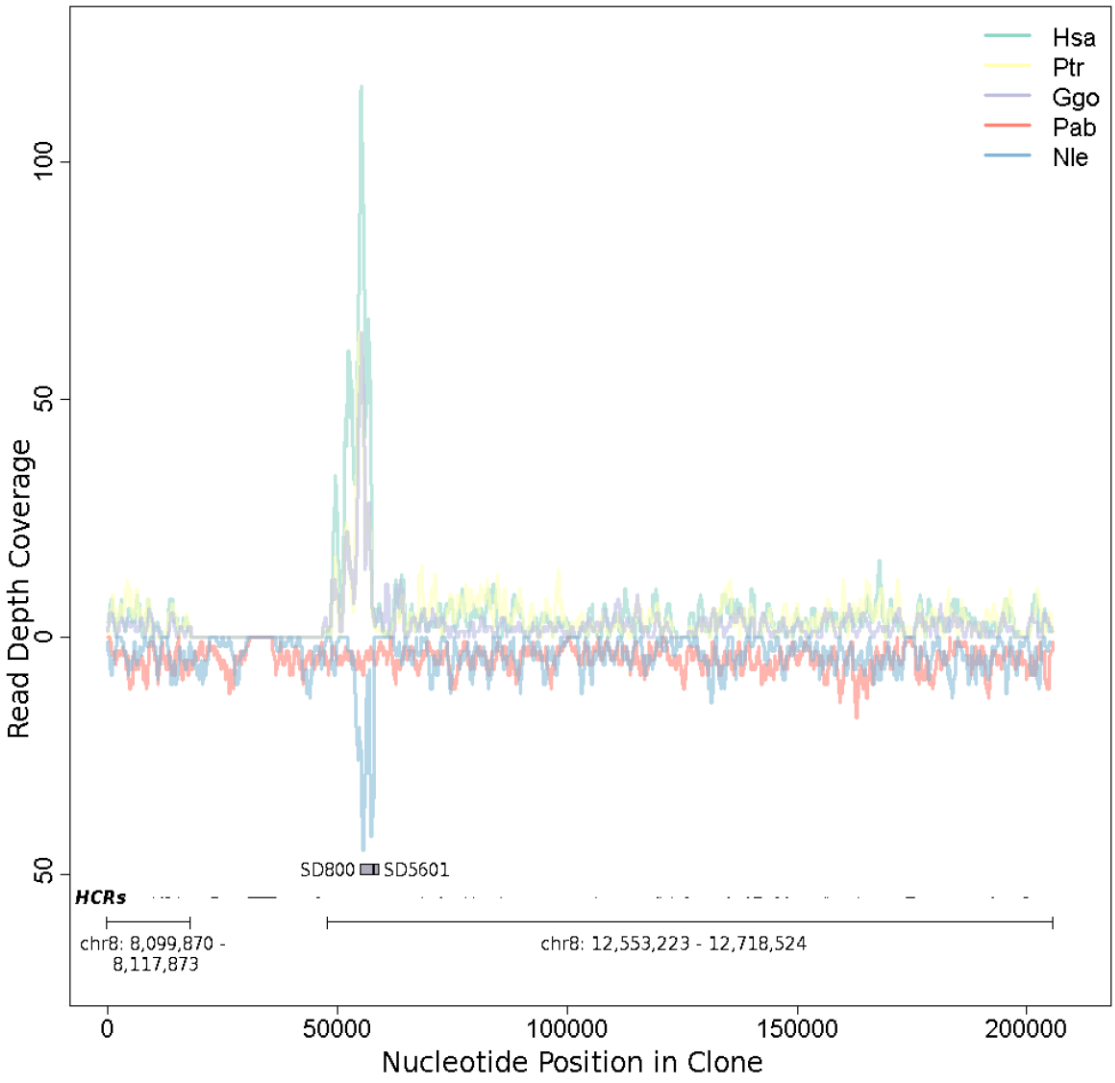
Supplemental Figure S9



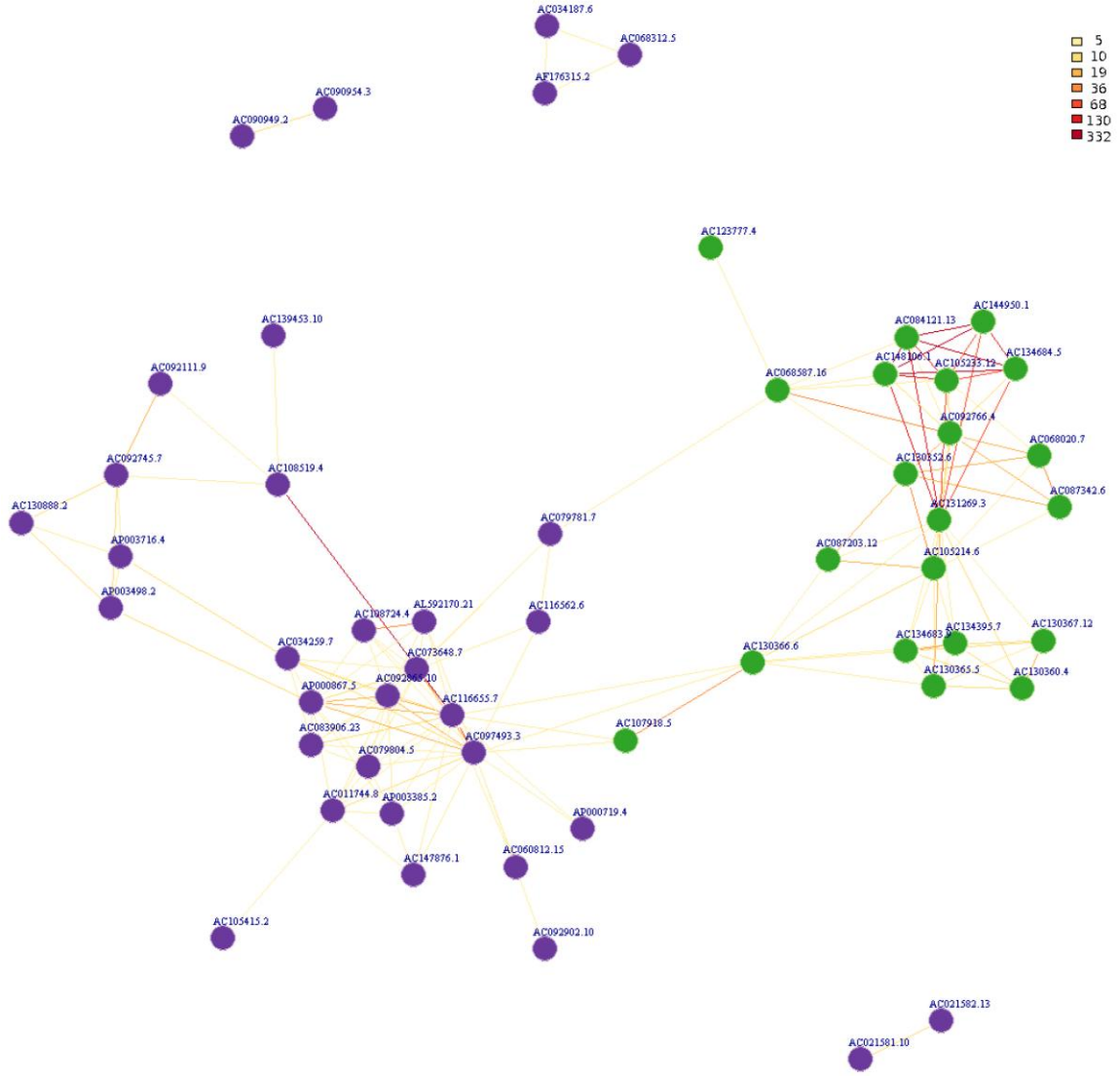
