nature genetics

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41588-020-0646-x

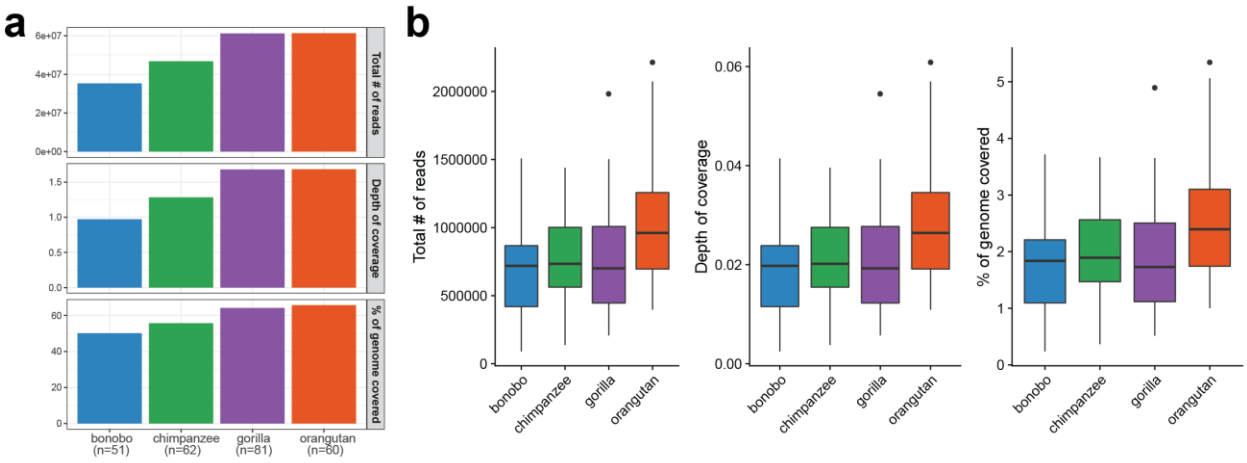
In the format provided by the authors and unedited.

Recurrent inversion toggling and great ape genome evolution

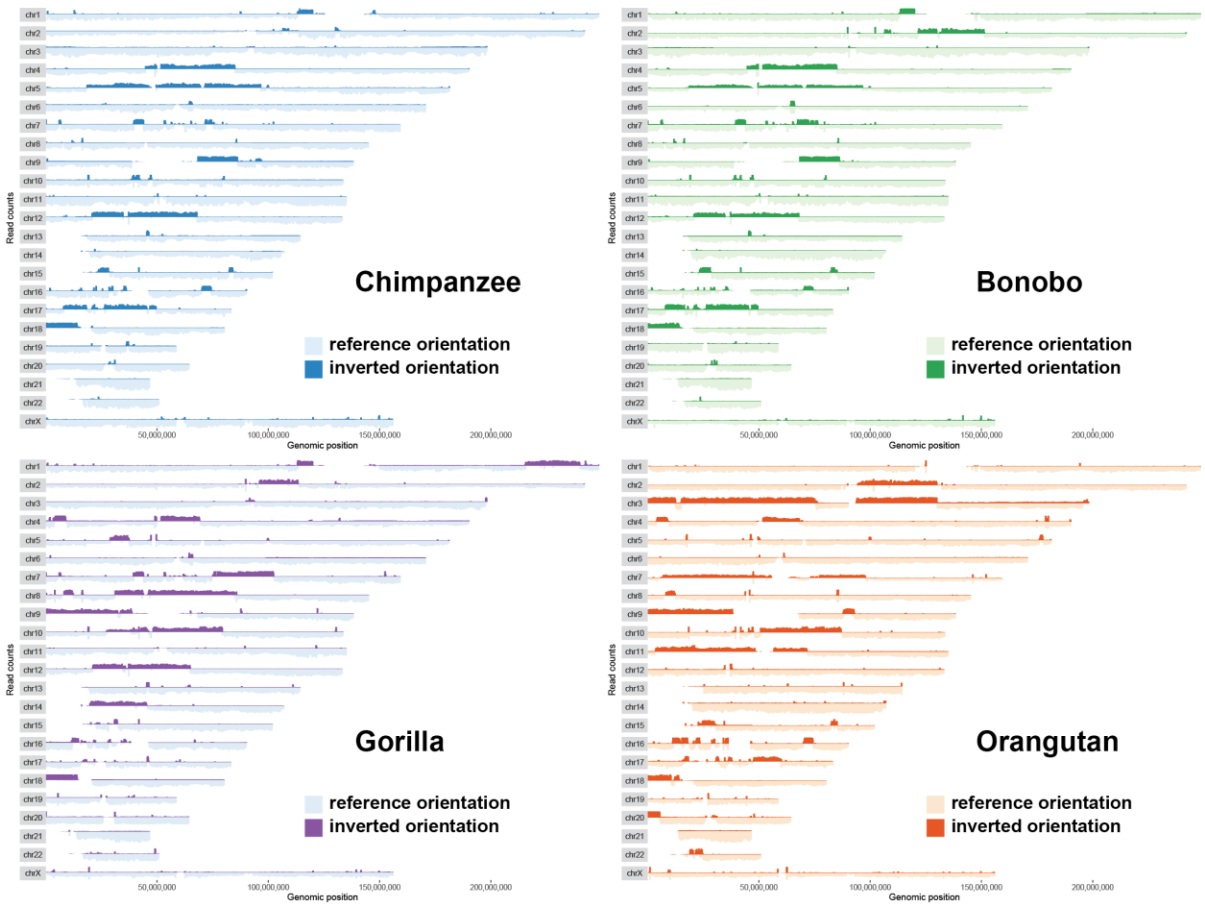
David Porubsky[1,2,9], Ashley D. Sanders[3,9], Wolfram Höps[3], PingHsun Hsieh[1], Arvis Sulovari[1], Ruiyang Li[1], Ludovica Mercuri[4], Melanie Sorensen[1], Shwetha C. Murali[1,5], David Gordon[1,5], Stuart Cantsilieris[1,6], Alex A. Pollen[7], Mario Ventura[4], Francesca Antonacci[4], Tobias Marschall[8], Jan O. Korbel[3] and Evan E. Eichler[1,5] ✉

[1]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA. [2]Max Planck Institute for Informatics, Saarland Informatics Campus, Saarbrücken, Germany. [3]European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany. [4]Dipartimento di Biologia, Università degli Studi di Bari Aldo Moro, Bari, Italy. [5]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA. [6]Centre for Eye Research Australia, Department of Surgery (Ophthalmology), University of Melbourne, Royal Victorian Eye and Ear Hospital, Melbourne, Victoria, Australia. [7]Department of Neurology, University of California, San Francisco, San Francisco, CA, USA. [8]Institute for Medical Biometry and Bioinformatics, Medical Faculty, Heinrich Heine University, Düsseldorf, Germany. [9]These authors contributed equally: David Porubsky, Ashley D. Sanders.
✉e-mail: eee@gs.washington.edu

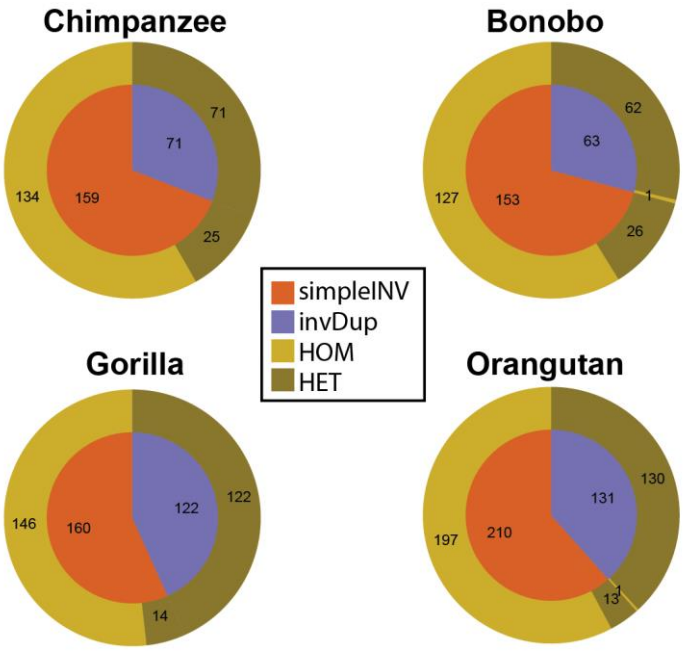
SUPPLEMENTARY FIGURES 1-32



Supplementary Figure 1: Summary statistics of Strand-seq libraries produced for NHPs.
a, Summary statistics for all selected Strand-seq libraries merged together. From the top to the bottom: Total number of reads (mapq≥10); depth of coverage given the length of GRCh38 without gaps and a percentage of GRCh38 positions covered by at least one Strand-seq read. **b**, Boxplot showing summary statistics from (a) as a distribution of values per Strand-seq library (bonobo n=51; chimpanzee n=62; gorilla n=81; orangutan n=60). Distribution of the total number of reads (median values for bonobo: 719,444; chimpanzee: 733,749; gorilla: 700,624; orangutan: 961,310), depth of coverage (median values for bonobo: 0.0198; chimpanzee: 0.0202; gorilla: 0.0193; orangutan: 0.0264) and % of genome covered (median values for bonobo: 1.84; chimpanzee: 1.89; gorilla: 1.73; orangutan: 2.39) is shown.

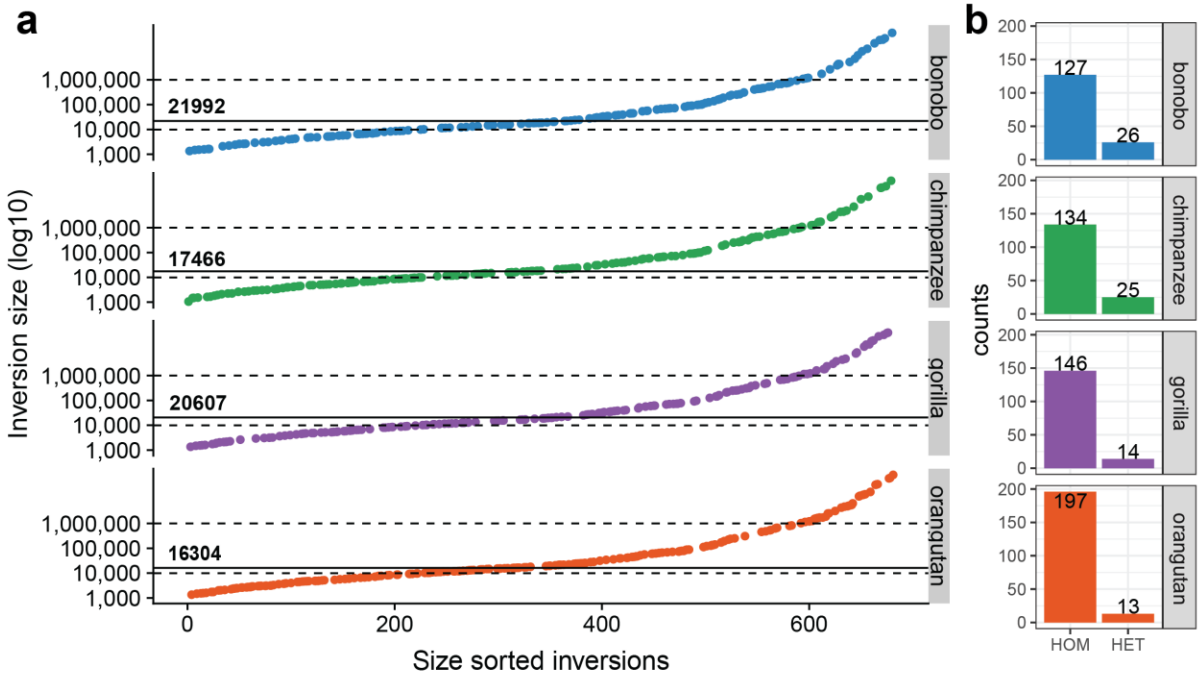


Supplementary Figure 2: Genome-wide plot of composite files constructed for each NHP.
Genome of each individual is binned into 200 kb bins and the number of reads mapped in forward (reference orientation - light color) and reverse (inverted orientation - dark color) orientation is depicted as a length of a bar along each chromosome.



Supplementary Figure 3: Summary of inversions detected in Strand-seq data.

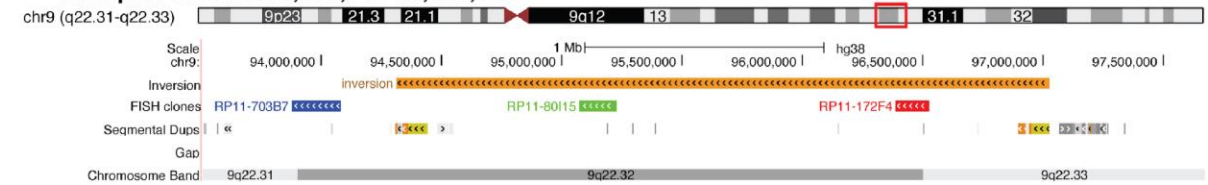
Summary of all inversions and inverted duplications mapped for each NHP. Inner circle projects a number of simple inversions ('INV') and inverted duplications ('invDup'). Outer circle projects genotype counts (homozygous - 'HOM', heterozygous - 'HET') within each inversion class (inner circle).

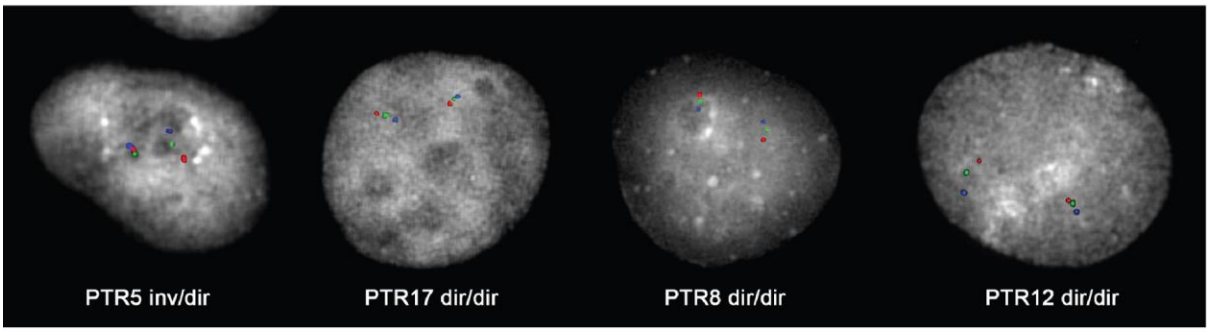


Supplementary Figure 4: Size distribution of simple inversions.

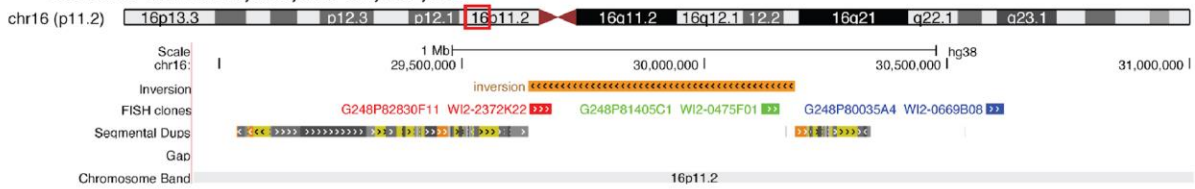
a, Size distribution of simple inversions for each NHP (rows) sorted by increasing size from the left to the right. Solid lines mark median inversion size for each NHP. **b**, Total counts of homozygous (HOM) and heterozygous (HET) inversions per NHP.

