nature genetics

ARTICLES

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

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] ✉

Inversions play an important role in disease and evolution but are difficult to characterize because their breakpoints map to large repeats. We increased by sixfold the number ($n = 1{,}069$) of previously reported great ape inversions by using single-cell DNA template strand and long-read sequencing. We find that the X chromosome is most enriched (2.5-fold) for inversions, on the basis of its size and duplication content. There is an excess of differentially expressed primate genes near the breakpoints of large (>100 kilobases (kb)) inversions but not smaller events. We show that when great ape lineage-specific duplications emerge, they preferentially (approximately 75%) occur in an inverted orientation compared to that at their ancestral locus. We construct megabase-pair scale haplotypes for individual chromosomes and identify 23 genomic regions that have recurrently toggled between a direct and an inverted state over 15 million years. The direct orientation is most frequently the derived state for human polymorphisms that predispose to recurrent copy number variants associated with neurodevelopmental disease.

Inversions play an important role in disease and genome evolution since they suppress recombination[1] and predispose to nonallelic homologous recombination (NAHR) associated with cancer and neurodevelopmental disease[2]. They are notoriously difficult to detect using both long- and short-read sequencing technologies[3,4] because inversion breakpoints are typically embedded within highly identical segmental duplications (SDs)[5–7] exceeding 50–100 kb in size[8]. It is estimated that more than 50% of inversions within human genomes are flanked by such inaccessible SDs[4,6,9,10]. True inversions are also difficult to distinguish from repeat sequences that have mobilized and inserted in an inverted orientation[11]. As a result, inversions are now recognized as one of the most underascertained forms of structural variation in human[4] and nonhuman primate genomes, limiting our understanding of their evolution[12].

Among apes, the largest cytogenetically visible inversions were first documented by Yunis and Prakash[13] and most subsequent studies have inferred a subset of events by using indirect genomic approaches[12,14–18]. For example, smaller inversions embedded in a unique sequence were readily detected by using paired-end sequencing[19], linked reads[12,20] and assembly-based approaches[12,21]. These approaches especially fail to detect events that are flanked by SDs exceeding the length of the library inserts or sequence read length[17].

In this study, we applied single-cell DNA template strand sequencing (Strand-seq)[22,23] to discover a comprehensive set of inversions in the great ape lineage and leverage long-read sequencing data to validate new events that could not be confirmed by other approaches. Strand-seq is a single-cell sequencing technique that preserves directionality of single-stranded DNA at chromosome-length scale, allowing inversions to be readily detected and genotyped[4,7]. We applied this approach to provide a comprehensive framework for understanding the evolution and recurrence of inversions in the ape lineage.

Results

Great ape inversion discovery. To systematically detect inversions in nonhuman primates (NHPs), we generated Strand-seq data from a representative of each great ape species[22,23]. We selected NHP individuals that differed from those where whole-genome assemblies were recently generated[12], although this complicates the validation of heterozygous events not fixed in each species. We generated 62 high-quality single-cell libraries for chimpanzee (Dorien), 51 for bonobo (Ulindi), 81 for gorilla (GGO 9) and 60 for orangutan (PPY-10) (Table 1, Supplementary Fig. 1a and Methods). Because genome coverage for each single-cell Strand-seq library is low (approximately 0.02×) (Supplementary Fig. 1b), we increased the resolution for smaller inversions (1–50 kb) by concatenating all directional reads across all selected Strand-seq libraries into NHP-specific composite files[4,7] (Fig. 1a, Supplementary Fig. 2 and Methods). Using composite files aligned to the human reference assembly (GRCh38), we detected inverted NHP loci as short as 1 kb in length by tracking changes in read directionality along each chromosome[24] (Methods).

We distinguished three classes of inversions. A homozygous inversion present on both homologs appears as a complete switch in reads mapping in reference orientation to all reads mapping in an inverted orientation (Fig. 1b). A heterozygous inversion resides on

[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

Table 1 | Summary of the Strand-seq inversion callset

Species	Number of Strand-seq libraries	Depth of coverage	Number of simple inversions	Number of inverted duplications
Chimpanzee	62	1.28	159	71
Bonobo	51	0.97	153	63
Gorilla	81	1.68	160	122
Orangutan	60	1.68	210	131

a single homolog and results in a 1:1 ratio of reads mapping in reference and inverted orientation. If there was no associated change in the underlying copy number, we grouped heterozygous and homozygous inversions as 'simple' inversions (Fig. 1b). These are distinct from inverted duplications, where a change in copy number accompanies the localized change in read directionality. This class is often associated with lineage-specific SDs where at least one copy of a given locus resides in the genome in inverted orientation (Fig. 1b).

Among the NHPs, we detected 682 simple inversions and 387 inverted duplications (Fig. 1c, Supplementary Fig. 3 and Methods) with the number of events increasing with phylogenetic distance (Table 1). The vast majority of simple inversions ($n=604$) are homozygous and probably represent fixed differences between humans and NHPs. The remainder ($n=78$) are heterozygous and indicative of inversion polymorphisms within a great ape lineage (Supplementary Fig. 4b). As expected, nearly all (385 out of 387) inverted duplications appeared 'heterozygous', suggesting the duplicated locus occurs in an inverted orientation compared to that at the human locus and is on the same chromosome. These were easily distinguishable from simple heterozygous inversions by the increased sequence read depth over ancestral loci.

We performed extensive validations of both simple inversions and inverted duplications, using a variety of orthogonal sequencing and mapping technologies (Supplementary Table 1 and Methods). Using FISH, for example, we tested five large inversions (between 500 kb and 2.7 megabases (Mb) in size) and confirmed that all were inverted in the predicted great ape (Supplementary Fig. 5, Supplementary Table 2 and Methods). We considered an inversion validated if it overlapped (50% reciprocal overlap) with an inversion call made by an orthogonal technology or an inversion that was already published[12,13,18,21,25] (Supplementary Fig. 6). Additionally, we attempted to assemble the breakpoints of 119 inversions by using a recent phased long-read assembly approach[4,26]. This approach confirmed 27 inversions and provided sequence resolution of the inversion breakpoints (Supplementary Fig. 7, Supplementary Table 3 and Supplementary Note). Altogether, we validated 88% of our simple inversions (Fig. 1d), including most fixed events. Of the inversions that lacked validation, 80% were either heterozygous (and therefore probably polymorphic in the lineage) or flanked by SDs and thus difficult to ascertain by other technologies. We estimate that we increased the number of validated simple inversions more than sixfold (78 versus 521) when compared to those in previous studies (Fig. 1d).

Size and chromosomal distribution. Simple inversions ranged from 1,055 base pairs (bp) to 9.1 Mb in length (Supplementary Fig. 4a). Those flanked by SDs ($n=227$; median 71,873 bp) were significantly larger (two-sided Wilcoxon rank-sum test, $P=1.21\times10^{-19}$) (Fig. 1e) when compared to inversions not flanked by SDs ($n=455$; median 12,476 bp). We note that this difference is probably not due to ascertainment biases associated with previous studies[12,27]. Additionally, we found that inversion size correlated positively with the size of SDs flanking the inversion[28] (Supplementary Fig. 8). Strand-seq detection is much more sensitive than short-read pair mapping approaches because inversion detection does not depend on the mapping of discordant reads to the reference genome[17,29]. Instead, directionality with regard to the reference is embedded in every sequencing read, allowing for the unambiguous detection of inversions even when flanked by complex SDs[7]. Similarly, inverted duplications, which probably arise by duplicative transposition, show a wide size distribution (range 10,171–1,708,343 bp; median 48,421 bp) but rarely exceed 1 Mb in length, suggesting an upper bound for SD formation (Fig. 1e). Of note, we set a lower limit for inverted duplication calls at 10 kb.