Supplemental Figure S10



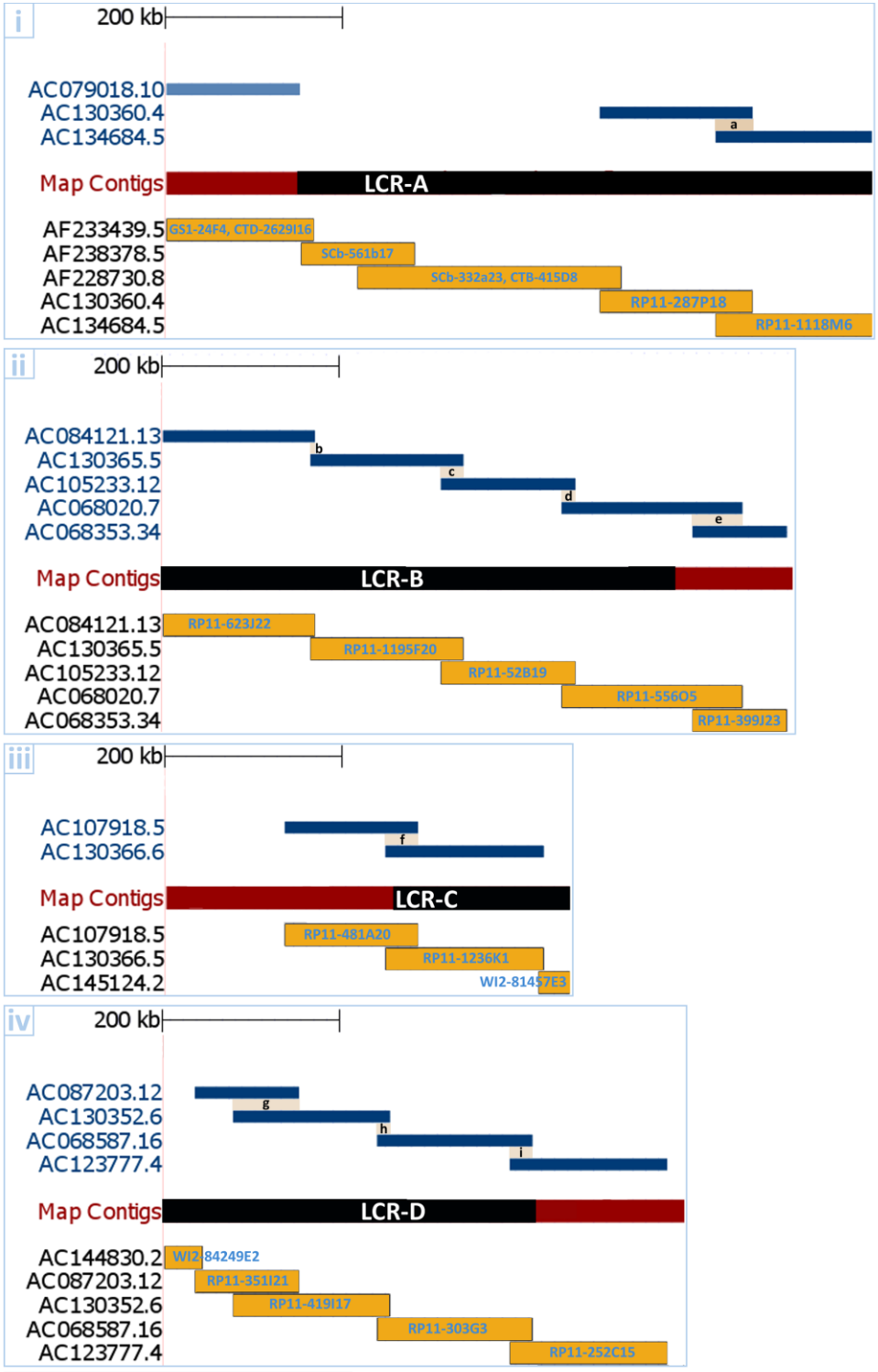
Supplemental Figure S11



Supplemental Figure S12