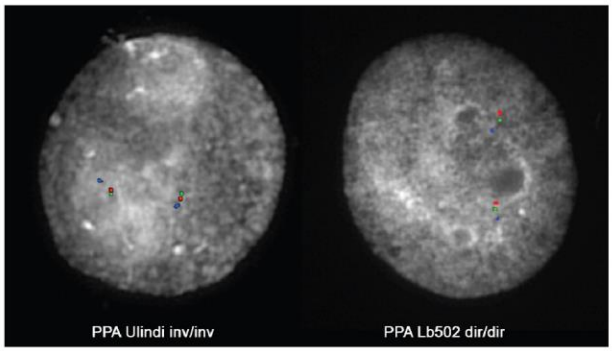
a **Chimpanzee chr9:94,308,147-97,028,185**



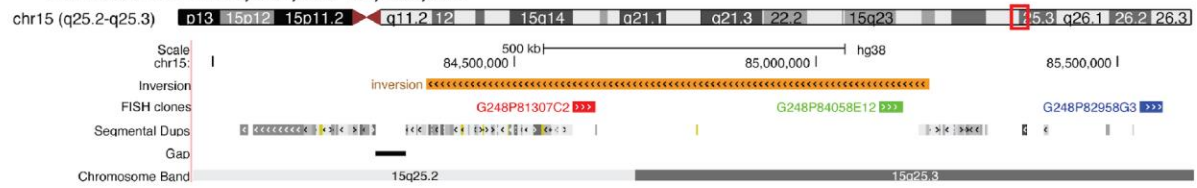


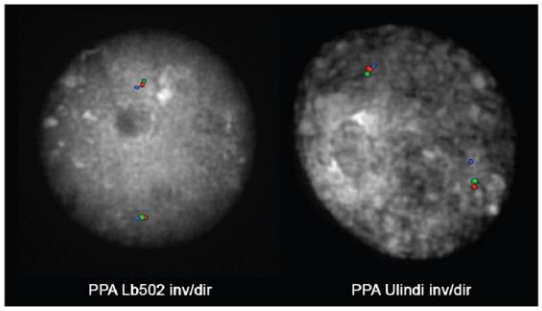
b **Bonobo chr16:29,638,605-30,187,613**

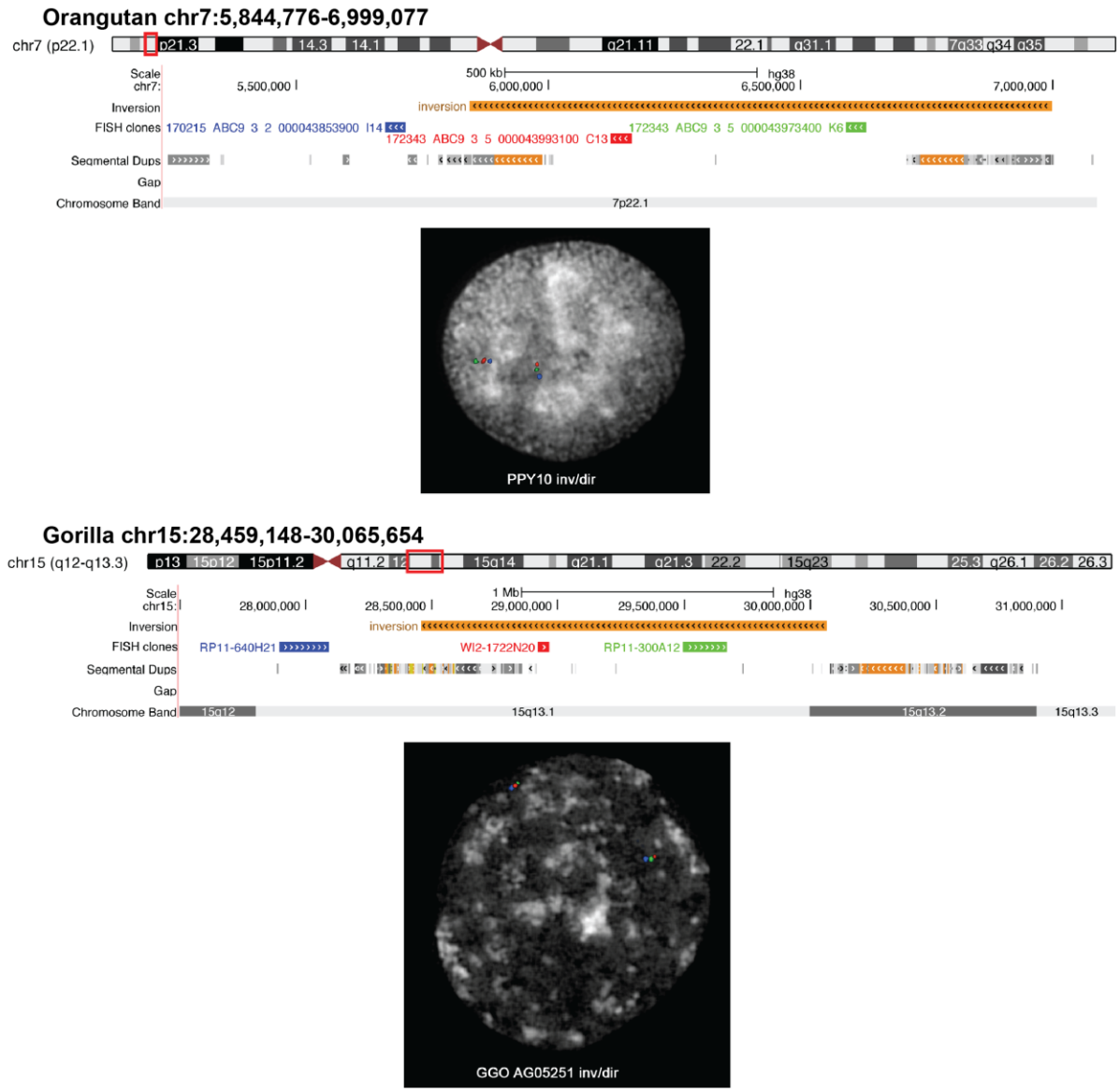




Bonobo chr15:84,355,083-85,188,393

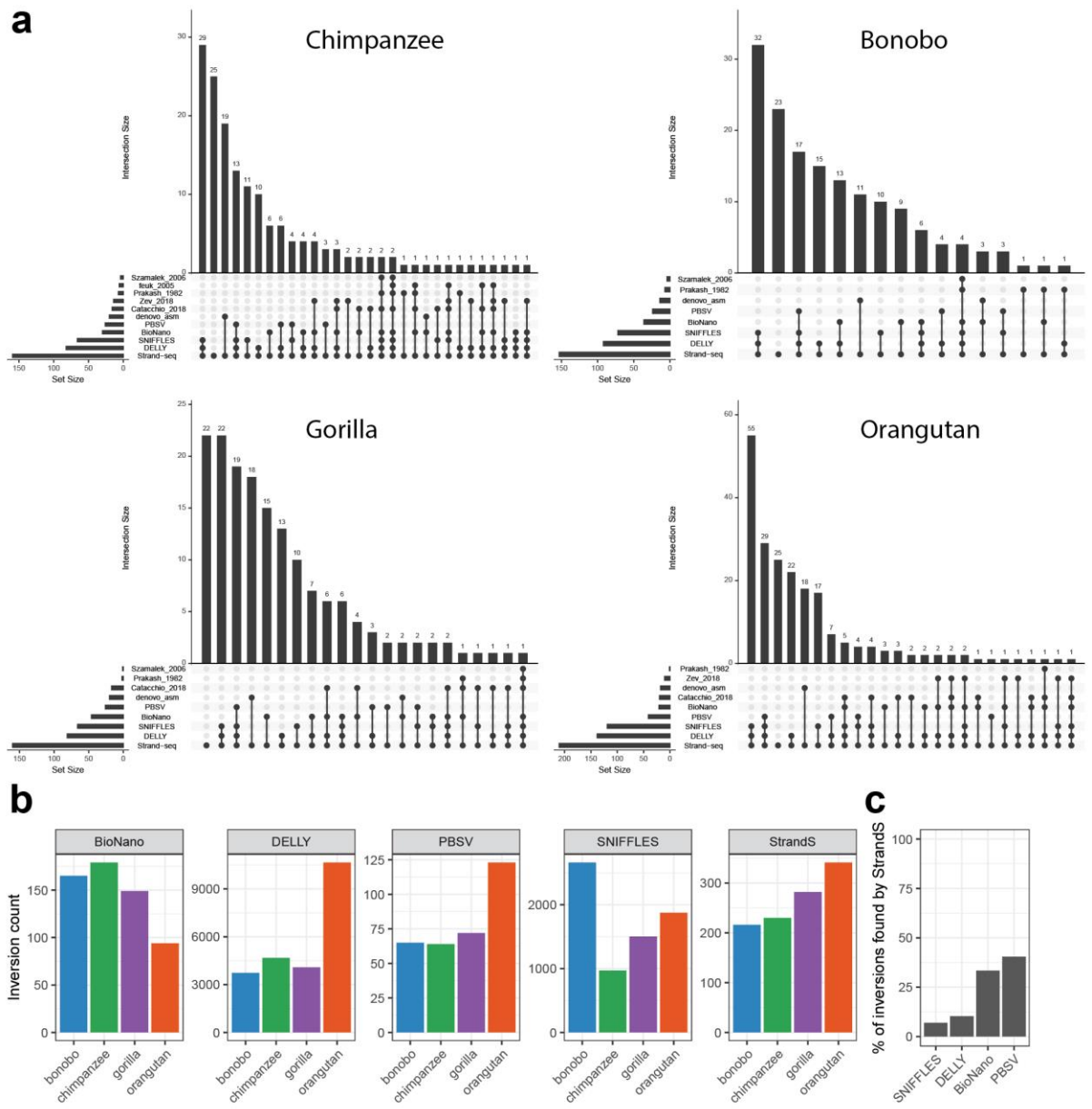






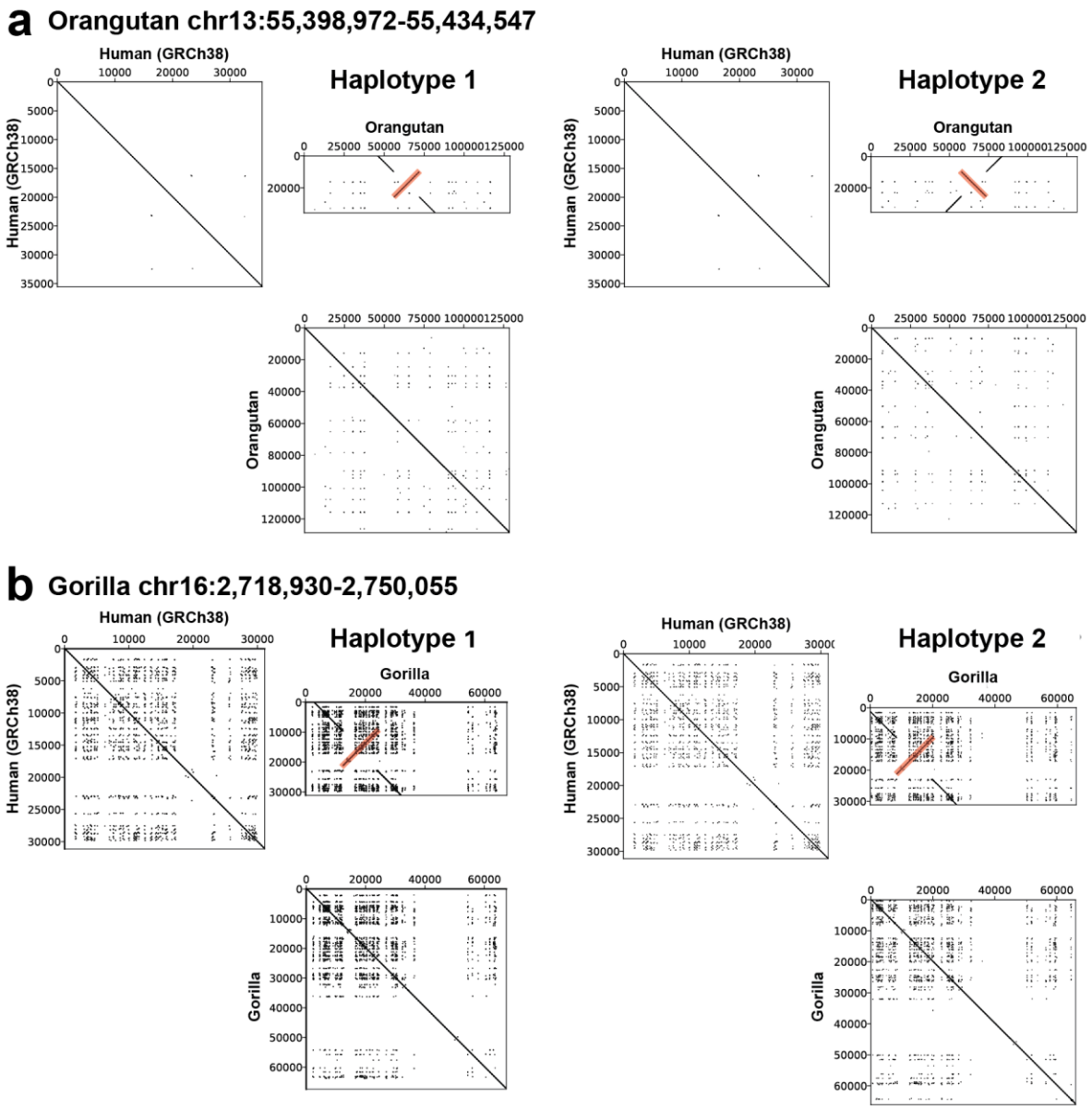
Supplementary Figure 5: Experimental validation by FISH.

a, The top panel shows a UCSC Genome Browser view of a 2.7 Mb inversion at chr9:94308147-97028185 predicted by Strand-seq in a heterozygous state in chimpanzee. FISH clones are shown with colored boxes. The bottom panel shows FISH results on four chimpanzee individuals, with three (PTR12, PTR17, PTR8) being homozygous for the direct haplotype and one (PTR5) carrying the inversion in heterozygous state. **b**, Other FISH validated inversions for bonobo, orangutan, and gorilla (**Supplementary Table 2**). High-resolution FISH images are available at https://github.com/daewoooo/ApeInversion_paper/FISHimages. A minimum of 50 interphase cells were scored for each region in order to determine if the pattern observed was casual or due to a real inversion. A region was considered homozygously inverted if scoring of the probes in inverted orientation exceeded 80% of the total count, and heterozygously inverted if probes in direct and inverted orientation were equally scored.