While the number of simple inversions generally correlates with chromosome length ($R^2=0.3$) (Fig. 1f,g), the X chromosome is an exception with approximately 2.5-fold more inversions when compared to autosomal length (z-score = 3.57, one-sided $P=0.000177$) (Fig. 1f). This difference is even more pronounced for heterozygous inversions (4.6-fold), consistent with elevated rates of inversion polymorphism on the X chromosome (Supplementary Fig. 9). X chromosome inversions also showed a tighter size distribution (up to approximately 100 kb) compared to autosomes (Supplementary Fig. 10), possibly due to differences in the underlying architecture of SDs.

Unlike simple inversions, the number of inverted duplications correlates less strongly with chromosome size ($R^2=0.1$) (Supplementary Fig. 11) but instead with SD content in the human genome ($R^2=0.301$) (Fig. 1g). This is expected since lineage-specific duplications are tenfold more likely to arise adjacent to ancestral duplicated sequences shared between two ape species[30,31]. If such a duplication arises in an inverted orientation, it will appear as an inverted duplication. For example, human chromosomes 5, 7, 10, 16 and 17 are among the most SD-rich chromosomes and similarly showed the greatest density of ape inverted duplications, often in close proximity to known human SDs (Supplementary Fig. 12). Once again, the clear exception is the X chromosome, which shows an excess of inverted duplications (Fig. 1g) with regard to autosomes given the chromosomal SD content.

Phylogenetic reconstruction. We compared the distribution of inversions among all great apes, including an African human sample[4] (NA19240) (Fig. 2a, colored bars). Human-specific inversions are identifiable as loci that were inverted in all NHPs compared to the 'direct' orientation in the human genome (Fig. 2b, left). We identified 26 total human-specific inversions, of which only 6 were previously reported (Fig. 2b, right, and Methods)[12,18]. Excluding human reference genome misassemblies[4], we classified all human-specific inversions as ancestral or lineage-specific, parsimoniously assigning them to an ape phylogenetic tree (Methods). We placed 60 inversions on ancestral branches of the great ape phylogeny, with the majority ($n=45$) occurring on the ancestral *Pan* lineage (Fig. 2c). This was expected due to the recent divergence of chimpanzee and bonobo. Approximately 27% (16 out of 60) of all ancestral inversions are heterozygous in one or more ape species and are probably polymorphic.

Using a nonredundant dataset of simple autosomal inversions ($n=358$), we constructed a Bayesian evolutionary tree (Fig. 2d and Methods) and estimated the rate of fixation of simple inversions as approximately 7 autosomal inversions per million years of evolution. No ape lineage showed evidence of inversion acceleration, with branch rates ranging from 0.0075 to 0.0093 inversions per locus (Supplementary Table 4); however, we observed variable inversion rates after accounting for the number of inverted bp per single-base-pair substitution (range: 0.05–17.13) (Supplementary Table 4). Interestingly, we identified 27 inverted loci that showed evidence of homoplasy (Fig. 2a, asterisks) either due to recurrent mutation or incomplete lineage sorting. For instance, 5 inversions shared between human, gorilla and orangutan were absent in the

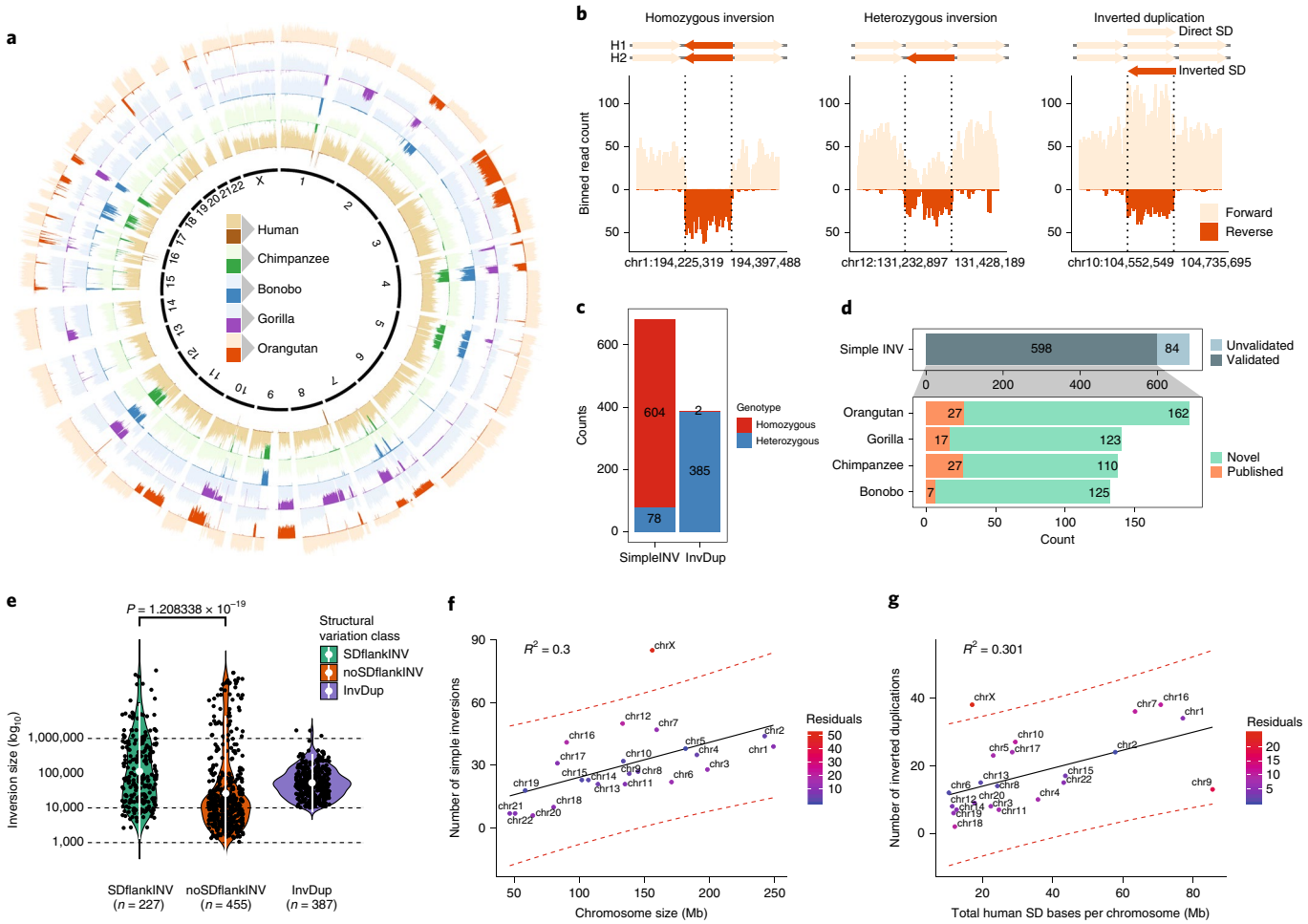


Fig. 1 | Inversion call summary. a, Circular representation of composite files for each member of a great ape family. The genome of each individual is divided into 500-kb bins and the number of reads mapped in forward (light color) and reverse (dark color) orientation in each bin is depicted as a bar along each chromosome. **b**, Example of inversion classes mapped in orangutan. Directional reads are binned into 10-kb bins (step 5 kb) and the number of reads mapped in forward (light color) and reverse (dark color) orientation is depicted as a vertical bar along a given genomic region. Inverted loci are highlighted by dashed lines. **c**, Summary of all inversions and inverted duplications mapped in this study. The inner circle summarizes the number of events found for each structural variation class (simple inversion, INV; inverted duplication, invDup). **d**, Summary of validated simple inversions by other orthogonal technologies (top) and summary of validated simple inversions (bottom) that appear to be new in comparison to previously published data. **e**, Size distribution of simple inversions flanked by SDs (SDflankINV, $n=227$), simple inversions not flanked by SDs (noSDflankINV, $n=455$) and inverted duplications (invDup, $n=387$). The white dot shows the mean of each distribution along with its interquartile range (Wilcoxon rank-sum test). **f**, Scatterplot of the number of simple inversions ($n=682$) given the chromosome size. **g**, Scatterplot showing the number of inverted duplications ($n=387$) given the total length of known human SDs per chromosome. **f**,**g**, The regression line is added as a solid black line and the 95% confidence intervals are highlighted as red dashed lines. Deviation from an expected number of inversions is expressed in the number of residuals.