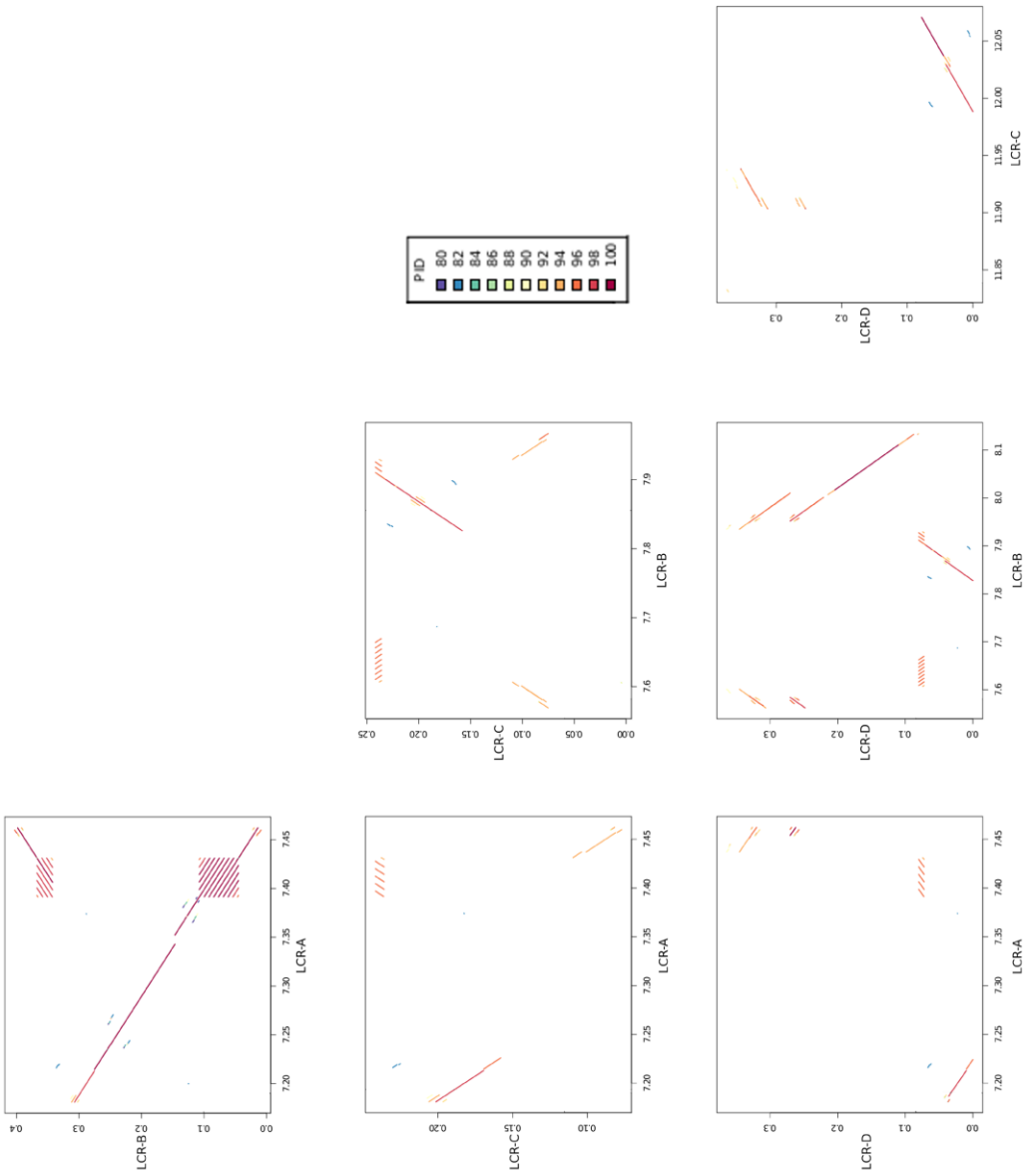
Supplemental Figure S13



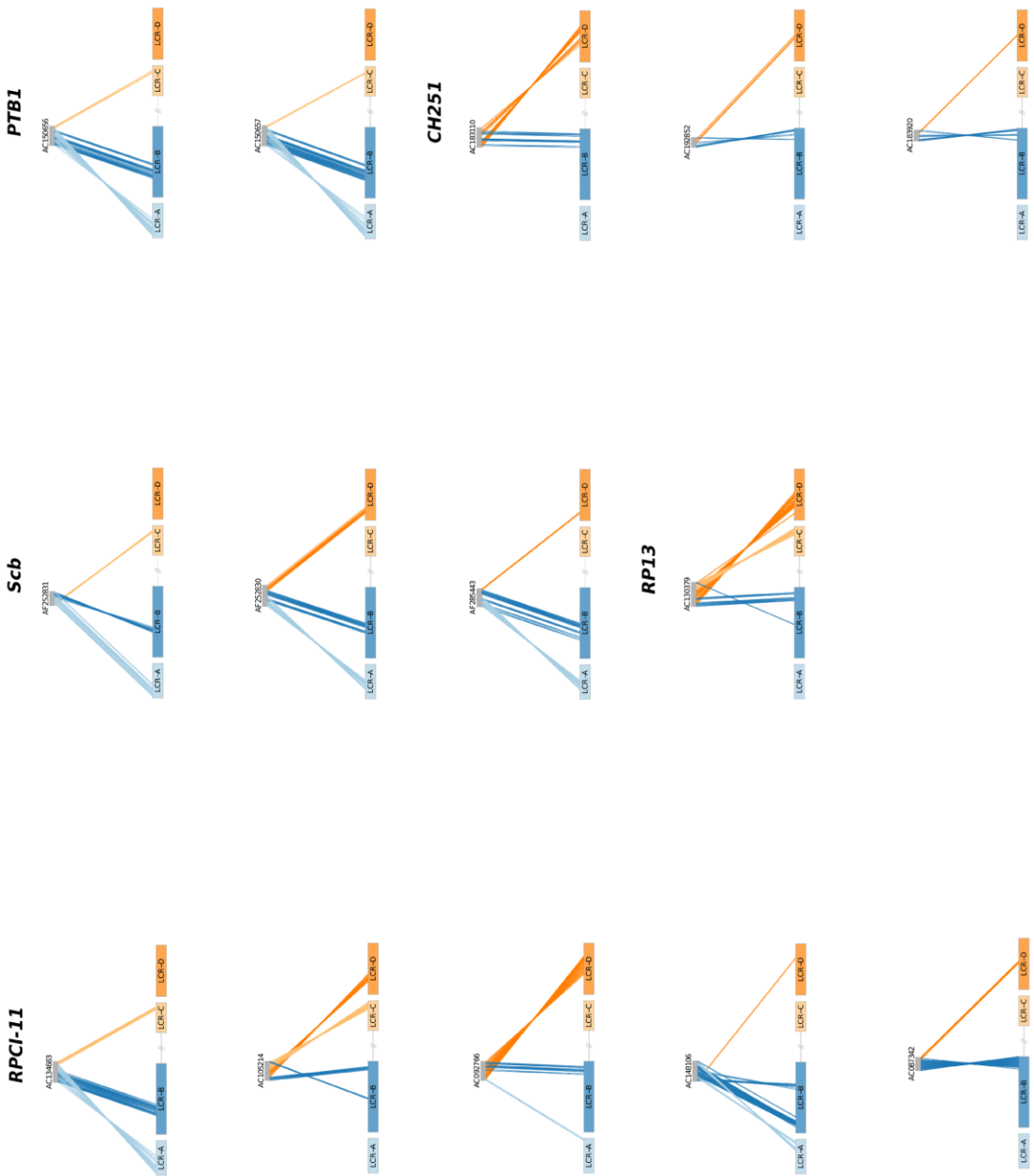