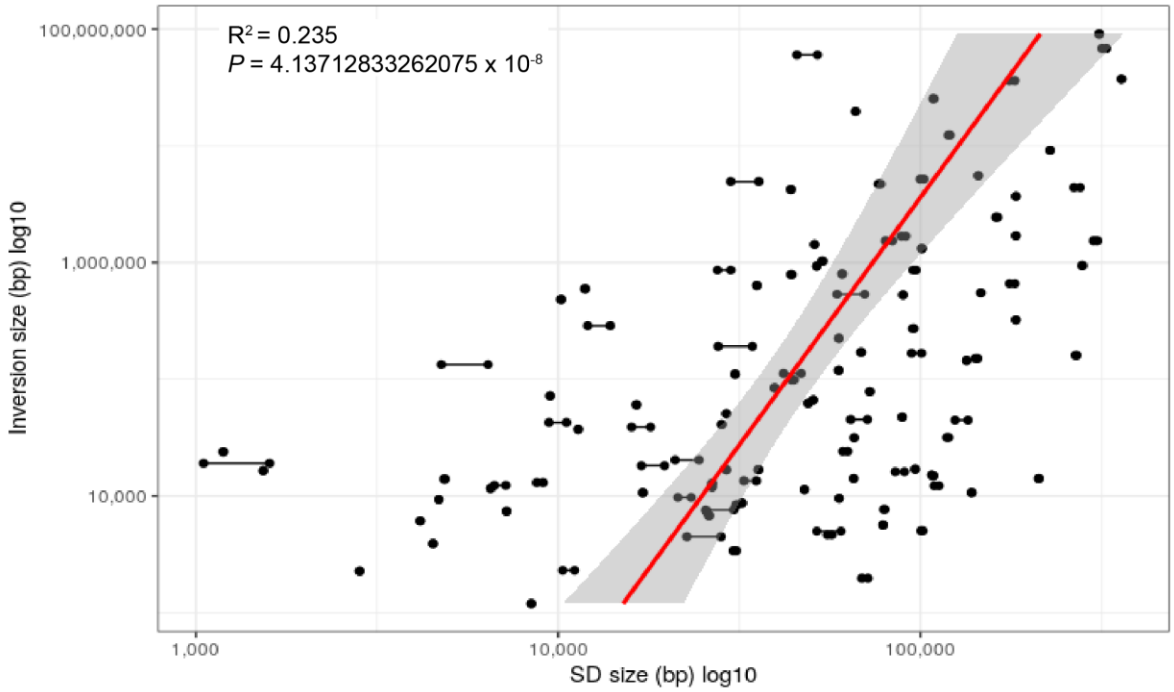
Supplementary Figure 6: Inversion validation summary.

a, UpsetR[1] plots (per NHP) showing the number of Strand-seq detected inversions that have been found by other SV callers (e.g., PBSV and SNIFFLES) or were already published. **b**, Total number inversion calls per NHP for Illumina (DELLY), Bionano, PacBio (PBSV, SNIFFLES) and Strand-seq (StrandS) data. **c**, Percentage of inversion calls from other technologies that have 50% reciprocal overlap with Strand-seq callset.



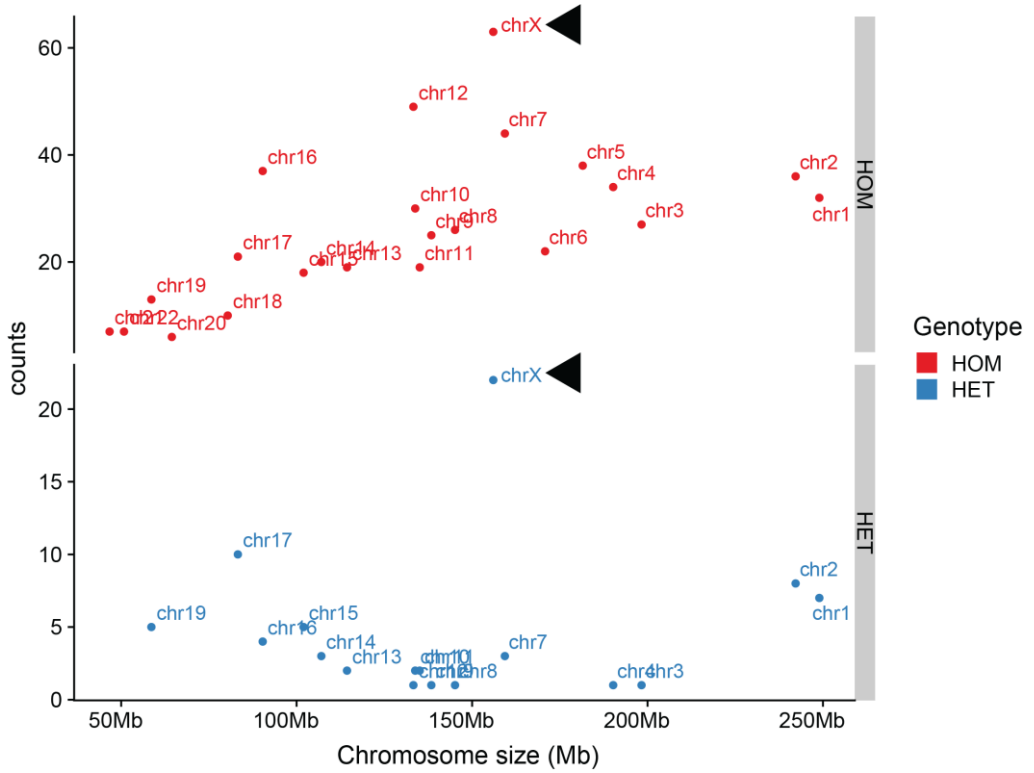
Supplementary Figure 7: Inversion validation by phased PacBio assembly.

Pairwise comparison dotplots of NHP phased contigs against human reference (GRCh38). In both regions, self-comparisons of human reference or NHP assembled contigs show no large duplication event or palindromic sequence while comparisons between human reference and assembled contigs indicate an inversion event. **a**, Orangutan region chr13:55398972-55434547 shows homozygous inversion events. **b**, Gorilla region chr16:2718930-2750055 shows homozygous inversions compared to human. Inverted regions are highlighted by a red rectangle.



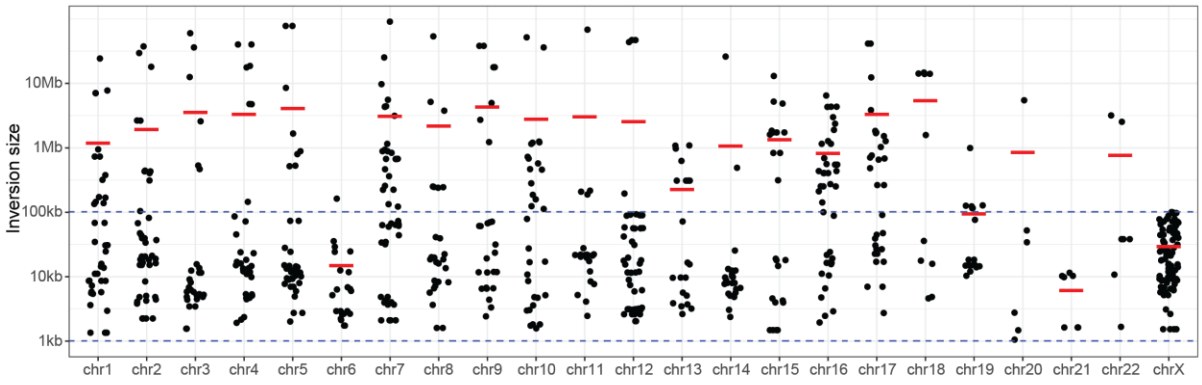
Supplementary Figure 8: Correlation between inversion size and flanking SD size.

Here we calculated the distance of each inversion breakpoint to the closest SD using the primatR function called 'range2rangeDistance'. Inversions flanked by SDs have been defined as those that overlap SD regions or the SD region is within 5 kb of both breakpoints. Based on this, we defined 115 of 388 nonredundant simple inversions as flanked by SDs. Then, we selected all intrachromosomal SD pairs within 1 kb distance from either inversion breakpoint. We retained the longest SD on either side of the inversions as the flanking SD. Lastly, we calculated the correlation between inversion size and the mean SD size flanking the inversion by fitting a linear model of inversion.size ~ mean.SD.size using raw lengths of both variables by R function 'lm'. The scatterplot shows the relationship between inversion size (n=115) and corresponding SDs flanking the inversion. The red line shows the correlation between inversion size and mean SD size flanking any given inversion along with a 95% CI shown in the gray area around the line. GRCh38-specific SD annotation for this analysis was obtained from UCSC Genome Browser[2].



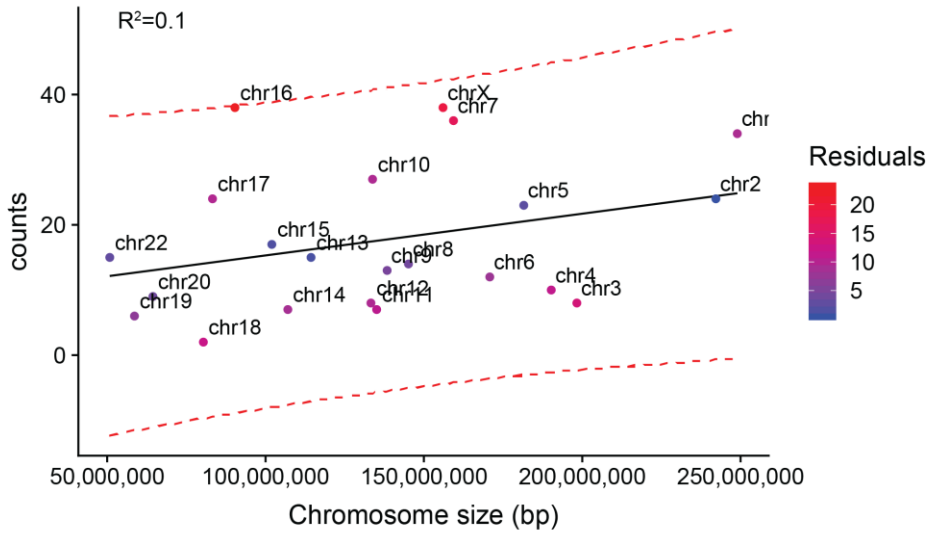
Supplementary Figure 9: Number of simple inversions for each chromosome given the chromosome length and genotype (HOM vs. HET).

Chromosome X has the highest count of HOM and HET inversions given the chromosome size (black arrowheads).



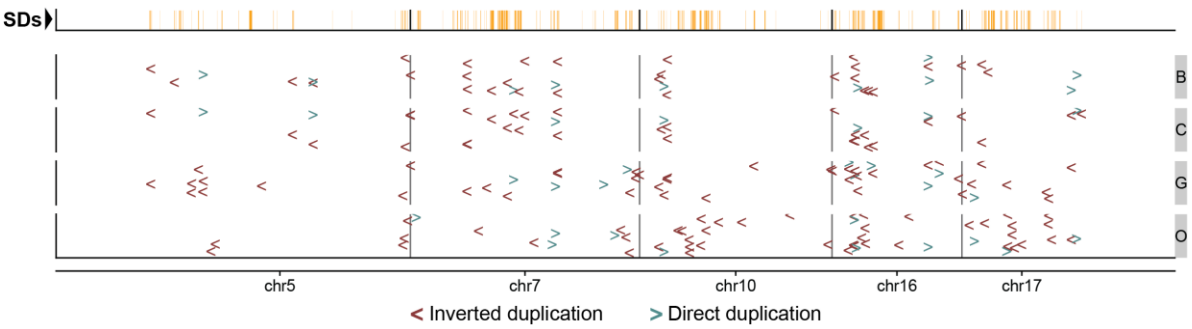
Supplementary Figure 10: Simple inversion size distribution per chromosome.

Each dot represents an inversion with size shown on the y-axis (log10 scale). Red lines indicate mean inversion size for each chromosome. Blue dashed lines highlight that the inversion size distribution for X chromosome is bounded between 1 kb and 100 kb (NOTE: 1 kb is the lower bound of inversion detection in this study).



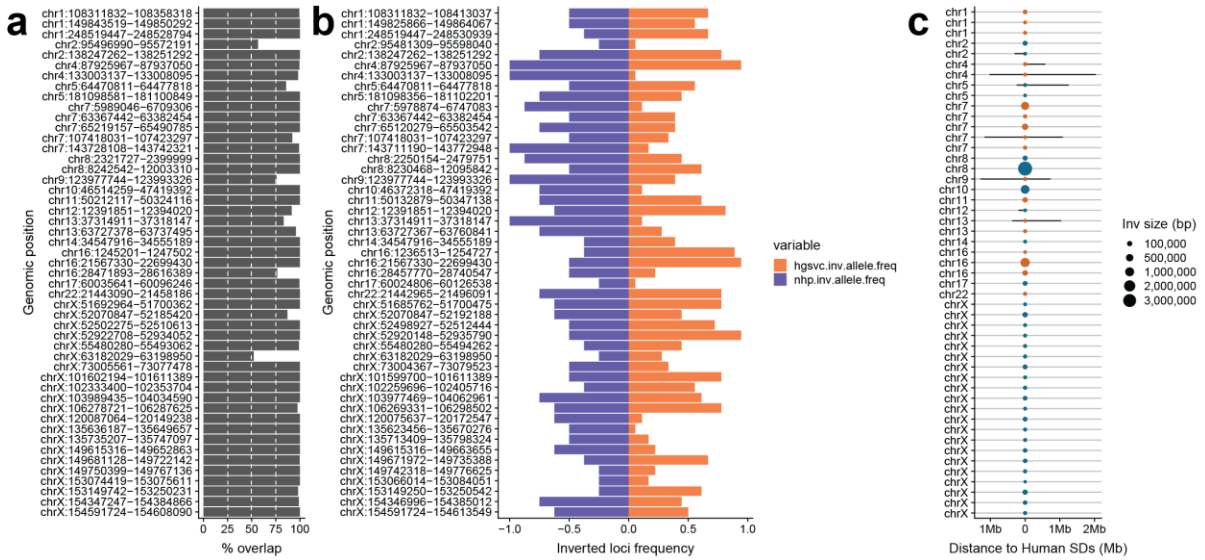
Supplementary Figure 11: Inverted duplication counts per chromosome size.

The scatterplot shows the number of inverted duplications (n=387) given the chromosome length. The regression line is added as a solid black line and 95% confidence intervals are highlighted as red dashed lines. Deviation from an expected number of inverted duplications is enumerated in the number of residuals.



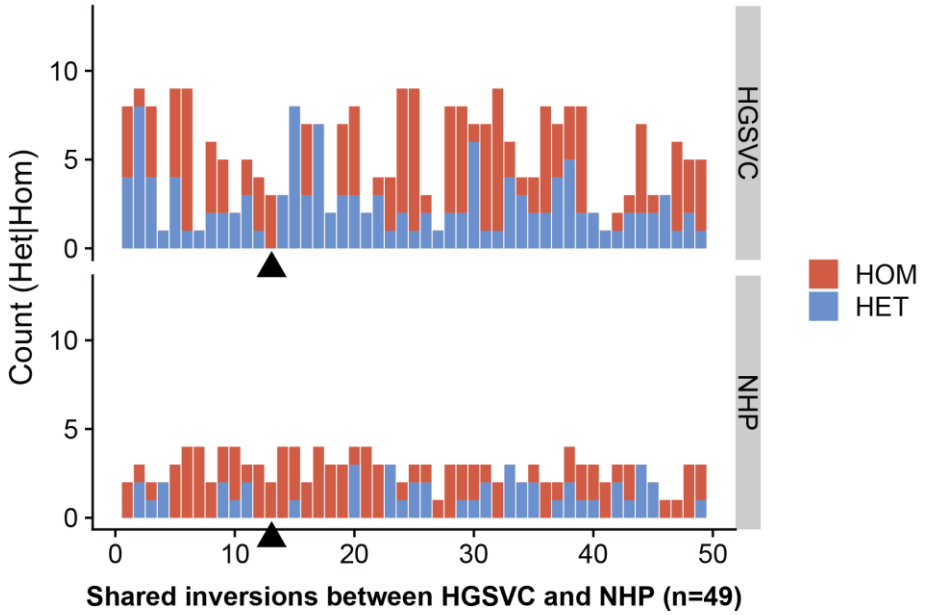
Supplementary Figure 12: Distribution of inverted duplications for selected chromosomes.

Distribution of inverted ('<' - dark red) and direct ('>' - blue) duplications along chromosomes 5, 7, 10, 16 and 17. The upper panel plots the distribution of known human SDs (orange). Each NHP (B - bonobo, C - chimpanzee, G - gorilla, O - orangutan) is plotted in a single row.