Pan lineage, 11 out of the 27 likely recurrent loci resided on chromosome X and 85% (23 out of 27) were flanked by known human SDs.

Human polymorphism and inversion hotspots. We compared the 388 nonredundant simple ape inversions (including the X chromosome) to 150 simple human inversions recently described for six humans of diverse ancestry[4] (Supplementary Fig. 13 and Methods). Strikingly, we found that one-third (49 out of 150) of the human polymorphic inversions overlapped with an inversion detected in an NHP (Fig. 3a). Of these, 43% (21 out of 49) mapped to the X chromosome (Fig. 3b, top track) and 31% (15 out of 49) corresponded to the aforementioned recurrent ape inversion sites ($n=27$). Notably, more than half (27 out of 49) of these loci were heterozygous in an NHP lineage (Supplementary Fig. 14), which is evidence of polymorphism across multiple ape lineages. The majority (38 out of 49) of these inversions were flanked by highly homologous SDs (Fig. 3b, bottom track), with 10 of these regions being polymorphic in a larger genotyping panel of the human population (Supplementary Fig. 15).

Inversion breakpoints were not randomly distributed but clustered into 23 discrete genomic regions (median size of 5.5 Mb) (Fig. 3c and Supplementary Fig. 16) enriched for human female meiotic recombination hotspots ($P=0.021$, z-score$=2.438$) (Supplementary Fig. 17a). Twelve of these clusters harbored half (25 out of 49) of the inversions shared between humans and NHPs. As expected, breakpoint clusters were enriched approximately 5.6-fold for SDs (Fig. 3c inset), with chromosomes 16, 17 and X harboring the greatest number. For example, we observed three distinct inversion clusters on chromosome X that encompassed 21 inversions shared between humans and NHPs (Supplementary Fig. 18). Using the phase information embedded in the Strand-seq data, we ordered and phased all 21 inversions along the entire

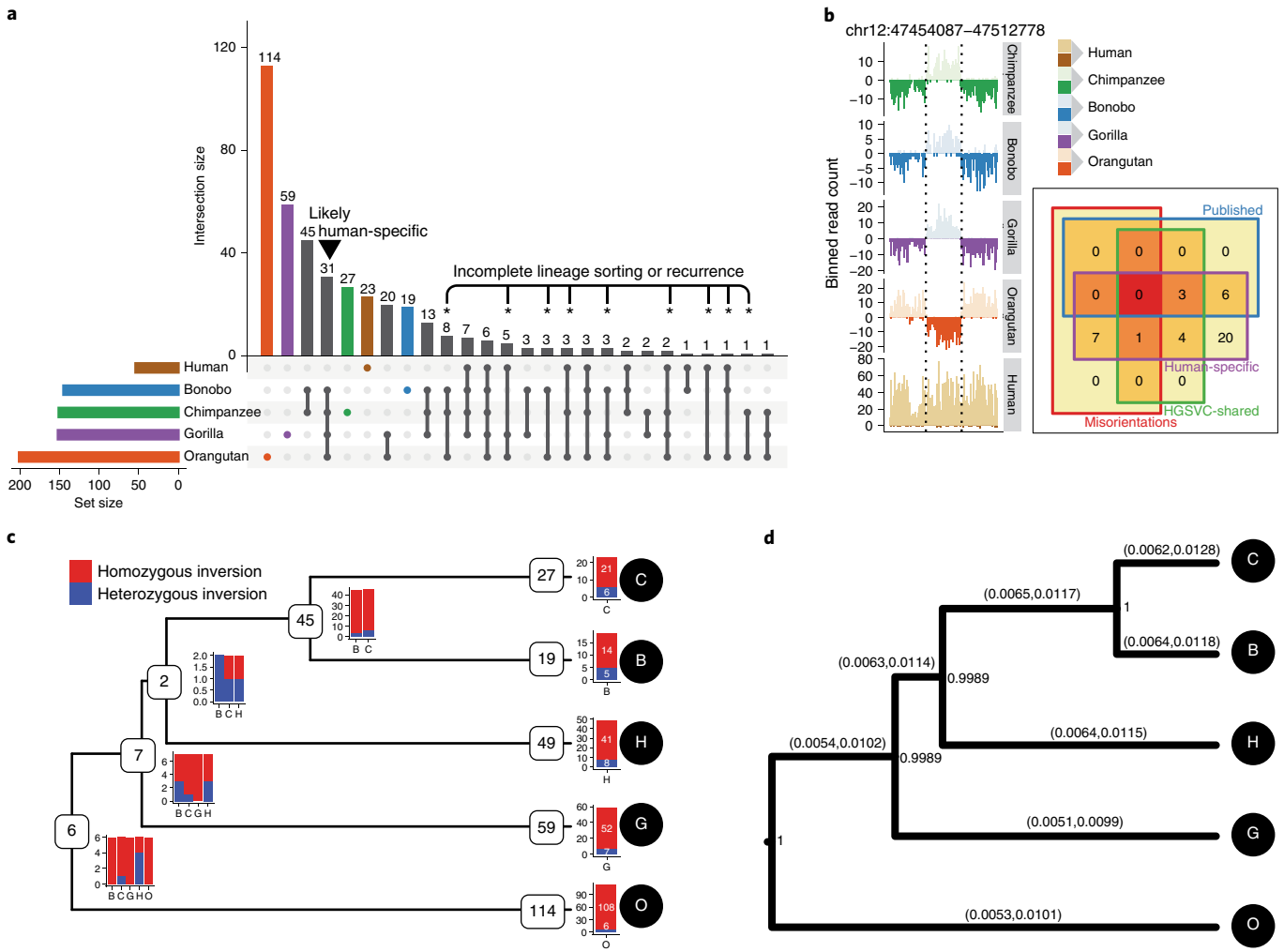


Fig. 2 | Lineage-specific simple inversions and their evolutionary rates. a, An UpSetR[62] plot showing the number of shared inversions between members of the great ape family (≥50% reciprocal overlap). The black arrowhead points to putative human-specific inversions. The asterisks highlight inversions with recurrent or incomplete lineage sorting signatures. **b**, Example of a human-specific inversion, predicted based on Strand-seq data. The inverted region is highlighted by dashed lines. Human-specific inversion is deemed as a region inverted in all NHPs with regard to the flanking region, but in direct orientation in humans. Inset: Venn diagram showing predicted human-specific inversions with regard to known genome minor alleles/misorientations, human inversion polymorphisms[4] and already published human-specific inverted loci. **c**, A tree constructed based on shared simple inversions (≥50% reciprocal overlap) by using hierarchical clustering. Each branching node contains a number of shared inversions in a given subtree together with a barplot showing inversion genotypes per individual (B, bonobo; C, chimpanzee; G, gorilla; H, human; O, orangutan). The tips of the tree contain the number of inversions without a significant overlap (<50%) with any other inversion and are probably species-specific. The barplot showing the inversion genotypes for such species-specific inversions is plotted at each tip of the tree (Methods). **d**, A rooted Markov Chain Monte Carlo (MCMC) evolutionary tree constructed based on a nonredundant set of 358 autosomal simple inversions among great apes. Inversion rates are reported for each branch as the 95% highest posterior density confidence intervals. The numbers at each branching node provide posterior support for this tree topology based on 10,000 MCMC trees sampled from an MCMC chain of 10,000,000 samples constructed from these data.

length of chromosome X (Fig. 3d, Supplementary Fig. 19 and Methods), which revealed a remarkable degree of evolutionary toggling between humans and NHPs with SDs bracketing recurrently inverting regions and frequently containing protein-coding genes (Fig. 3d, top track, and Supplementary Fig. 20). Each human haplotype in these regions showed a unique combination of inverted and directly orientated loci ($n=21$) (Supplementary Fig. 21) and was not significantly different from a random inversion state at these loci (Mantel statistic, $P=0.162$; low bootstrap support; Supplementary Note). A similar pattern of inversion toggling was observed in two regions on chromosome 16 (Fig. 3e). Interestingly, both the X chromosome and the reported regions on chromosome 16 were biased toward female meiotic recombination (Supplementary Fig. 17b).