v

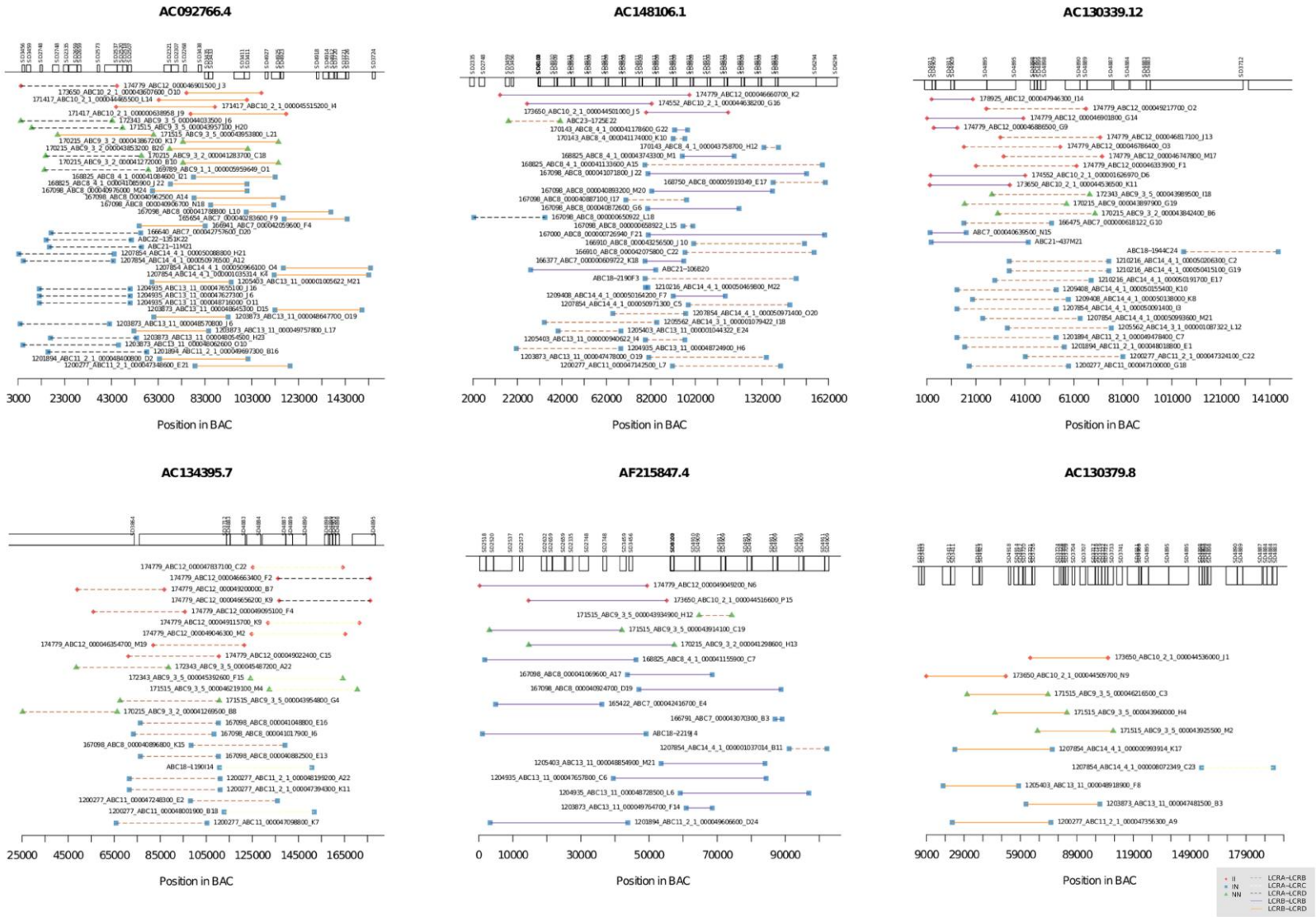
Overlap	SNP	Gap	Length	PID
a	0	0	40096	100
b	0	0	5700	100
c	0	0	25864	100
d	0	1	15948	99.9937
e	0	0	55752	100
f	0	0	36950	100
g	0	3	75317	99.996
h	0	0	15252	100
i	0	0	24720	100