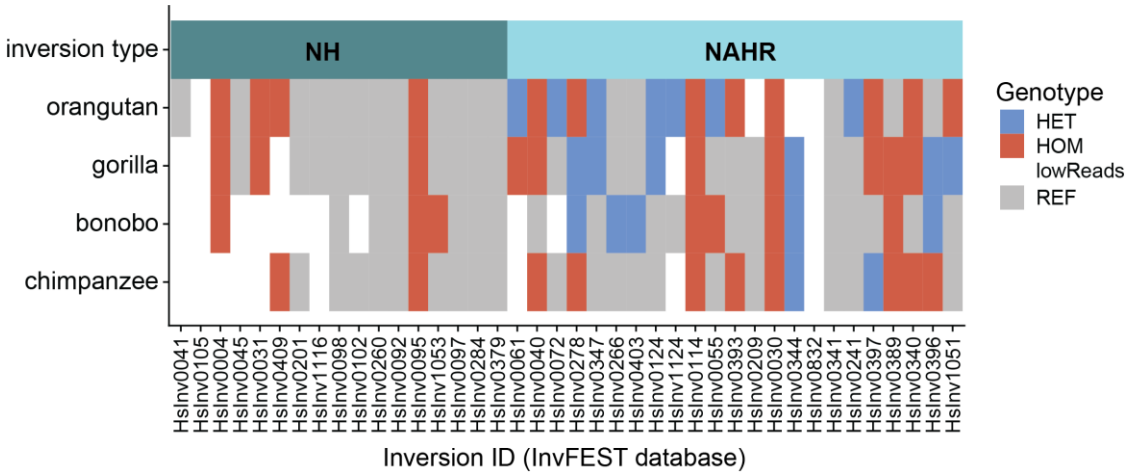
Supplementary Figure 13: Summary of shared inversions between HGSVC and NHP datasets.

a, Percentage of overlap between the HGSVC and NHP datasets for 49 regions that pass the 50% reciprocal overlap threshold. **b**, Inverted loci frequency of overlapping inversions (n=49) between HGSVC (orange) and NHP (purple) from (a). **c**, The distance for each inversion from the closest human SD is shown as a horizontal black bar reaching to the left and right from the centrally positioned dot. The size of each dot represents the size of a given inversion. Alternating orange and blue colors distinguish different chromosomes.



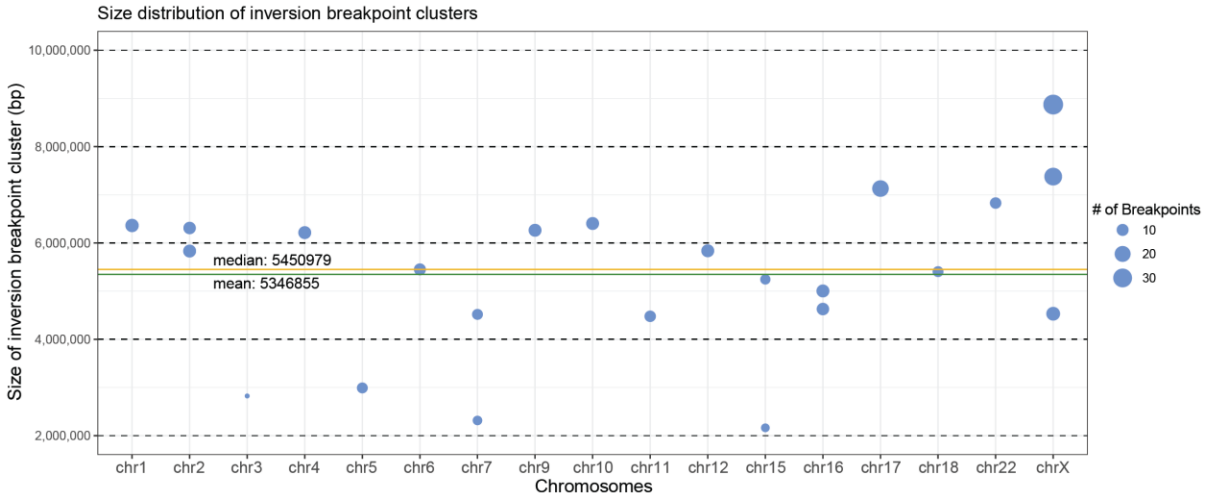
Supplementary Figure 14: Genotype counts for shared inversions between NHP and HGSVC inversions.

Barplot showing the total counts of homozygous (red - HOM) and heterozygous (blue - HET) inversions for each of the 49 sites shared between HGSVC and NHP individuals. Each count represents a sum of genotypes among HGSVC and NHP individuals. The black arrowhead points to an inversion that appears to be homozygous in both HGSVC and NHP individuals and is likely a false positive call.



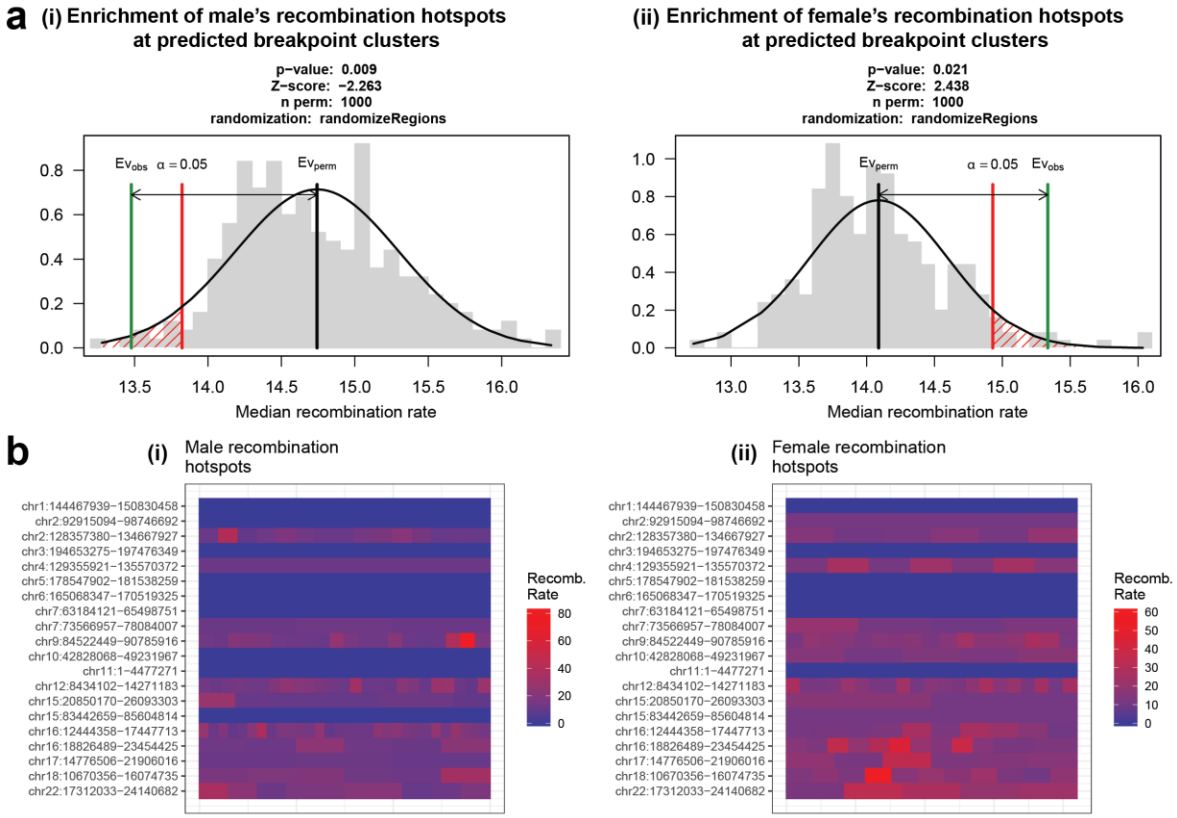
Supplementary Figure 15: Strand-seq genotype for a selected number of inversions from Giner-Delgado et al. (2019).

From the total of 44 inversions genotyped in multiple humans and NHPs[3], we selected inversions longer than 1 kb (n=40) and re-genotyped them based on Strand-seq data. All inversions were divided into two categories: those initiated by non-homologous (NH) mechanism and those initiated by non-allelic homologous recombination (NAHR) based on the data from the InvFEST database. Inversions are ordered in columns and each row represents an NHP. Genotypes are colored as follows: blue - heterozygous (HET), red - homozygous (HOM), and gray - reference orientation (REF) or no-call because of low read coverage. In line with Giner-Delgado et al. (2019), we observe the highest number of heterozygous inversions among NAHR initiated inversions.



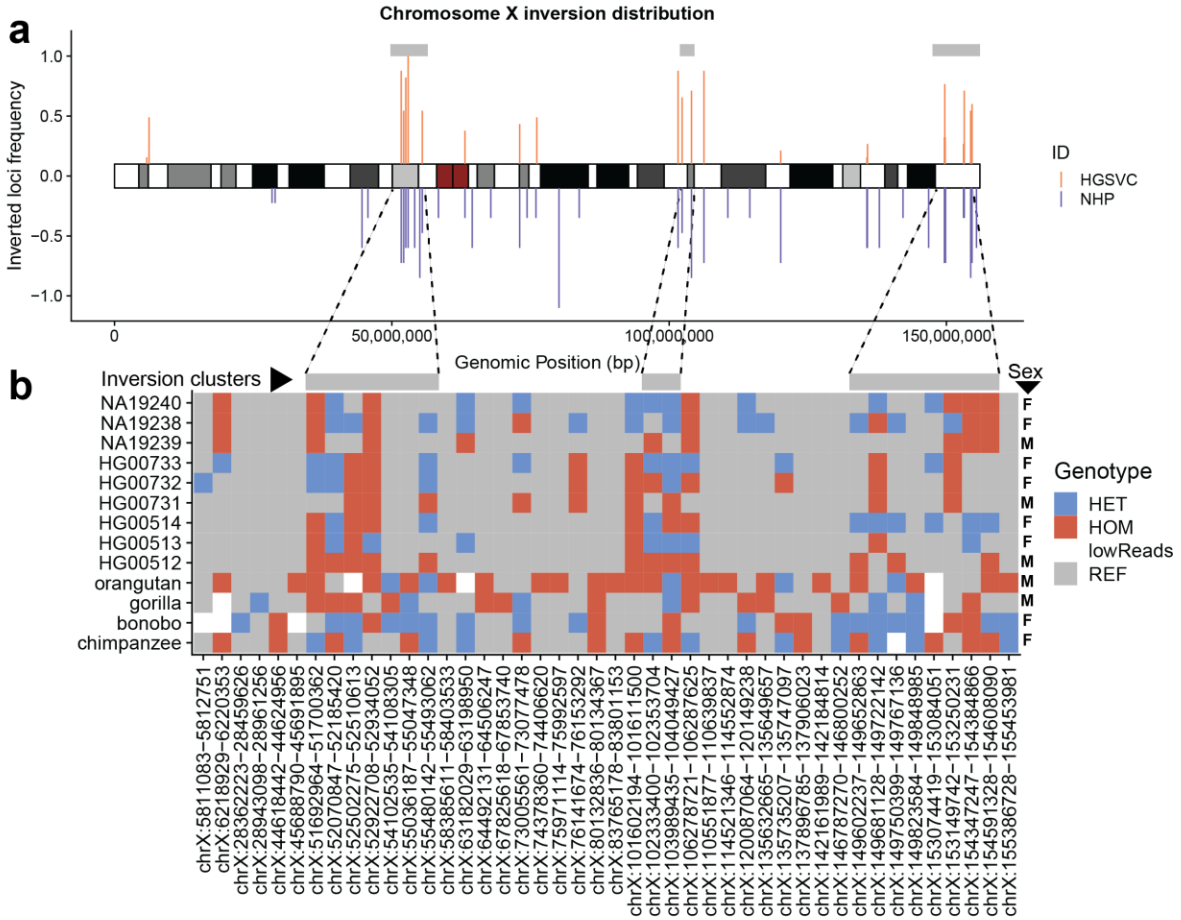
Supplementary Figure 16: Size distribution of detected inversion breakpoint clusters.

The size distribution of each inversion breakpoint cluster is plotted as a dot per chromosome. The size of the dot is scaled to the number of inversion breakpoints within each cluster. Mean (green) and median (yellow) sizes are shown as solid horizontal lines.



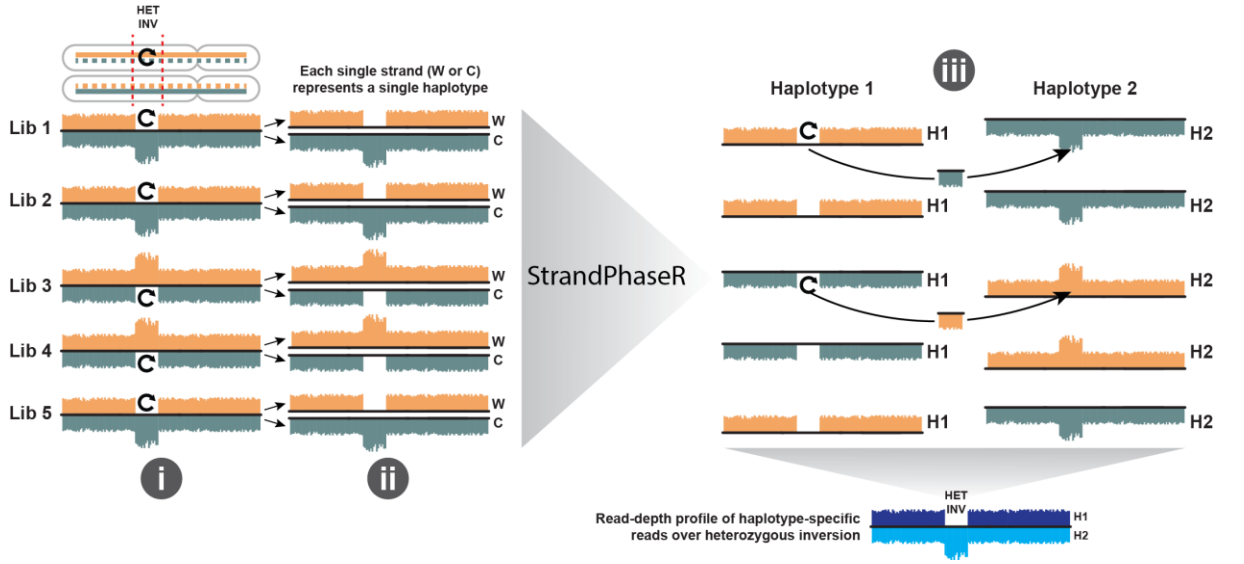
Supplementary Figure 17: Male and female recombination rates in predicted inversion breakpoint clusters.

a, Enrichment analysis of male and female recombination hotspots at predicted inversion breakpoint clusters. Green line shows observed median recombination rate over all breakpoint clusters while barplot shows permuted values after 1000 iterations performed using regioneR[4] function 'permTest'. **b**, A comparison of the overlap of the inversion breakpoint clusters (rows) predicted in this study to the male's and female's meiotic recombination hotspots from the deCODE project[5]. The size of each breakpoint cluster is scaled to the largest breakpoint cluster. Recombination hotspots are scaled accordingly. Red and blue shades depict recombination hot and cold regions, respectively.



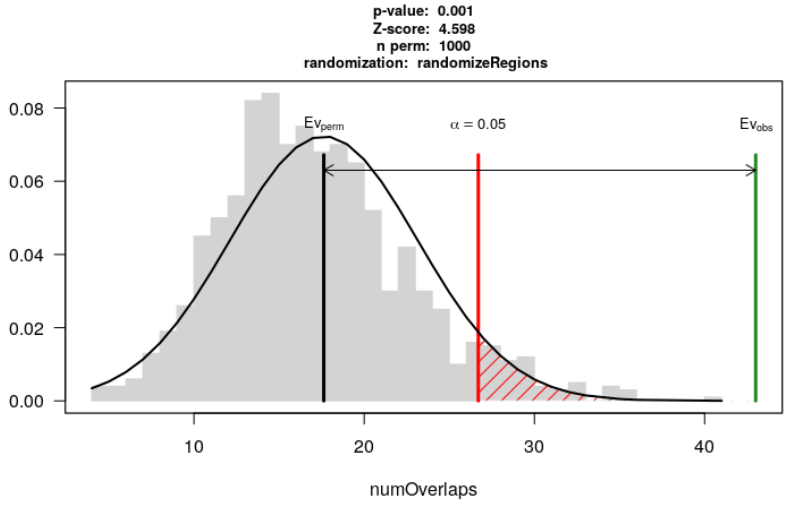
Supplementary Figure 18: Distribution of simple inversions on chromosome X and their genotypes.

a, Distribution of inverted allele frequencies for HGSVC (orange) and NHP (purple) along chromosome X. Gray rectangles at the top of the plots show positions of predicted inversion breakpoint clusters (from Fig. 3C). **b**, Heatmap of inversion genotypes for each inversion detected on chromosome X (columns) and for each HGSVC and NHP individual (rows). Heterozygous (HET) inversions are colored in blue and homozygous (HOM) inversions are colored in red. Reference orientation (REF) is colored in gray.