Because inversion polymorphic regions have been associated with particular recurrent rearrangements[8,32,33], we investigated 36 recurrent large-scale copy number variants (CNVs) associated with neurodevelopmental disorders in humans[34] and found that 47% (17 out of 36) of these overlap (50% reciprocal overlap) with our map of NHP inversions. This represents an approximately 14-fold enrichment when compared to a random simulation of pathogenic CNVs (z-score = 17.2, two-sided $P=2.09\times10^{-66}$, 100 iterations) (Supplementary Fig. 22a, Supplementary Table 5 and Methods). Two of these inversions are classified as occurring specifically in the human lineage, two inversions are known to be polymorphic in humans, while the remaining 13 are observed as simple NHP inversions (Supplementary Fig. 22b). At the species level, orangutan showed the greatest correspondence with

42% (15 out of 36) of recurrent CNVs overlapping an inversion and the highest frequency (0.88) of inverted loci at these regions (Supplementary Fig. 22c). In about half of these cases (8 out of 15), orangutan represents the ancestral configuration based on synteny analysis with macaque and mouse models (Supplementary Table 6). Interestingly, most of these CNV hotspot regions are in an inverted orientation in at least one NHP while most human haplotypes are in direct orientation with regard to the human reference (Supplementary Figs. 23 and 24).

Inverted orientation bias for lineage-specific duplications. A relatively unique feature of the Strand-seq assay is the ability to distinguish simple inversions from inverted duplications associated with a copy number change[4] (Fig. 1b). In contrast to simple inversions, which accumulate relatively uniformly between ape lineages, we observed a slight excess of inverted duplications in gorilla and orangutan when compared to bonobo and chimpanzee (Supplementary Fig. 25a), although the number of lineage-specific duplications generally recapitulated the ape phylogeny (Supplementary Fig. 25b). Taking advantage of short-read sequencing data from 286 human, ape and archaic hominin genomes, we genotyped copy numbers and assayed lineage specificity for 387 inverted duplications (Methods). The majority of orangutan (93%) and gorilla (79%) copy number increases are lineage-specific in comparison to chimpanzee and bonobo, where >50% of the inverted duplications are shared (50% reciprocal overlap) due to their more recent divergence (Fig. 4a and Supplementary Fig. 25d). We highlighted a human-specific duplication of *GPRIN2* that was recently shown to be missing from the human reference (GRCh38) (ref. [35]) (Supplementary Fig. 26). Using an independent map of great ape-specific duplications[31,36], we investigated if SDs showed a preferential bias in their orientation (Methods). Excluding interchromosomal events (Supplementary Fig. 27 and Methods), we found that approximately 78% of lineage-specific duplications map in an inverted orientation (Bonferroni-corrected $P<0.005$) (Fig. 4b). If we limit the analysis to only those lineage-specific duplications with no more than 1 or 2 additional copies ($n=3$ or 4 copy number estimate in a diploid genome), this bias remains significant with approximately 75% of lineage-specific duplications occurring in an inverted orientation. In addition to this orientation bias, we predicted an enrichment of inverted duplications mapping near the ends of chromosomes (last 5% of a chromosomal arm) with this difference being the most pronounced in gorilla ($P=0.001$) (Fig. 4c)[37].

Rearrangement and NHP gene expression differences. The association of inversions and SDs creates the potential for the formation of previously unknown fusion transcripts and genes during evolution. We examined all NHP inverted regions, searching for the presence of previously unknown fusion genes based on a comparison of long-read genome sequence data and full-length nonchimeric (FLNC) transcripts generated for the different NHP species (Supplementary Table 8). We detected 15 putative fusion transcripts of which three were further supported by long-read Pacific Biosciences (PacBio) data (Supplementary Table 9 and Methods). We identified a fusion gene specific to the gorilla lineage that was created by inverted duplication and reintegration of a segment of DNA between chromosomes 4 and 7 (Fig. 4d). This fusion is supported by both split-read mapping of FLNC transcripts and long PacBio reads.

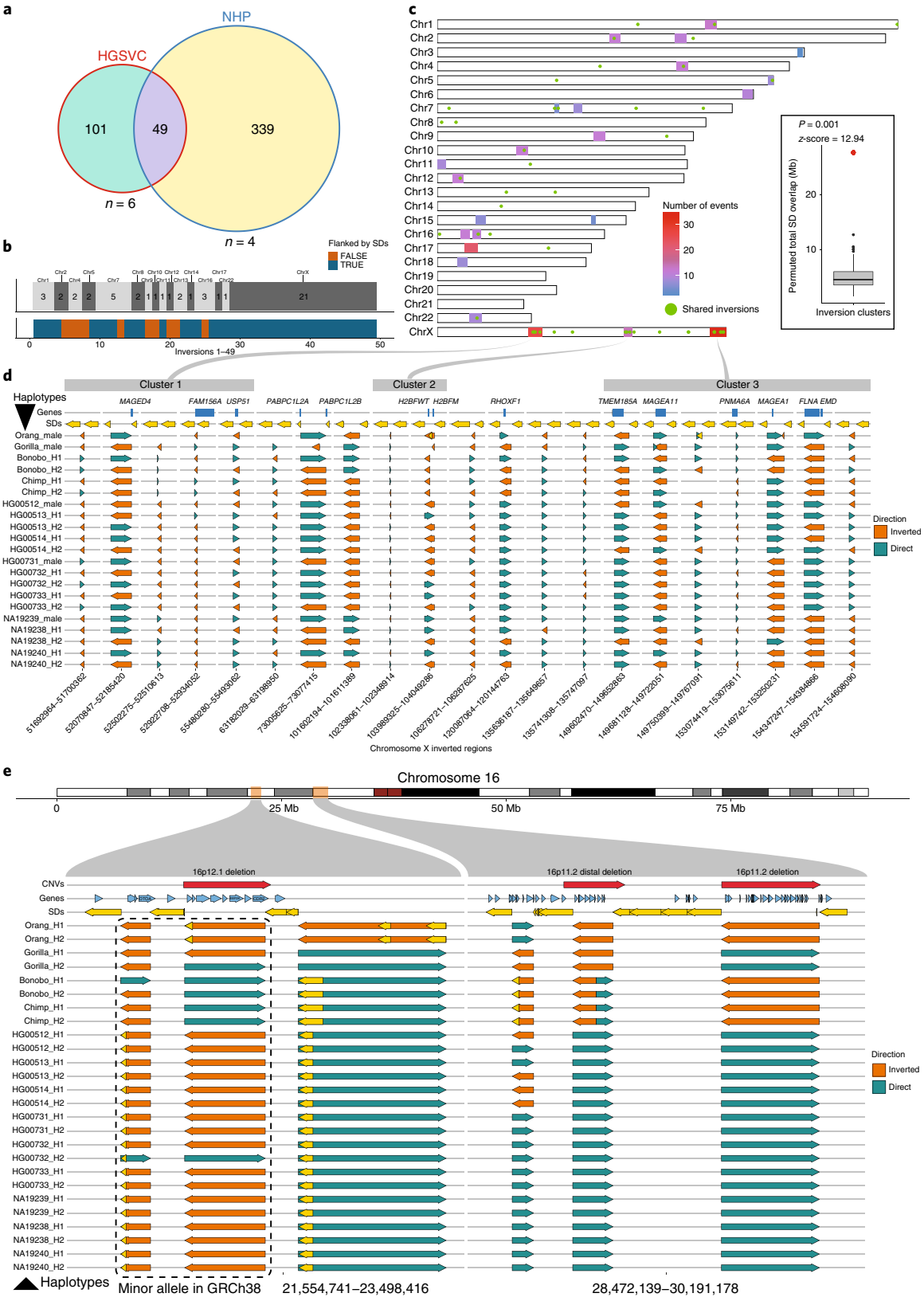
Inversions also carry the potential to rearrange gene regulatory regions and thus perturb gene–enhancer interactions, for example, by disrupting the structure of topologically associating domains (TADs), as reported previously in the context of human diseases[38,39] (Supplementary Fig. 28). Notably, we found that breakpoints of larger inversions (>100 kb) tended to colocalize with human-defined TAD boundaries[40] (Fig. 5a,b), whereas shorter (<100 kb) inversions do not show such tendency; instead, their breakpoints appear to be strongly depleted from TAD boundaries (Fig. 5b, inset). Next, we investigated the effect of these large-scale balanced rearrangements on primate gene expression by analyzing bulk RNA sequencing (RNA-seq) data from 21 human and 47 NHP samples spanning 6 tissues[41]. Per tissue, we observed a median of 1,499 differentially expressed genes in each NHP (compared to the corresponding human tissue). We found that differentially expressed genes were located more frequently (approximately 1.15-fold increase, $P=0.0048$, one-sided permutation test; Methods) in TADs disrupted by an inversion compared to intact TADs that did not contain an inversion breakpoint (Fig. 5c). When testing differential expression with respect to inversion breakpoints, we observed more differentially expressed genes near the breakpoints of large inversions (>100 kb), when compared to small inversions (<100 kb) (Fig. 5d). We further investigated this effect by using other recently published datasets[41–44] with a specific emphasis on brain genes. We preselected protein-coding genes with disrupted gene–enhancer interaction at the breakpoints of 388 nonredundant simple inversions. In total, we found 249 candidate genes, of which 102 are differentially expressed genes in at least one from the aforementioned datasets, with 30 genes confirmed by two datasets (Supplementary Fig. 30a, Supplementary Table 10 and Supplementary Note), including neurodevelopmental disease genes (for example, *SETD7* or *CTNNA3*). In line with the previous analysis[45], we continued to observe the trend of more differentially expressed genes located near the breakpoints of larger inversions (>100 kb) (Supplementary Fig. 30b, see the circle sizes).