Supplemental Figure S14



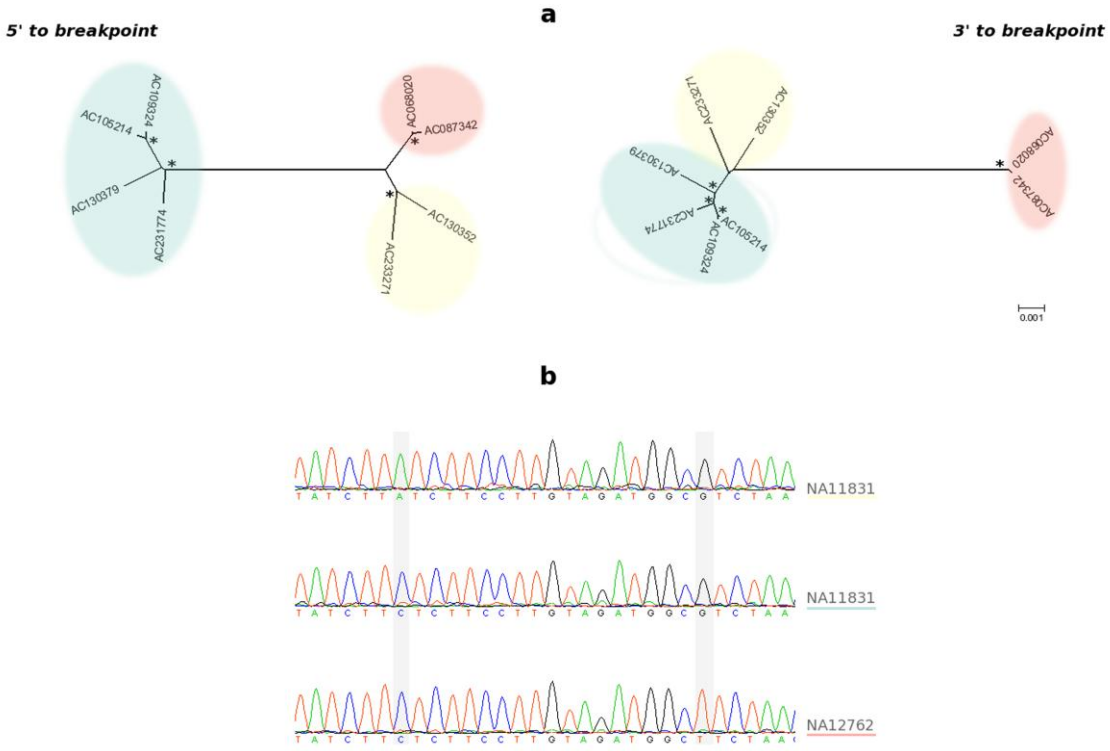
Supplemental Figure S15





Supplemental Figure S16

Supplemental Figure S17



Supplemental Figure S18