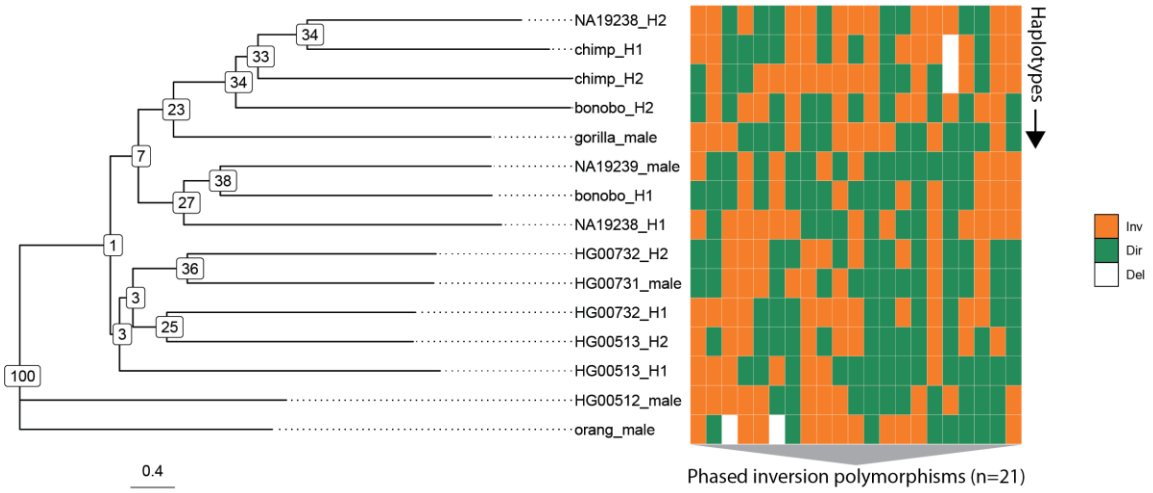
Supplementary Figure 19: Phasing of heterozygous inversions based on the depth of phased Strand-seq reads.

(i) An expected read distribution of a heterozygous inversion (red dashed lines) is shown for five example Strand-seq libraries with a random inheritance of maternal and paternal template strands. As was previously reported[6], genomic regions that inherited one Watson (W) and one Crick (C) template strand from each parent are haplotype informative. (ii) Essentially, W and C reads are inherited from a single parent. This allows us to physically separate heterozygous single-nucleotide variants (SNVs) sampled in W and C reads along each parental homologue. Using the strand directionality, larger structural variants, such as heterozygous inversions, can be anchored in their respective haplotype using these SNVs. (iii) We used StrandPhaseR[7] to assign template strands (W or C) from each Strand-seq library to their respective haplotypes (Haplotype 1 and 2). Inversion changes directionality of a piece of DNA and therefore reads overlapping an inverted region have flipped directionality and thus are assigned to the opposite strand (haplotype) (see arrows). This results in absence or very low coverage over an inverted allele (see Haplotype 1) while the coverage over a reference allele (Haplotype 2) is increased.



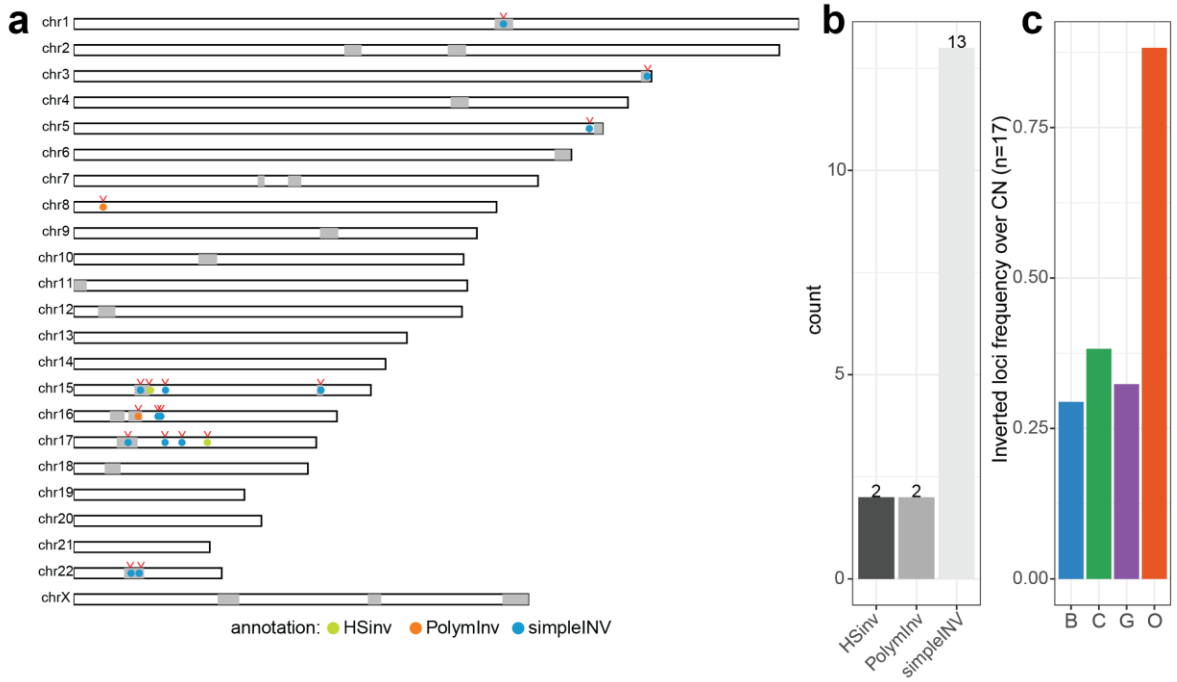
Supplementary Figure 20: Enrichment of protein-coding genes mapping to polymorphic X chromosome inversions.

Here each chromosome X inversion (n=21) range was first expanded by half of its original size to the left and to the right. The green vertical line shows the number of protein-coding genes (n=43) overlapping observed chromosome X inversions (n=21) while the red line marks the significance level. Gray bars show distribution of overlapping protein-coding genes after 1000 random shuffling of the original set of inversions (after length expansion). We used the R package regioneR[4] and function 'permTest' to perform the randomization experiment. The list of protein-coding genes in GRCh38 was obtained from the GENCODE database (v29).



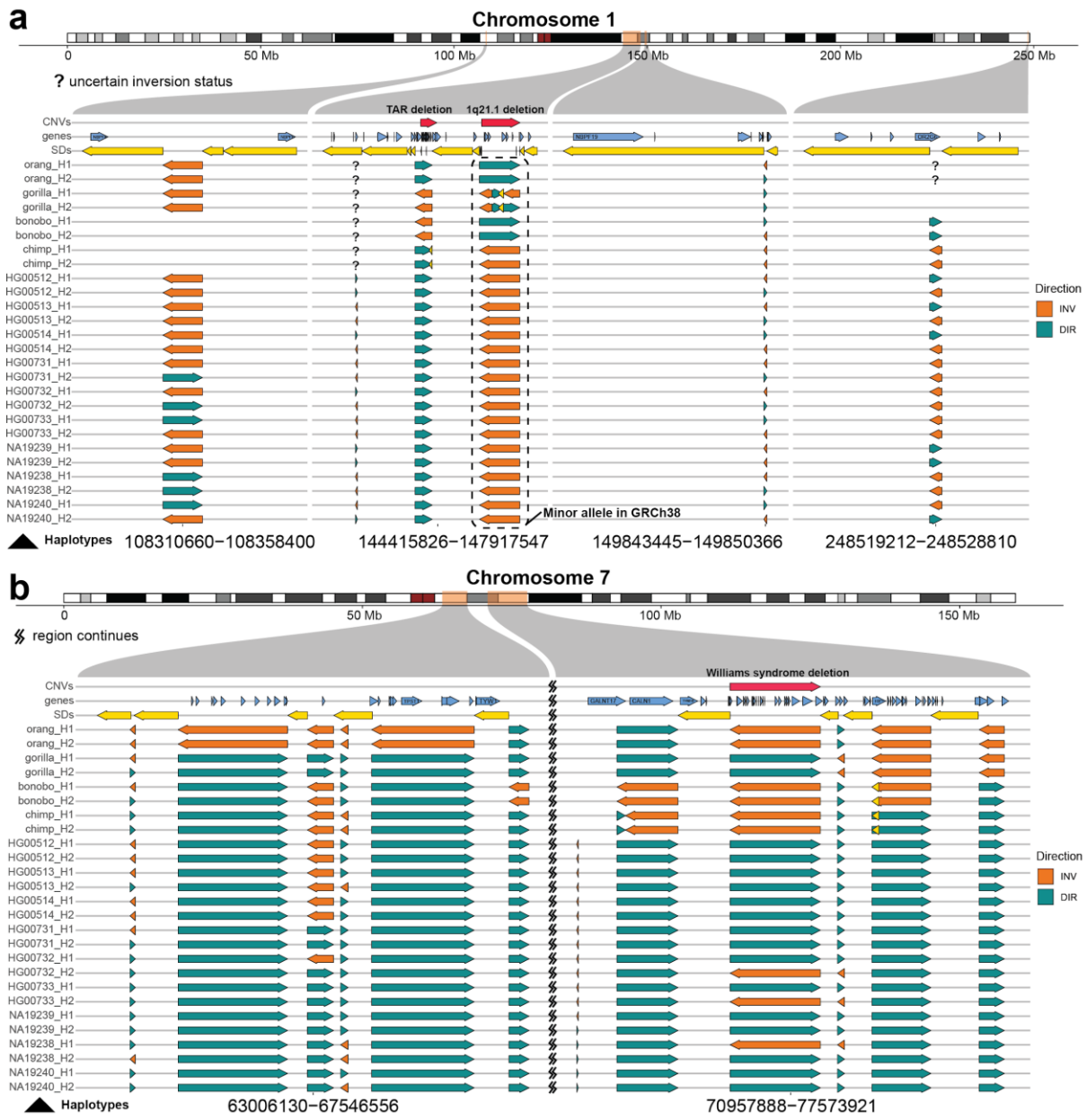
Supplementary Figure 21: Similarity of phased inversion polymorphisms on chromosome X.

Left plot: A neighbor-joining evolutionary tree based on Euclidean distances between unique haplotypes (n=15) in the great ape lineage. Bootstrap support is plotted over each node (10,000 iterations). Right plot: Inversion state (orange - inverted, green - direct/reference, white - deleted) at each polymorphic site (n=21).



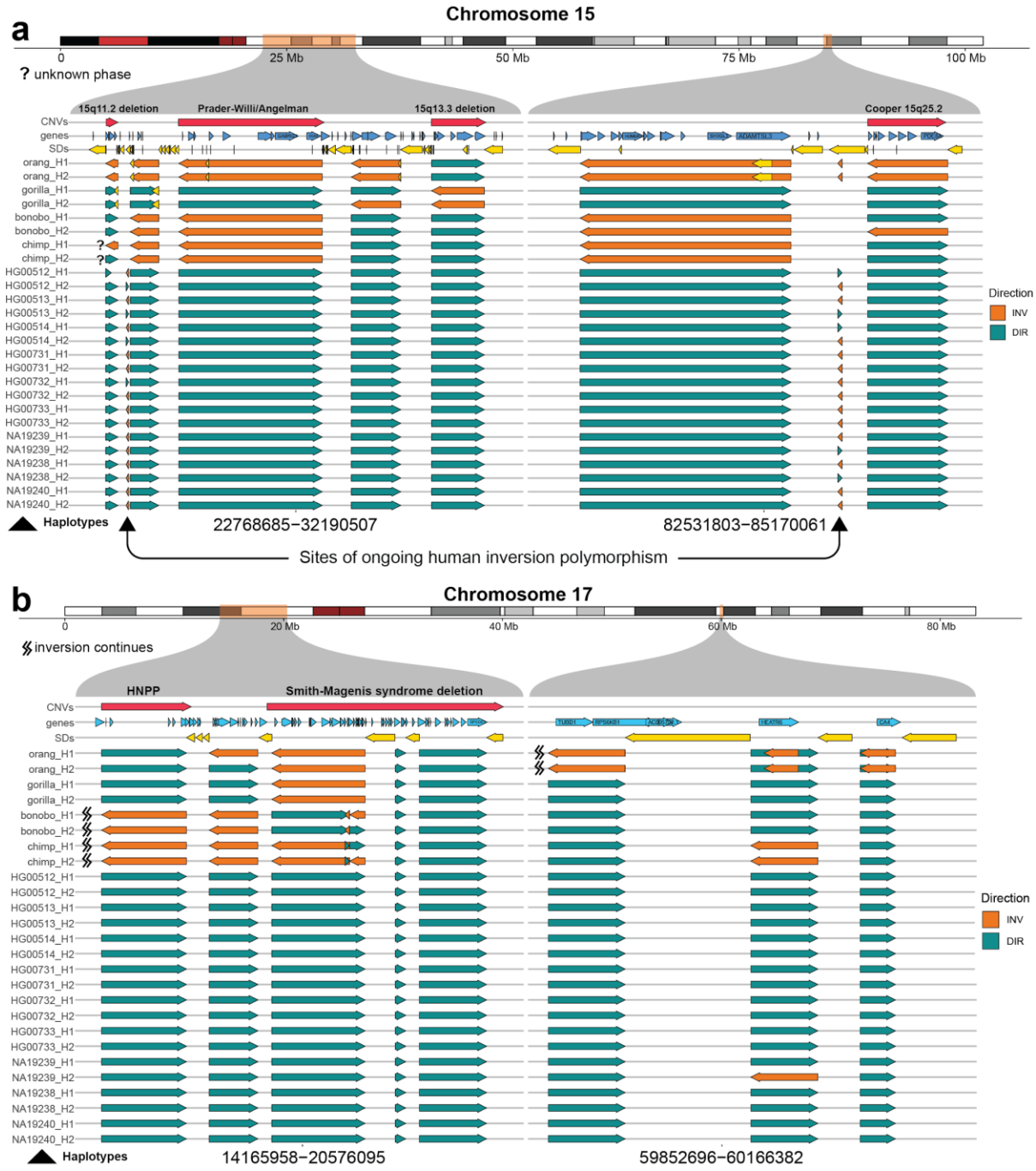
Supplementary Figure 22: Overlap of NHP inversions with CNV morbidity map.

a, Overlap (50% reciprocal) of NHP simple inversion calls with the gold-standard CNV morbidity map predicted to be involved with neurodevelopmental diseases[8]. Colored dots distinguish various classes of simple inversions detected in our NHP callset. Red arrowheads point to locations of a respective CNV. **b**, Barplot showing the total number of different classes of simple inversions that overlap with the CNV morbidity map. **c**, Frequency of inverted loci over the CNVs that have been shown to overlap with NHPs (n=17) in (a).