Discussion

Inversions have long been thought to be a driving force in human evolution with the potential to reduce recombination, create fusion genes and alter patterns of gene expression[46,47]. We assessed the latter by comparing regions of ape inversion with corresponding NHP bulk RNA-seq data[41] with previously defined TADs[40]. Notably, we observed evidence that large (>100 kb) inversions may

Fig. 3 | Shared inversions and inversion hotspots. a, Venn diagram showing overlapping simple inversions (50% reciprocal overlap) between the HGSVC nonredundant and NHP redundant datasets. **b**, Top tracks: number of shared inversions between the HGSVC and NHP datasets from **a**, shown as counts per chromosome. Bottom track: inversions flanked by SDs are colored blue; those not flanked are colored orange. **c**, A genome-wide map of detected inversion breakpoint clusters based on simple inversions from HGSVC and NHP datasets. A set of inversions ($n=49$) from **a** is plotted over this genome-wide map as green dots. Inset: comparison of the total number of SD bp mapping to the 23 breakpoint clusters (red dot, observed = 29,138,268) and a random genome-wide simulation ($n=1{,}000$ permutations; RegioneR[63] permTEST). Minimum: 1,653,112; first quantile: 3,757,630; median: 5,301,424; third quantile: 7,084,019; maximum: 11,940,452). **d**, Each row represents a haplotype with all tested inversions phased along the whole X chromosome. The inverted direction is shown by an orange arrow and direct orientation by a teal arrow. The top track plots protein-coding genes (blue rectangles) that overlap with either the inversion itself or with flanking SDs, which are shown as yellow arrows. Previously defined inversion breakpoint clusters are shown as gray rectangles at the top of the figure and are linked to their location on chromosome X in **c**. **e**, Each row represents a haplotype with all tested inversions phased along the whole of chromosome 16. The inverted direction is shown by an orange arrow and direct orientation by a teal arrow. The top track plots protein-coding genes (blue arrows) that overlap either the inversion itself or with a flanking SD, shown as yellow arrows. Previously published[34] pathogenic CNVs are shown as red arrows in the top track.

mediate gene regulatory changes in NHP evolution, unlike smaller inversions, which are rarely associated with NHP gene expression changes. Irrespective of this, inversions are responsible only for a relatively small number of differentially expressed changes (approximately 1.15-fold enrichment of differentially expressed genes), suggesting more complex gene regulatory relationships[45]. We further found that 15 out of 26 human-specific inversions have known enhancer regions[48] within 5 kb distance (Supplementary Fig. 31a). For example, a human-specific inversion on chromosome 12 repositioned an enhancer in the vicinity of *SLC48A1*, which was