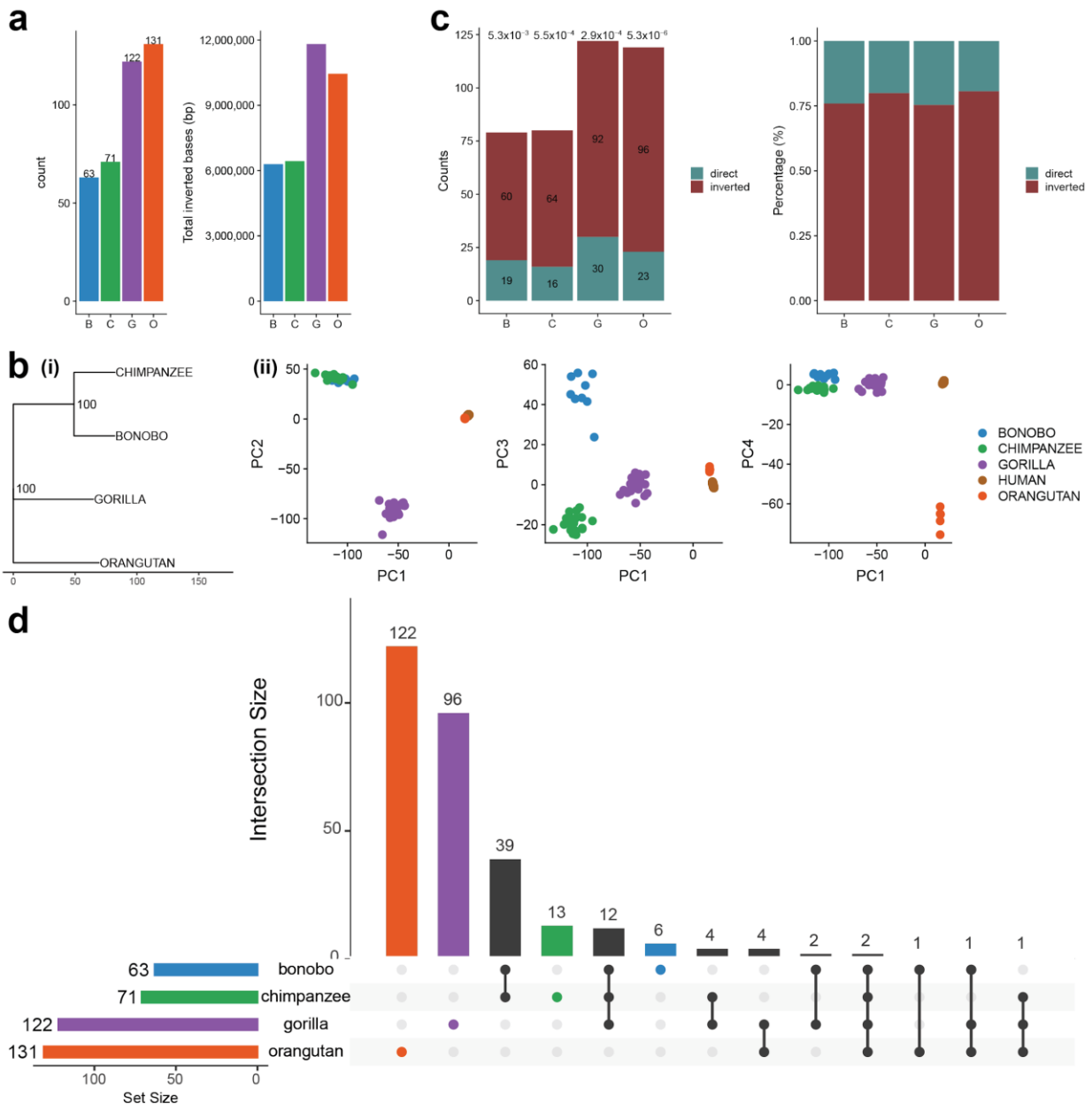
Supplementary Figure 23: Phasing of complex regions on chromosomes 1 and 7.

Inverted directions are shown by orange arrows and direct orientations by a teal arrows. Previously published[8] pathogenic CNVs are shown as red arrows in the top track. Track below shows protein-coding genes (blue arrows) that overlap with either the inversion itself or with a flanking SD, shown as yellow arrows. Dashed rectangle in panel a highlights deemed reference error in human assembly[9] that is likely a minor allele. **a** and **b** show complex genomic region overlapping with pathogenic CNV(s).



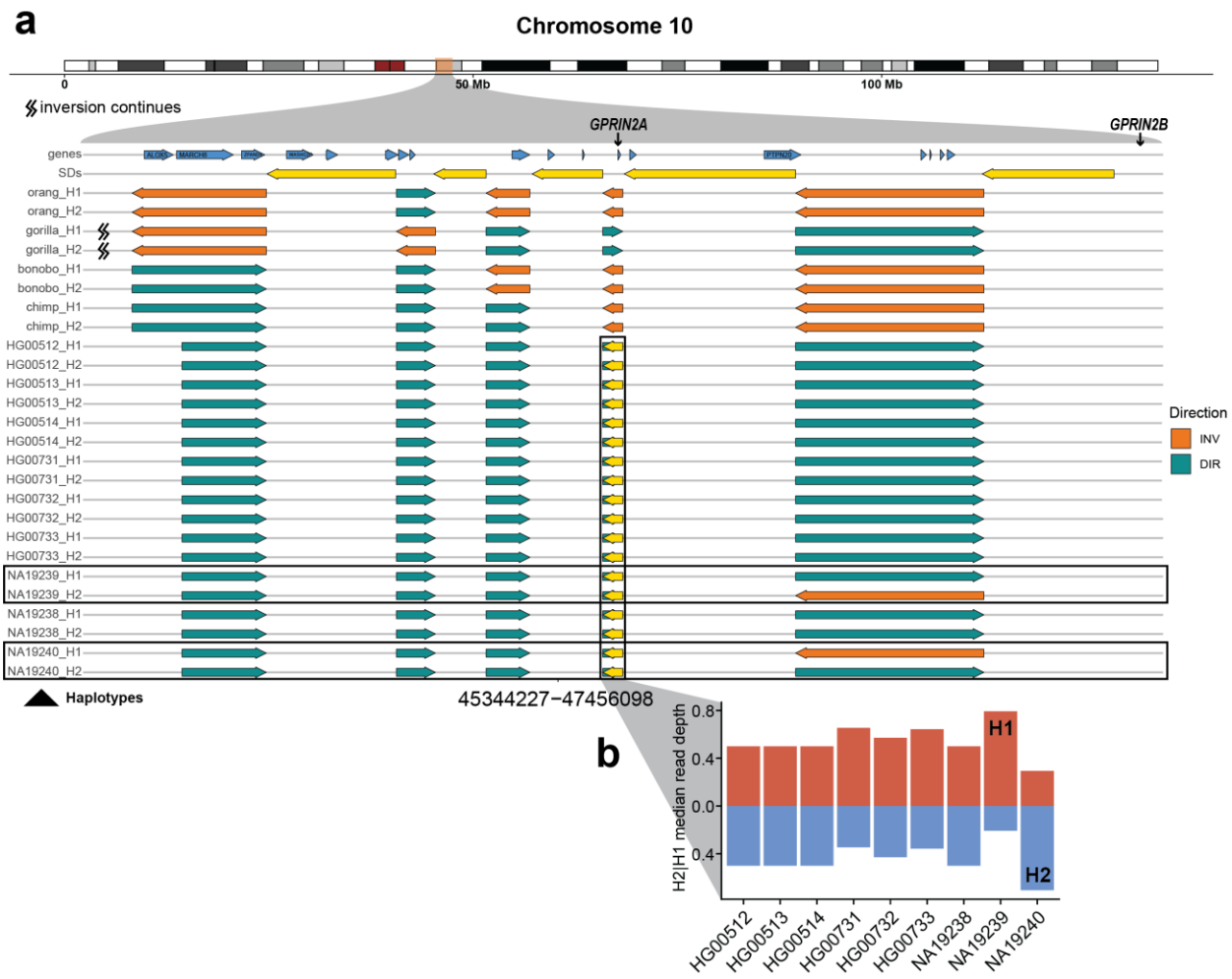
Supplementary Figure 24: Phasing of complex regions on chromosomes 15 and 17.

Inverted directions are shown by orange arrows and direct orientations by a teal arrows. Previously published[8] pathogenic CNVs are shown as red arrows in the top track. Track below shows protein-coding genes (blue arrows) that overlap with either the inversion itself or with a flanking SD, shown as yellow arrows. Dashed rectangle in panel a highlights deemed reference error in human assembly[9] that is likely a minor allele. **a** and **b** show complex genomic region overlapping with pathogenic CNV(s).



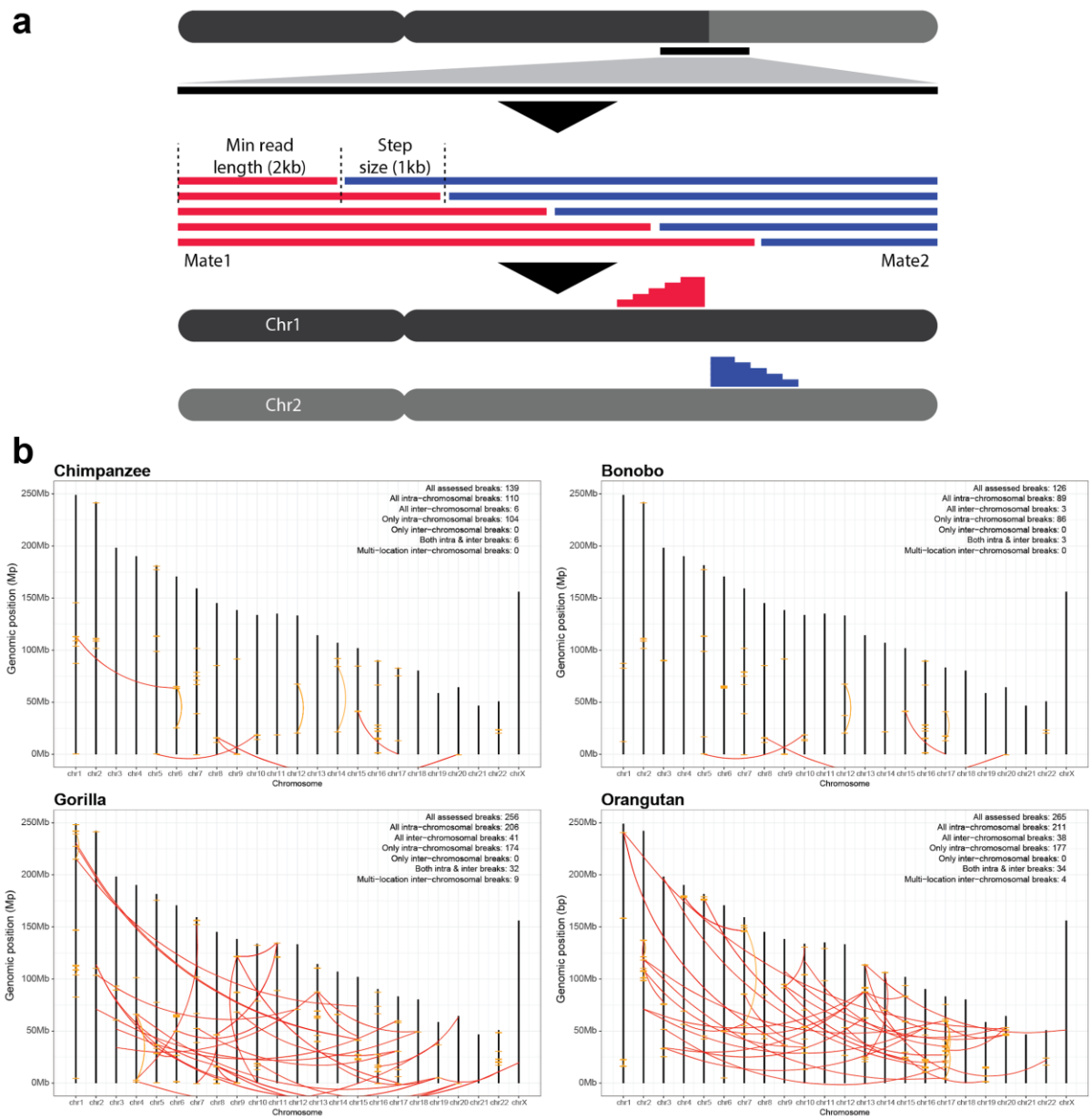
Supplementary Figure 25: Abundance and distribution of inverted duplications.

a, Left: Number of inverted duplications mapped in each NHP (B - bonobo, C - chimpanzee, G - gorilla, O - orangutan). Right: Total number of inverted bases that fall in the inverted duplication callset for each NHP. The excess of inverted duplications in gorilla and orangutan might be partly caused by an elevated number of predicted interchromosomal duplications in gorilla (n=41) and orangutan (n=38) (**Supplementary Figure 21b**). **b**, (i) A neighbor-joining tree constructed (bootstrap iterations: 10,000) based on a mean copy number of inverted duplications genotyped in multiple NHPs (bonobo n=9, chimpanzee n=20, gorilla=23, orangutan n=5) using WSSD. (ii) A PCA of mean copy number of inverted duplications genotyped in multiple NHPs (bonobo n=9, chimpanzee n=20, gorilla=23, orangutan n=5) and humans (n=229). Various combinations of principal components are shown to highlight the separation of different great ape individuals. **c**, Number of mapped duplicated regions in an inverted versus direct orientation from this study. The significance of observed differences between inverted and direct duplications is reported above each bar as p-value (chi-squared with Bonferroni correction). **d**, An upsetR[1] plot showing the number of shared inverted duplications between members of the great ape family (≥50% reciprocal overlap). Orangutan-specific inverted duplications account for 93% (122/131) while gorilla-specific inverted duplications account for 79% (96/122). The majority (n=53) of inverted duplications between bonobo and chimpanzee are shared.



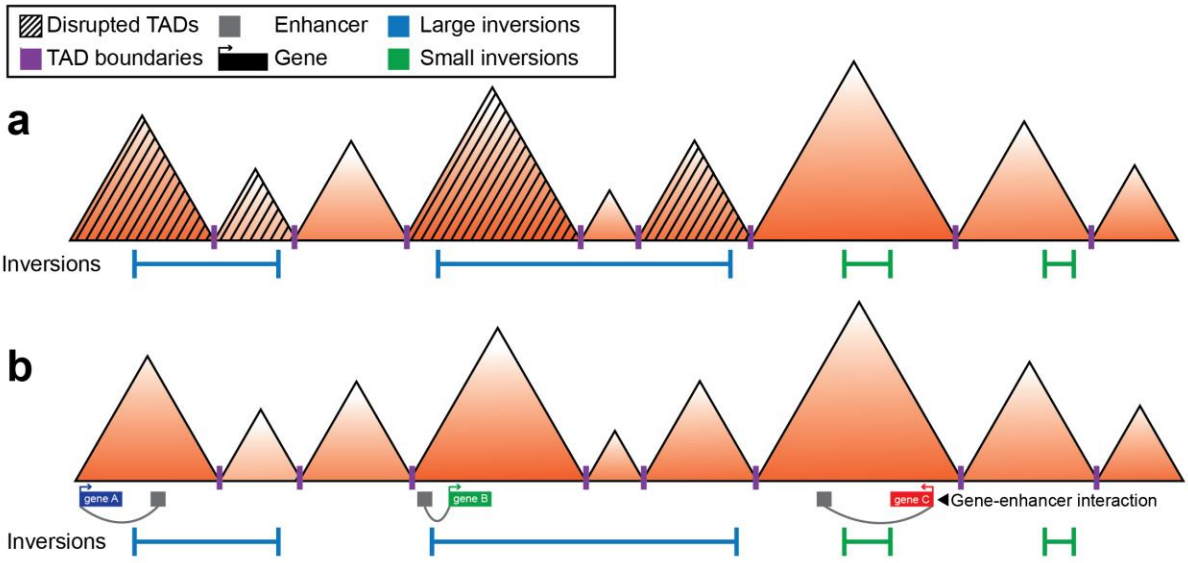
Supplementary Figure 26: Phasing of complex regions on chromosome 10.

a, Inverted directions are shown by orange arrows and direct orientations by a teal arrows. Top track plots protein-coding genes (blue arrows) that overlap with either the inversion itself or with a flanking SD. SDs are shown as yellow arrows. Inverted duplication in *GPRIN2A* region is highlighted by a black box. Heterozygous inversion, in between *GPRIN2* genes, present in NA19239 and NA19240 is highlighted by black boxes. **b**, Median read depth for haplotype-specific Strand-seq reads (red bar - H1, blue bar - H2) over the GRCh38 copy of *GPRIN2* for each HGSVC individual.



Supplementary Figure 27: Genome-wide map of PacBio-supported links between NHP-specific duplicated regions in respect to GRCh38.

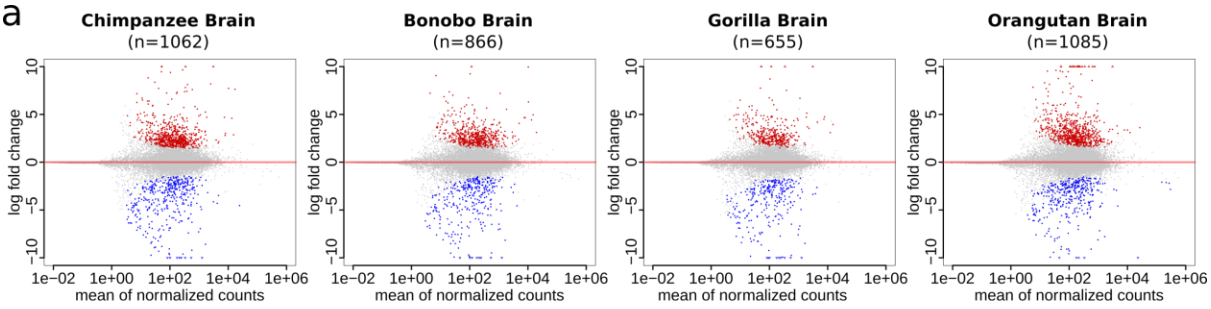
a, Graphical description of a method used to map links between discordantly mapped PacBio read ends. In this procedure every PacBio read is cut in half leaving the left ('red') and right ('blue') portion of the read, which differs by size and are not smaller than set (k-mer=2 kb). Reads are cut in an iterative fashion by moving the breakpoint by 1 kb at a time. All pairs of the left and right portions of the read are considered as mates and are mapped back to the reference genome in paired-end settings (**Methods**). **b**, Genome-wide map of links between paired-end reads produced as described in (a). Paired-end links that reside on different chromosomes are colored in red (interchromosomal) and those that map within the same chromosome are colored in yellow (intrachromosomal).



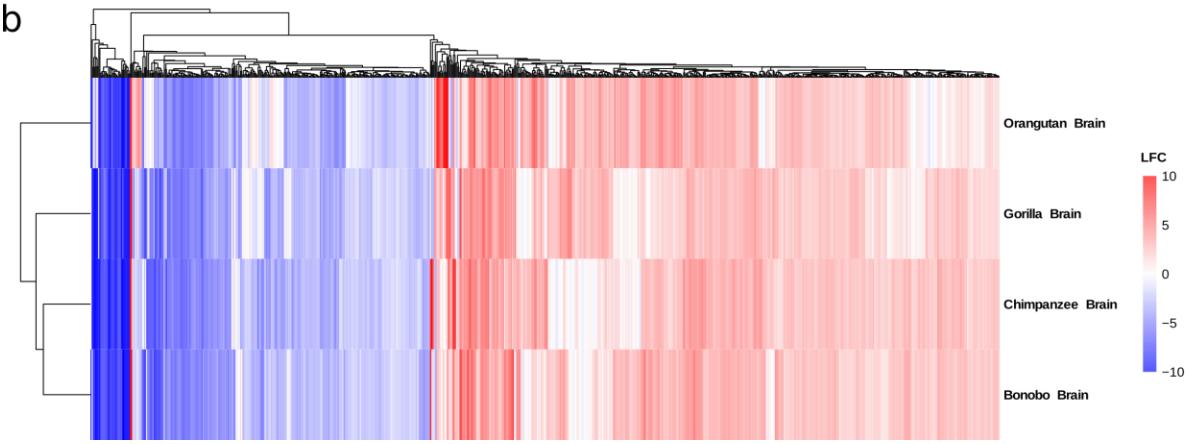
Supplementary Figure 28: Inversions disrupting TAD boundaries and gene-enhancer interactions.

a, An inversion disrupts a TAD boundary when one breakpoint lies within a TAD whereas the other breakpoint does not. The second breakpoint can lie within a separate TAD (as shown, blue) or else outside of the TAD space (e.g., within a centromere or a telomere). If both inversion breakpoints lie within the same TAD (green), it is not classified as disrupted. **b**, An inversion disrupts a gene-enhancer interaction when the gene lies within an inversion and the corresponding enhancer is positioned outside of the inversion, or vice versa.

a



b

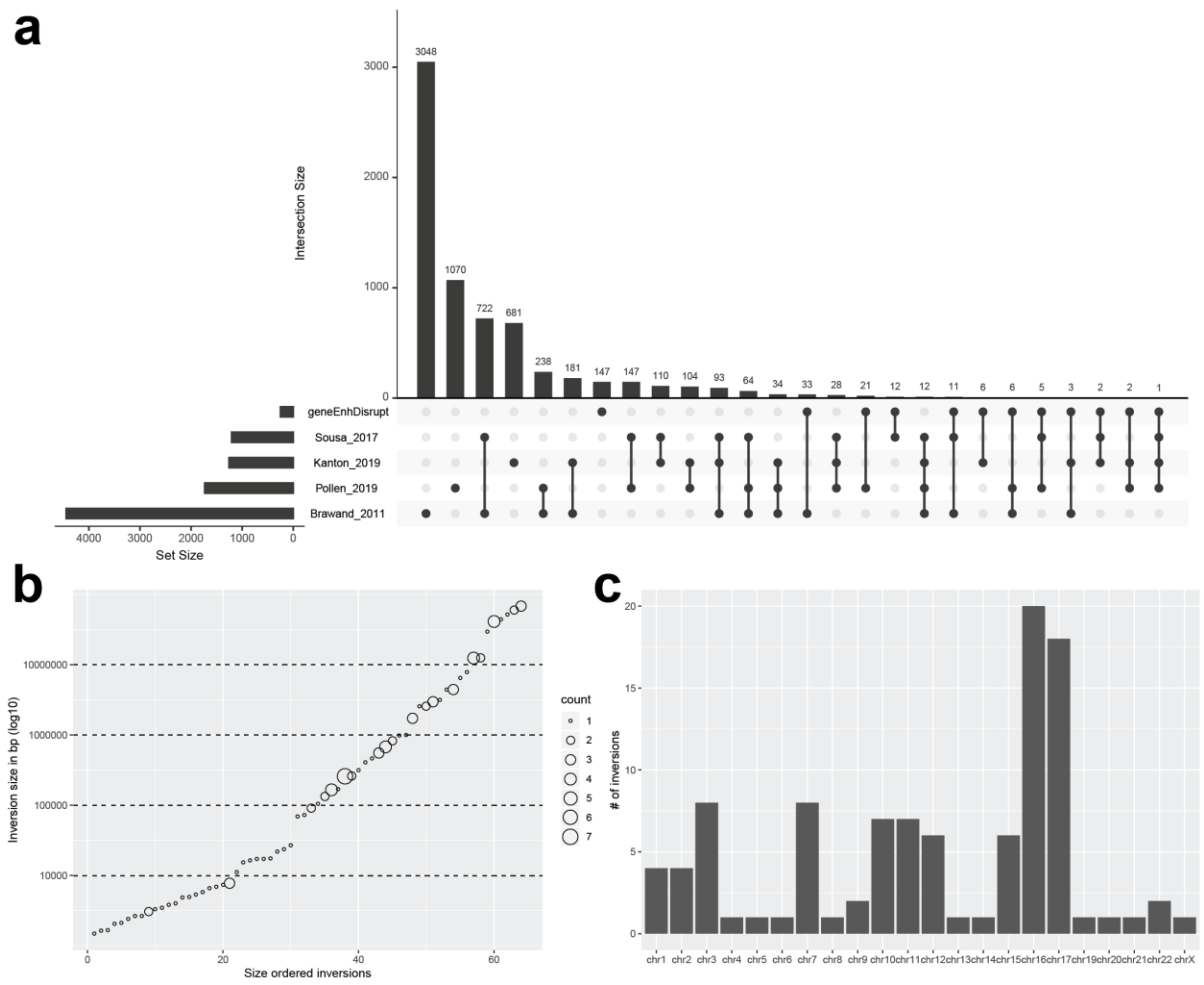


c

	Brain	Cerebellum	Heart	Kidney	Liver	Testis	total
Human	9	2	3	3	2	2	21
Chimpanzee	6	2	2	2	2	1	15
Bonobo	3	2	2	2	2	1	12
Gorilla	2	2	2	2	2	1	11
Orangutan	2	1	2	2	2	0	9

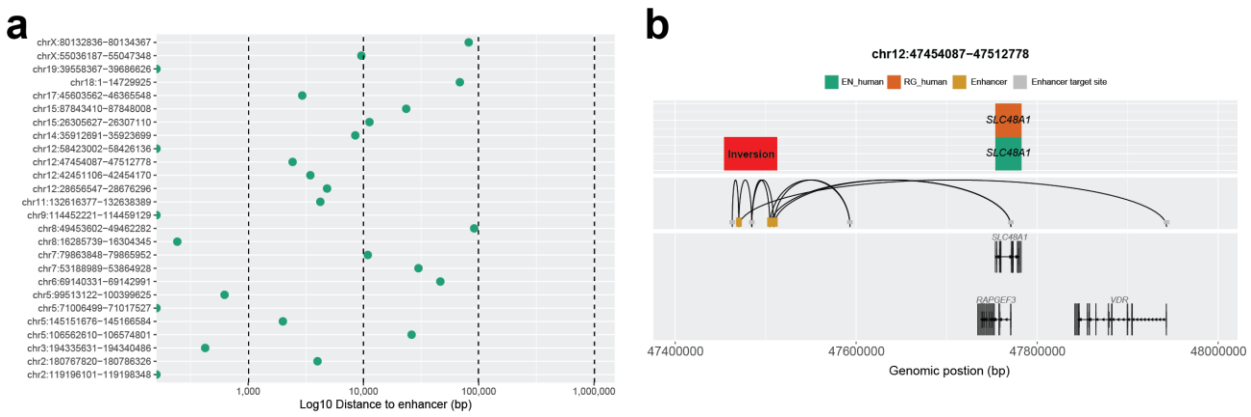
Supplementary Figure 29: DE genes between NHPs.

a, Scatterplots of RNA-seq data[10] show the log2 fold change versus mean read counts for orthologous genes, comparing the ape lineage to human lineage for brain tissue. In each comparison, differentially expressed (DE) genes that have an absolute fold change > 2 and a Shannon information s-value < 0.005 are highlighted, with upregulated genes shown in red, downregulated genes in blue, and the total number of DE genes (N) listed above. **b**, Clustered heatmap showing all DE genes (columns) found in brain tissue of at least one ape lineage (rows) LFC: Log2 fold change. **c**, Summary of tissue-specific sample numbers used for the analysis.



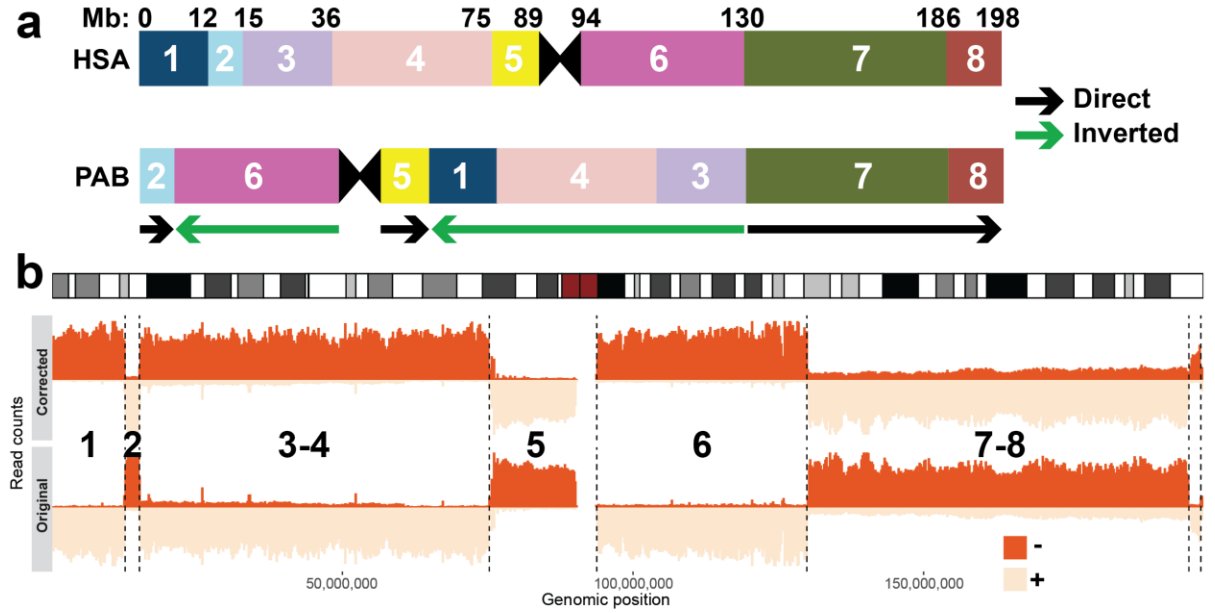
Supplementary Figure 30: Effect of simple inversion on gene expression.

a, An upsetR[1] plot showing the number of shared genes between those predicted to have disrupted gene-enhancer interaction by inversion breakpoint ('GeneEnhDisrupt') and various public dataset that have previously reported DE genes between human and other NHPs (mostly chimpanzee). **b**, Size distribution (ordered by size) of simple inversions that disrupt gene-enhancer interaction of genes that have been found differentially expressed in at least one public dataset from a. Size of the dot represents the number of DE genes, with gene-enhancer disruption, influenced by a given inversion. **c**, A barplot showing the number of inversions (from b) per chromosome.



Supplementary Figure 31: Overlap of human-specific inversions with known enhancers.

a, The shortest distance of known enhancer regions[11] to either left or right breakpoint of predicted human-specific inversions (n=26). **b**, Human-specific inversions (top track, red rectangle) that overlap with known enhancers (middle track in gold). Interacting partners of enhancers are colored in gray and are connected to a corresponding enhancer by a black line. Genes predicted to be upregulated in human are plotted in the top track (EN - excitatory neurons, RG - radial glia). Gene models that overlap with enhancer interacting regions are plotted in the bottom track.



Supplementary Figure 32: Correction of inversion on chromosome 3 in orangutan.

a, Shows cytogenetic derived model of chromosome 3 for human (HSA) and orangutan (PAB). Corresponding genomic segments (direct - black arrow, inverted - green arrow) share the same color and number. **b**, Shows binned (200 kb bin) reads counts for minus ('-' - dark orange) and plus ('+' - light orange) reads. Upper row shows 'corrected' read directionality while bottom row shows 'original' read directionality. Numbers of genomic segments from **a** are shown in the middle.

SUPPLEMENTARY TABLES 1-14

Supplementary Table 1: Summary of inversion validations using orthogonal sequencing technologies and published inversions. (XLSX file)

Supplementary Table 2: Inversion validation using FISH. (XLSX file)

Supplementary Table 3: Inversion validation using phased PacBio assembly. (XLSX file)

Supplementary Table 4: Estimate of inversion rates in great ape lineage.