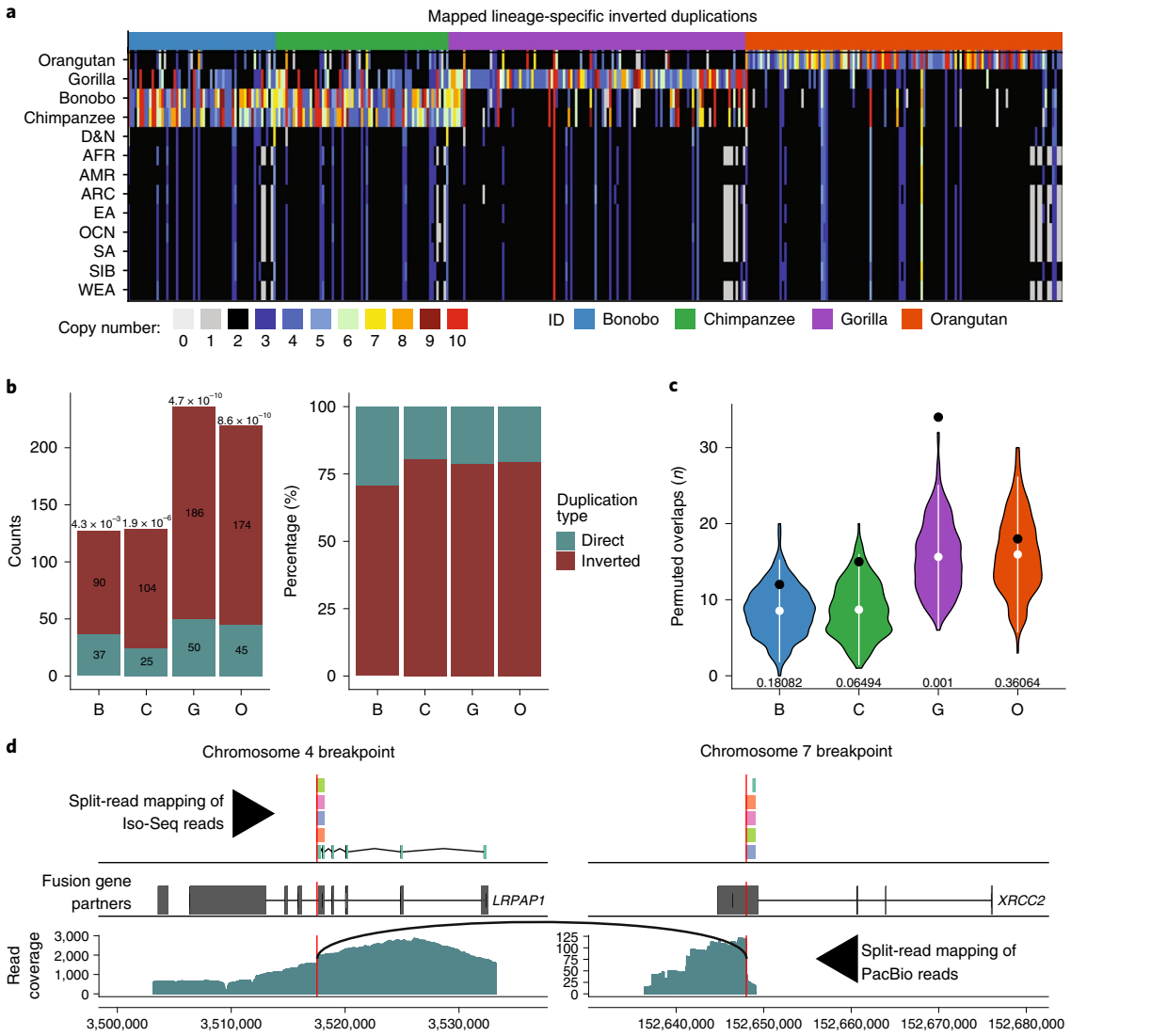


Fig. 4 | Evolutionary impact of inverted duplications. a, Heatmap of estimated copy number (mean copy number) per inverted duplication (columns) in multiple human populations and NHPs (rows). **b**, Left: number of mapped duplicated regions in inverted versus direct orientation. The significance of observed differences between inverted and direct duplications is reported above each bar as a *P* value (chi-squared with Bonferroni correction). Right: each bar shows the proportions of inverted and direct duplications per NHP (colored as shown in **a**). **c**, Enrichment analysis of inverted duplication in a 0.05 fraction of each chromosome end (1–22 and X). The observed counts are shown by a black dot; the distributions of permuted counts ($n = 1{,}000$ permutations; RegioneR[63] permTEST) are depicted by violin plots. The white dots show the mean of each distribution (B-8.54, C-8.69, G-15.6, O-16). At the bottom of each distribution, there is a *P* value showing the significance of the difference between observed (B-12, C-15 G-34 O-18) and permuted counts. **d**, Predicted gene fusion between *XRCC2* on chromosome 7 and *LRPAP1* on chromosome 4. Upper track: split-read mappings of Iso-Seq reads over the predicted breakpoint (red vertical line). Iso-Seq reads that belong to the same transcript share the same color. Middle track: gene models of aforementioned genes (exons, wide boxes; introns, lines in-between). Lower track: split-read mapping of PacBio reads over the fusion breakpoint on chromosome 4. The black arc line connects the ends of PacBio reads with the split-read mappings.

previously shown to be upregulated in human neuronal cells (excitatory and inhibitory neurons) and radial glia[43]. *SLC48A1* shows the highest expression in the spinal cord and enhances tumorigenic functions of non-small-cell lung cancer cells and tumor growth[49] (Supplementary Fig. 31b).

Our analysis shows that inversions are among the most biased forms of genetic variation showing a highly nonrandom distribution. The X chromosome is the greatest outlier with approximately 2.5-fold more inversions based on its size and duplication content when compared to ape autosomes (Fig. 1f). This difference is most pronounced for heterozygous inversions, suggesting elevated rates of inversion polymorphism for the X chromosome (Supplementary Fig. 9). It has been hypothesized that X chromosome hemizygosity and the absence of male recombination (outside the pseudoautosomal region) may be responsible for the abundance of X chromosome inversions by promoting NAHR for unpaired X chromosomes during meiosis[50]. It is possible that regions of sex-biased recombination may be particularly prone to inversions if such regions are more likely to fail to pair homologously during meiosis, allowing preferential intrachromosomal or interchromatidal exchange of genomic regions between duplicated sequences. Importantly, regions of inversion toggling, such as chromosome 16, are also known to be hotspots for NAHR associated with recurrent rearrangement, which is commonly seen in neurodevelopmental delay[34] (Fig. 3e, red arrows). It is also interesting that the size distribution of simple inversions on the X

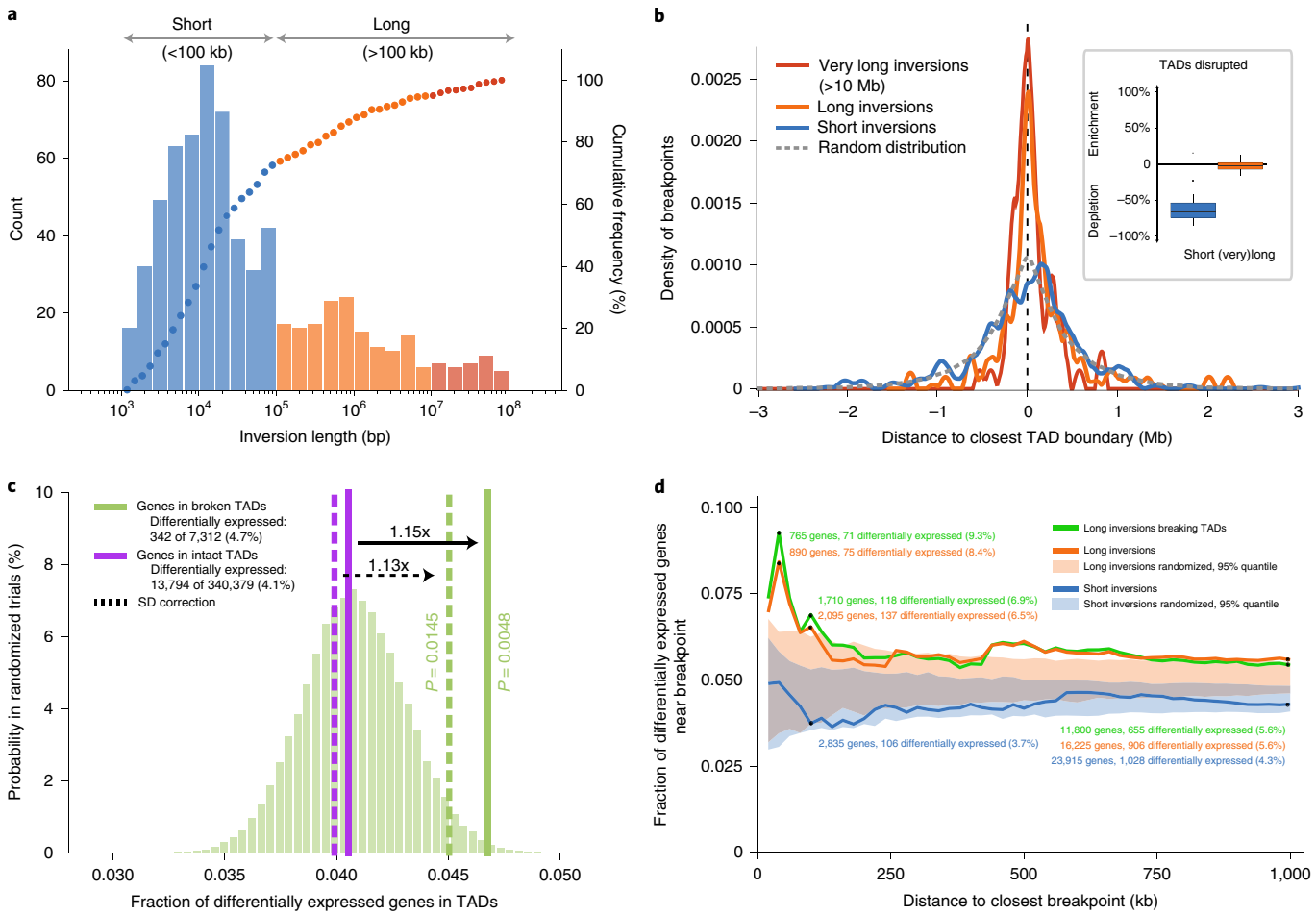


Fig. 5 | Impact of copy-neutral inversions on genome topology and differential gene expression. a, Length distribution of all 388 nonredundant simple inversions, classified as 'short' (<100 kb; blue) or 'long' (>100 kb, orange). The histogram illustrates the absolute counts of binned inversion lengths; the overlaid dots represent the cumulative frequency of inversions corresponding to each bin. **b**, Distance of each inversion breakpoint (centered at 0) to the closest TAD boundary, stratified by inversion length (color coding according to **a**). The expected distance distribution for randomly placed breakpoints is indicated by the gray dashed line. The inset displays the proportion of inversions (stratified by length) that disrupt TADs (median short: −67.1%, median long: −2.4%). Percentage 'enrichment' or 'depletion' is shown as the ratio of observed over expected disruptions calculated after randomizing the inversion locations (Methods). **c**, Proportion of differentially expressed genes in TADs classified as either 'broken' (solid green line) or 'intact' (solid purple line). The underlying histogram depicts the expected differentially expressed frequency after randomizing TAD labels. The dotted lines represent the differentially expressed proportion after excluding genes in SDs. One-sided permutation testing was used to derive the *P* values (Methods). **d**, Proportion of differentially expressed genes relative to inversion breakpoints and stratified by inversion length or whether the inversion disrupted a TAD. The shaded areas show the expected differentially expressed proportion measured in matched randomized breakpoints.

chromosome is more tightly distributed than autosomes with an upper limit of approximately 100 kb (Supplementary Fig. 10). This size constraint may be the consequence of the relatively unique SD organization on the X chromosome, where closely distributed pairwise SDs provide the substrates for NAHR as opposed to autosomes where recent duplications are more interspersed[51]. Alternatively, selective effects on sex chromosomes may be playing a role[52,53], eliminating such events in males.

Within a chromosome, there is also clear regional clustering and we identified 23 discrete regions where we observed an excess of ape inversions. These inversion breakpoint clusters are enriched approximately sixfold for the presence of SDs (Fig. 3b, inset), with regions on chromosomes 16, 17 and X showing some of the largest intervals (Supplementary Fig. 22). Interestingly, chromosomes 16 and X are particularly biased for female recombination where genetic estimates suggest a tenfold reduction in male recombination[54] (Supplementary Fig. 17b). Targeted sequencing of large-insert BAC clones from orangutan, chimpanzee and human confirm an excess of fixed and inverted polymorphisms, with breakpoints mapping to these SDs[55]. Related to this feature, we also observed 27 shared inversions among the different ape species, which suggests either recurrent inversions or incomplete lineage sorting during evolution (Fig. 2a) (ref. [56]). Several lines favored recurrent hotspots of mutation—85% (23 out of 27) of these hotspots, for example, were flanked by SDs that would promote recurrent mutation by NAHR. We found that inversions flanked by SDs are much more likely to be polymorphic when compared to ape inversions not flanked by SDs. When we analyzed 150 validated human inversion polymorphisms separately[4], we found that 33% (49 out of 150) overlap those detected in NHPs with 77% flanked by SDs and many mapping to the predicted 23 inversion breakpoint clusters. Once again, the X chromosome is disproportionately enriched, carrying more than a third of these likely recurrent sites (11 out of 27).

Phasing of individual human and NHP haplotypes reveals a remarkable pattern of inversion toggling extending previous observations of individual loci[8,50,57] to entire chromosomal regions

(Fig. 3d). One of these inversion hotspots on the X chromosome, for example, corresponds to the previously described FLNA-EMD inversion, which has been estimated to have undergone at least 10 independent inversion events based on a comparative sequencing study of 27 eutherian mammals[50]. The dynamics of recombination, linkage disequilibrium and allele frequency of such ancient evolutionary polymorphisms can now be evaluated more systematically and is especially interesting in light of the fact that these inversion hotspots frequently contain protein-coding genes (for example, *MAGEA11*, *H2BFWT* and *H2BFM*) (Fig. 3d).

The association between SDs, inversion polymorphisms, and microdeletion and microduplication syndromes is long-standing[2,4,8,32,33]. Recurrent inversions on the X chromosome have also been associated with factor VIII deficiency observed in both humans and dogs, which appears to be mediated by inverted repeats arising independently or homogenized by conversion at the same regions[57]. In a few cases where the underlying mechanism has been investigated[58–61], individuals carrying inverted haplotypes appear predisposed to higher rates of NAHR either because of SDs evolved in the flanking regions in direction orientation or because SDs become configured to predispose to interchromosomal rearrangement in the heterozygous state[33,60]. Related to this, one of the important findings of this study is the ability to distinguish simple inversions from inverted duplications by comparing SD and Strand-seq datasets[31]. In so doing, we determined that the preferred (75%) orientation for emergence of lineage-specific duplications is in the inverted orientation as opposed to the direct one. While this is selectively advantageous in the short term to reduce NAHR-mediated copy number changes, inverted duplications set the stage for cascading and recurrent inversion toggling, which leads to simple inversions and ultimately more complex SDs predisposing to recurrent rearrangement and neurodevelopmental disease.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-020-0646-x.

Received: 10 September 2019; Accepted: 15 May 2020;
Published online: 15 June 2020

References

1. Sturtevant, A. H. Genetic factors affecting the strength of linkage in Drosophila. *Proc. Natl Acad. Sci. USA* **3**, 555–558 (1917).
2. Antonacci, F. et al. Characterization of six human disease-associated inversion polymorphisms. *Hum. Mol. Genet.* **18**, 2555–2566 (2009).
3. Chaisson, M. J. P., Wilson, R. K. & Eichler, E. E. Genetic variation and the de novo assembly of human genomes. *Nat. Rev. Genet.* **16**, 627–640 (2015).
4. Chaisson, M. J. P. et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1784 (2019).
5. Kidd, J. M. et al. Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64 (2008).
6. Kidd, J. M. et al. A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell* **143**, 837–847 (2010).
7. Sanders, A. D. et al. Characterizing polymorphic inversions in human genomes by single-cell sequencing. *Genome Res.* **26**, 1575–1587 (2016).
8. Zody, M. C. et al. Evolutionary toggling of the MAPT 17q21.31 inversion region. *Nat. Genet.* **40**, 1076–1083 (2008).
9. 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).
10. Giner-Delgado, C. et al. Evolutionary and functional impact of common polymorphic inversions in the human genome. *Nat. Commun.* **10**, 4222 (2019).
11. Tuzun, E. et al. Fine-scale structural variation of the human genome. *Nat. Genet.* **37**, 727–732 (2005).
12. Kronenberg, Z. N. et al. High-resolution comparative analysis of great ape genomes. *Science* **360**, eaar6343 (2018).
13. Yunis, J. & Prakash, O. The origin of man: a chromosomal pictorial legacy. *Science* **215**, 1525–1530 (1982).
14. Kehrer-Sawatzki, H., Sandig, C. A., Goidts, V. & Hameister, H. Breakpoint analysis of the pericentric inversion between chimpanzee chromosome 10 and the homologous chromosome 12 in humans. *Cytogenet. Genome Res.* **108**, 91–97 (2005).
15. Kehrer-Sawatzki, H. et al. Breakpoint analysis of the pericentric inversion distinguishing human chromosome 4 from the homologous chromosome in the chimpanzee (*Pan troglodytes*). *Hum. Mutat.* **25**, 45–55 (2005).
16. Ventura, M. et al. The evolution of African great ape subtelomeric heterochromatin and the fusion of human chromosome 2. *Genome Res.* **22**, 1036–1049 (2012).
17. Lucas Lledó, J. I. & Cáceres, M. On the power and the systematic biases of the detection of chromosomal inversions by paired-end genome sequencing. *PLoS ONE* **8**, e61292 (2013).
18. Catacchio, C. R. et al. Inversion variants in human and primate genomes. *Genome Res.* **28**, 910–920 (2018).
19. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* **12**, 363–376 (2011).
20. Rasekh, M. E. et al. Discovery of large genomic inversions using long range information. *BMC Genomics* **18**, 65 (2017).
21. Feuk, L. et al. Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. *PLoS Genet.* **1**, e56 (2005).
22. Falconer, E. et al. DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. *Nat. Methods* **9**, 1107–1112 (2012).
23. Sanders, A. D., Falconer, E., Hills, M., Spierings, D. C. J. & Lansdorp, P. M. Single-cell template strand sequencing by Strand-seq enables the characterization of individual homologs. *Nat. Protoc.* **12**, 1151–1176 (2017).
24. Porubsky, D. et al. breakpointR: an R/Bioconductor package to localize strand state changes in Strand-seq data. *Bioinformatics* **36**, 1260–1261 (2020).
25. Szamalek, J. M. et al. The chimpanzee-specific pericentric inversions that distinguish humans and chimpanzees have identical breakpoints in *Pan troglodytes* and *Pan paniscus*. *Genomics* **87**, 39–45 (2006).
26. Sulovari, A. et al. Human-specific tandem repeat expansion and differential gene expression during primate evolution. *Proc. Natl Acad. Sci. USA* **116**, 23243–23253 (2019).
27. Newman, T. L. et al. A genome-wide survey of structural variation between human and chimpanzee. *Genome Res.* **15**, 1344–1356 (2005).
28. Shao, H. et al. npInv: accurate detection and genotyping of inversions using long read sub-alignment. *BMC Bioinformatics* **19**, 261 (2018).
29. Rausch, T. et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
30. Cheng, Z. et al. A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* **437**, 88–93 (2005).
31. Sudmant, P. H. et al. Evolution and diversity of copy number variation in the great ape lineage. *Genome Res.* **23**, 1373–1382 (2013).
32. Osborne, L. R. et al. A 1.5 million-base pair inversion polymorphism in families with Williams–Beuren syndrome. *Nat. Genet.* **29**, 321–325 (2001).
33. Giglio, S. et al. Olfactory receptor–gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements. *Am. J. Hum. Genet.* **68**, 874–883 (2001).
34. Coe, B. P. et al. Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nat. Genet.* **46**, 1063–1071 (2014).
35. Vollger, M. R. et al. Long-read sequence and assembly of segmental duplications. *Nat. Methods* **16**, 88–94 (2019).
36. Marques-Bonet, T. et al. A burst of segmental duplications in the genome of the African great ape ancestor. *Nature* **457**, 877–881 (2009).
37. Ventura, M. et al. Gorilla genome structural variation reveals evolutionary parallelisms with chimpanzee. *Genome Res.* **21**, 1640–1649 (2011).
38. Spielmann, M., Lupiáñez, D. G. & Mundlos, S. Structural variation in the 3D genome. *Nat. Rev. Genet.* **19**, 453–467 (2018).
39. Lupiáñez, D. G. et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012–1025 (2015).
40. Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
41. Brawand, D. et al. The evolution of gene expression levels in mammalian organs. *Nature* **478**, 343–348 (2011).
42. Sousa, A. M. M. et al. Molecular and cellular reorganization of neural circuits in the human lineage. *Science* **358**, 1027–1032 (2017).
43. Pollen, A. A. et al. Establishing cerebral organoids as models of human-specific brain evolution. *Cell* **176**, 743–756.e17 (2019).
44. Kanton, S. et al. Organoid single-cell genomic atlas uncovers human-specific features of brain development. *Nature* **574**, 418–422 (2019).