Lineage	Length	#loci	TMRCA (Mya)	Branch rate estimate	Inverted bp per substitution
bonobo-chimpanzee-gorilla-NA19240-orangutan	610022	6	12.21 – 16.33	-	-
bonobo-chimpanzee-gorilla-NA19240	1135009	3	8.66 – 11.63	0.0078	0.05
bonobo-chimpanzee-NA19240	0	0	7.13 – 10.23	0.0087	N/A
bonobo-chimpanzee	175436893	42	2.30 – 3.97	0.0089	6.87
orangutan	378455756	103	-	0.0077	17.13
gorilla	207205905	53	-	0.0075	9.63
NA19240	2819443	23	-	0.0087	0.11
chimpanzee	8099375	26	-	0.0093	0.30
bonobo	37994601	17	-	0.0089	1.49
gorilla-orangutan	67167875	19	-	-	-
bonobo-chimpanzee-gorilla	54842651	13	-	-	-
bonobo-chimpanzee-gorilla-orangutan	16849479	29	-	-	-
bonobo-chimpanzee-orangutan	5773567	4	-	-	-
bonobo-orangutan	621144	3	-	-	-
bonobo-gorilla	386465	3	-	-	-
bonobo-gorilla-orangutan	72334	3	-	-	-
gorilla-NA19240-orangutan	75687	5	-	-	-
chimpanzee-gorilla	24867	1	-	-	-
chimpanzee-gorilla-orangutan	9705	1	-	-	-
chimpanzee-NA19240-orangutan	9528	1	-	-	-
bonobo-chimpanzee-NA19240-orangutan	6774	1	-	-	-
chimpanzee-orangutan	2269	1	-	-	-
NA19240-orangutan	3037	1	-	-	-

Supplementary Table 5: Summary of NHP simple inversions overlapping with the list of pathogenic CNVs (n=36). (XLSX file)

Supplementary Table 6: Summary of synteny between orangutan, macaque, and mouse for 14 pathogenic CNV regions inverted in orangutan. (XLSX file)

Supplementary Table 7: List of inverted and direct duplications detected based on Strand-seq and WSSD data. (XLSX file)

Supplementary Table 8: Iso-Seq (FLNC) reads summary.

ID	Chimpanzee	Bonobo	Gorilla	Orangutan
Total reads	568684	873053	886776	532888
q10 reads	522328	833166	835637	479528

Supplementary Table 9: List of putative gene fusions in NHPs. (XLSX file)

Supplementary Table 10: DE genes with disrupted gene-enhancer interaction(s). (XLSX file)

Supplementary Table 11: PacBio datasets.

	Accession ID	Est. Depth	Subread length N50 (kb)
Chimpanzee	PRJNA369439	107	17.3
Bonobo	PRJNA526933	85	19.3
Gorilla	PRJNA369439	66.9	20.9
Orangutan	PRJNA369439	81	16.6

Supplementary Table 12: List of NHP assemblies used in this study.

Species	Chimpanzee	Orangutan	Gorilla
Assembly ID	Clint_PTRv1	Susie_PABv1	Kamilah_GGO_v0
BioProject ID	PRJNA369439	PRJNA369439	PRJNA369439
Assembly accession ID	GCA_002880755.3	GCA_002880775.3	SRLZ00000000.1

Supplementary Table 13: List of previously validated misassemblies in GRCh38.

chr	start	end	width	Strand-seq support	Valid
chr11	1894045	1915665	21621	TRUE	Vicente-Salvador et al. 2017; Chaisson et al. 2019
chr12	17770977	17858410	87434	TRUE	Vicente-Salvador et al. 2017; Chaisson et al. 2019
chr12	86846348	86859295	12948	TRUE	Vicente-Salvador et al. 2017; Chaisson et al. 2019
chr16	75206043	75223587	17545	TRUE	Chaisson et al. 2019
chr19	38773571	38791551	17981	TRUE	Chaisson et al. 2019
chr21	40024025	40038796	14772	TRUE	Chaisson et al. 2019
chr4	87926058	87937206	11149	TRUE	Chaisson et al. 2019
chr3	187418672	187425934	7263	TRUE	Chaisson et al. 2019

Supplementary Table 14: Bionano datasets.

For supplementary files like Bionano, first a .csv file is downloaded where each Accession ID and supplementary file contains an FTP download location.

	Accession ID	Supplementary files
Chimpanzee	PRJNA369439	bspq1: SUPPF_0000001269, bsss1: SUPPF_0000001270
Bonobo	PRJNA526933	bspq1: SUPPF_0000003185, bsss1: SUPPF_0000003186
Gorilla	PRJNA369439	dle1: SUPPF_0000003184
Orangutan	PRJNA369439	bspq1: SUPPF_0000001271, bsss1: SUPPF_0000001272

SUPPLEMENTARY NOTE

Artifacts caused by human reference errors

One region of about 20 kb mapping on chr11 (chr11:1894043-1915787) and two regions of 11 kb (chr12:86847401-86858902) and 90 kb (chr12:17772492-17863145) on chr12 show that chimpanzee, bonobo, gorilla, and orangutan are inverted compared to human. However, these three regions have been previously shown to be an error in the orientation of the human reference genome[12,13]. Therefore, the DNA strand switch in these regions is an artifact and the four species are the opposite of what appears to be an inversion, and therefore in the same orientation as human (**Supplementary Table 10**).

The 16p12.1 region (**Fig. 3e**) is made of three regions that are inverted in one or more of the primates analyzed. The two distal inversions have been previously shown to be an error in the reference genome assembly[14] since all humans analyzed (n=10) were in inverted orientation compared to the human reference. Our Strand-seq data of nine human genomes shows that the inverted haplotype is actually a polymorphism and not an error in the reference genome as previously stated, although it represents the minor allele found with a frequency of 5.5%. Even if the reference genome is not misassembled at 16p12.1, it still represents the minor allele, and therefore the primate strand-seq data mapped against it should be the opposite of what appears. This is consistent with previous FISH experiments and capillary sequencing of chimpanzee and gorilla BAC clones[14] for both inversions. Orangutan is discordant but the Strand-seq data shows that the most proximal region is also inverted, suggesting a more complex rearrangement might have occurred in this species.

Inversion correction on chromosome 3 in orangutan

Chromosome 3 is a highly rearranged chromosome, where about half of the chromosome is in inverted orientation compared to human (**Supplementary Figure 32a**). Because of this we used large-scale cytogenetic data to define which synteny blocks are in direct or inverted orientation. While we were able to detect correct inversion boundaries (**Supplementary Figure 32b**, dashed lines), assignment of inverted and direct regions did not correspond to previous cytogenetic studies (**Supplementary Figure 32b**, 'original' read counts). Based on the inversion model published by Ventura et al.[15] and Catacchio et al.[9], inverted blocks should go from 1-12 Mb, 15-75 Mb and 94-130 Mb. We have adjusted our composite files and our inversion call for chromosome 3 in orangutan accordingly (**Supplementary Figure 32b**, 'corrected' read counts). Moreover, we were able to see a small inversion at the end of the q-arm of chromosome 3 that has not been detected by previous cytogenetic studies. By changing our inversion calls for chromosome 3 in orangutan, we went from three homozygous inversions to four homozygous inversions. We redid all analysis that might have been even slightly influenced by this change. This example nicely reflects that for highly rearranged chromosomes more than one source of information is often required to reliably assign inversion status to genomic segments.

Inversion validation by Bionano Genomics

The primate genome maps were constructed as described by Maggiolini et al.[16]. For the current

inversion study, the alignment and SV calling (against GRCh38) were modified. Generally, Bionano detects inversion breakpoints between adjacent match groups in opposite orientation. If a single genome map captures an entire inverted match group with two flanking non-inverted match groups (typically the case for inversions <70 kb), then both inversion breakpoints would be called together. If a genome map can span only one of the two inversion breakpoints—typically the case for inversion size >70 kb—then the breakpoints are called independently. For this study, we changed the parameters on calling the former inversion type. Specifically, the minimum log-score required for a match for the inverted alignment was increased by 100-fold and the label-interval sizing error lowered from 3 to 2.4 (from -hashgen 5 5 3.0 to -hashgen 5 7 2.4). Furthermore, the hash table maximum unresolved site interval decreased (-HSDrange 2.0 0.8 to -HSDrange 1.0). Finally, the overall alignment score (p-value) required of the inverted match group increased from 1E-4 to 1E-6.

Inversion validation by FISH

Interphase nuclei were obtained from lymphoblast cell lines from four chimpanzees (PTR5, PTR12, PTR17, and PTR8), two bonobos (Ulindi and Lb502), and one orangutan (PPY10), and a fibroblast cell line from one gorilla (AG05251). FISH experiments were performed using human fosmid and BAC clones (**Supplementary Table 2**) directly labeled by nick-translation with Cy3-dUTP (Perkin-Elmer), Cy5-dUTP (Perkin-Elmer), and fluorescein-dUTP (Enzo) as described by Lichter et al.[17], 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 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 were simultaneously DAPI stained. Digital images were obtained using a Leica epifluorescence microscope equipped with a cooled CCD camera controlled by iVision-Mac software (v. 4.0.14; BioVision Technologies). DAPI, Cy3, Cy5 and fluorescein fluorescence signals, detected with specific filters and exposure time (0.2sec for DAPI, 3.0sec for Cy3, Cy5 and fluorescein), were recorded separately as gray-scale images with 2x2 binning. Brightness levels were manually adjusted, while pseudocoloring and merging of images were automatically obtained by custom scripts, both using Adobe Photoshop CS6 software.

Inversion validation by targeted long-read assembly

The 10x Genomics linked-read data were generated for three NHPs (chimpanzee, gorilla and orangutan) and processed through the Long Ranger (v.2.2.2) pipeline using the human assembly (GRCh38), and FreeBayes (10x provided version: v0.9.21-7-g7dd41db-dirty) for single-nucleotide variant (SNV) calling. The final VCF contains phased SNV genotypes for three above-mentioned NHP samples with heterozygous-phased SNVs being the most informative for long-read partitioning. We extracted continuous long reads (CLRs) mapped to the breakpoint regions and tagged each read with haplotype 1 or haplotype 2 using the phasing information. We then assembled each haplotype separately using Canu (version 1.9) (Canu parameter: “canu -pacbio-raw corMinCoverage=1 contigFilter="1 500 1.0 .75 1" cnsThreads=1 ovlThreads=1 gnuplot=undef stopOnLowCoverage=1 corMhapSensitivity=high useGrid=false mhapThreads=4”) followed by Arrow polishing. We visualized the breakpoints using dot matrix analysis (DottedPython: https://github.com/ruiyangl/Dottedpython). The binary alignment map

and the phased VCF files from 10x Genomics data were analyzed in combination with the CLR data of the same three individuals as described previously[18].

Evaluating the randomness of chromosome X inversion haplotypes

We created a perfectly ordered set of inversion haplotypes (n=15) for all previously reported polymorphic inversion sites on chromosome X (n=21). We compared dissimilarity matrices of such ordered haplotypes with observed haplotypes using Mantel statistics and observed a significant difference between these two sets of haplotypes (Mantel statistic - Mantel test computes correlation coefficient between two matrices: $P < 0.05$). To simulate random inversion toggling, we flipped the inversion state at a random position in each haplotype for putative 10,000 generations. We compared such randomized haplotypes to our observed haplotype and did not observe a significant difference between these two sets of haplotypes (Mantel statistic: $P > 0.1$).

Effect of inversions on gene-enhancer interrupted genes

To obtain a confident list of protein-coding genes that are differentially expressed likely due to the inversion variant, we searched multiple independent datasets to support our observations. We first preselected protein-coding genes based on interruption of gene-enhancer interaction caused by an inversion breakpoint. All genes that reside on the opposite side from the inversion breakpoint with respect to their corresponding enhancer were included in the analysis. We next searched through gene-expression profiles obtained from Brawand et al.[19] including other datasets reporting DE genes, in various brain regions, between human and at least one NHP (mostly chimpanzee)[19–22]. We required that each preselected gene based on gene-enhancer interaction must appear as differentially expressed in at least one of the above-mentioned datasets either in human or other NHP.

Reusable code

All functionalities used in this study were implemented in the R package called primatR (**Code availability**). Some useful functions utilized in the Methods section are listed below. See package documentation for more details.

getDisjointOverlapsWeighted - calculates percentage of overlap between multiple sets of genomic ranges.

getReciprocalOverlaps - calculates percentage of overlap between two sets of genomic ranges.

genotypeRegions - genotypes set of ranges based on Strand-seq composite files.

processReadLinks - exports a set of interchromosomal links based on PacBio split-read mappings.

getTransChromFusions - exports putative gene fusions based on split-read mappings of FLNC reads.

hotspotter - predicts locations where genomic ranges (breakpoints) cluster expecting random distribution of inversion breakpoints around the genome.

REFERENCES

1. Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**, 2938–2940 (2017).
2. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
3. Giner-Delgado, C. *et al.* Evolutionary and functional impact of common polymorphic inversions in the human genome. *Nat. Commun.* **10**, 4222 (2019).
4. Gel, B. *et al.* regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. *Bioinformatics* **32**, 289–291 (2016).
5. Kong, A. *et al.* Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467**, 1099–1103 (2010).
6. Porubský, D. *et al.* Direct chromosome-length haplotyping by single-cell sequencing. *Genome Res.* **26**, 1565–1574 (2016).
7. Porubsky, D. *et al.* Dense and accurate whole-chromosome haplotyping of individual genomes. *Nat. Commun.* **8**, 1293 (2017).
8. Coe, B. P. *et al.* Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nat. Genet.* **46**, 1063–1071 (2014).
9. Catacchio, C. R. *et al.* Inversion variants in human and primate genomes. *Genome Res.* (2018) doi:10.1101/gr.234831.118.
10. Brawand, D., Soumillon, M., Necsulea, A. & Julien, P. The evolution of gene expression levels in mammalian organs. *Nature* (2011).
11. Fishilevich, S. *et al.* GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database* **2017**, (2017).
12. Vicente-Salvador, D. *et al.* Detailed analysis of inversions predicted between two human genomes: errors, real polymorphisms, and their origin and population distribution. *Hum. Mol. Genet.* **26**, 567–581 (2017).
13. Chaisson, M. J. P. *et al.* Multi-platform discovery of haplotype-resolved structural variation

in human genomes. *Nat. Commun.* **10**, 1784 (2019).

14. Antonacci, F. *et al.* A large and complex structural polymorphism at 16p12.1 underlies microdeletion disease risk. *Nature Genetics* vol. 42 745–750 (2010).
15. Ventura, M. *et al.* Recurrent sites for new centromere seeding. *Genome Res.* **14**, 1696–1703 (2004).
16. Maggiolini, F. A. M. *et al.* Genomic inversions and GOLGA core duplicons underlie disease instability at the 15q25 locus. *PLoS Genet.* **15**, e1008075 (2019).
17. Lichter, P. *et al.* High-resolution mapping of human chromosome 11 by in situ hybridization with cosmid clones. *Science* **247**, 64–69 (1990).
18. Sulovari, A. *et al.* Human-specific tandem repeat expansion and differential gene expression during primate evolution. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 23243–23253 (2019).
19. Brawand, D. *et al.* The evolution of gene expression levels in mammalian organs. *Nature* **478**, 343–348 (2011).
20. Sousa, A. M. M. *et al.* Molecular and cellular reorganization of neural circuits in the human lineage. *Science* **358**, 1027–1032 (2017).
21. Pollen, A. A. *et al.* Establishing Cerebral Organoids as Models of Human-Specific Brain Evolution. *Cell* **176**, 743–756.e17 (2019).
22. Kanton, S. *et al.* Organoid single-cell genomic atlas uncovers human-specific features of brain development. *Nature* **574**, 418–422 (2019).