45. Ghavi-Helm, Y. et al. Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. *Nat. Genet.* **51**, 1272–1282 (2019).
46. Hey, J. Speciation and inversions: chimps and humans. *Bioessays* **25**, 825–828 (2003).
47. Navarro, A. & Barton, N. H. Chromosomal speciation and molecular divergence—accelerated evolution in rearranged chromosomes. *Science* **300**, 321–324 (2003).
48. Fishilevich, S. et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database (Oxford)* **2017**, bax028 (2017).
49. Sohoni, S. et al. Elevated heme synthesis and uptake underpin intensified oxidative metabolism and tumorigenic functions in non-small cell lung cancer cells. *Cancer Res.* **79**, 2511–2525 (2019).
50. Cáceres, M., Sullivan, R. T. & Thomas, J. W. A recurrent inversion on the eutherian X chromosome. *Proc. Natl Acad. Sci. USA* **104**, 18571–18576 (2007).
51. Bailey, J. A. et al. Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).
52. Corbett-Detig, R. B. & Hartl, D. L. Population genomics of inversion polymorphisms in *Drosophila melanogaster*. *PLoS Genet.* **8**, e1003056 (2012).
53. Natri, H. M., Merilä, J. & Shikano, T. The evolution of sex determination associated with a chromosomal inversion. *Nat. Commun.* **10**, 145 (2019).
54. Kong, A. et al. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467**, 1099–1103 (2010).
55. Nuttle, X. et al. Emergence of a *Homo sapiens*-specific gene family and chromosome 16p11.2 CNV susceptibility. *Nature* **536**, 205–209 (2016).
56. Fuller, Z. L., Leonard, C. J., Young, R. E., Schaeffer, S. W. & Phadnis, N. Ancestral polymorphisms explain the role of chromosomal inversions in speciation. *PLoS Genet.* **14**, e1007526 (2018).
57. Lozier, J. N. et al. The Chapel Hill hemophilia A dog colony exhibits a factor VIII gene inversion. *Proc. Natl Acad. Sci. USA* **99**, 12991–12996 (2002).
58. Itsara, A. et al. Population analysis of large copy number variants and hotspots of human genetic disease. *Am. J. Hum. Genet.* **84**, 148–161 (2009).
59. Antonacci, F. et al. Palindromic *GOLGA8* core duplicons promote chromosome 15q13.3 microdeletion and evolutionary instability. *Nat. Genet.* **46**, 1293–1302 (2014).
60. Mohajeri, K. et al. Interchromosomal core duplicons drive both evolutionary instability and disease susceptibility of the Chromosome 8p23.1 region. *Genome Res.* **26**, 1453–1467 (2016).
61. Maggiolini, F. A. M. et al. Genomic inversions and *GOLGA* core duplicons underlie disease instability at the 15q25 locus. *PLoS Genet.* **15**, e1008075 (2019).
62. 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).
63. 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).

Methods

Strand-seq library preparation and sequencing. Strand-seq libraries were prepared from B-cell lymphoblastic cell lines previously generated for a female western chimpanzee *(Pan troglodytes verus*; Dorien), female bonobo (*Pan paniscus*; Ulindi), male western gorilla (*Gorilla gorilla*; GGO 9) and male Sumatran orangutan (*Pongo abelii*; PPY-10). All lines were maintained in Roswell Park Memorial Institute 1640 with 10% FCS, 1% GlutaMAX and 1% penicillin/streptomycin. Bromodeoxyuridine (catalog no. B5002; Sigma-Aldrich) was added to log phase cell cultures at 40 or 100 µM concentrations for a period of 18 or 24 h. Single nuclei were prepared and sorted using the FACSMelody cell sorter (BD Biosciences) into 96-well plates for Strand-seq library production, as described previously[22,23]. The Strand-seq protocol was implemented on a Biomek FX^P liquid-handling robotic system; pooled single-cell libraries were sequenced on the NextSeq 500 platform (MID-mode, 75 bp paired-end protocol; Illumina). After demultiplexing, Strand-seq reads were aligned to the human reference assembly GRCh38 (GCA_000001405.15_GRCh38_no_alt_analysis_set.fna) using the default parameters of the Burrows–Wheeler Aligner-MEM v.0.7.15-r1140. Aligned BAM files were sorted by genomic position using SAMtools v.1.7 and duplicated reads marked using sambamba v.0.6.6. After alignment, each single library was evaluated to select only high-quality Strand-seq data for downstream analyses. Specifically, libraries with visible background reads (that is, reads mapped to the opposite direction on chromosomes that inherited template strands with the same directionality) and libraries with low (<50,000 reads) or uneven coverage were excluded, as detailed previously[23,64].

Inversion detection from Strand-seq data. To increase the sensitivity of inversion calling and inversion breakpoint resolution, we constructed composite files for each individual great ape genome. As described previously[4,7], composite files were generated by merging Strand-seq data for each chromosome based on shared strand inheritance patterns to produce a high-coverage directional file for the genome. Briefly, we concatenated reads from multiple Strand-seq libraries from each chromosomal region genotyped as either Watson-Watson (inverted orientation) or Crick-Crick (reference orientation) state. To match the reference orientation, we reverse-complemented regions genotyped as Watson-Watson before merging subsequent Strand-seq libraries. From a composite file, read directionality could then be assigned as either 'reference' and in the same (forward) orientation as the reference assembly or 'inverted' and in the opposite (reverse) orientation of the assembly. The ape-specific composite files produced in this study are available as University of California Santa Cruz (UCSC)-formatted BED files.

In each composite file, we then called inversions using the Bioconductor package breakpointR devel v.0.99.10 (ref. [24]). We used the runBreakpointr function with the following parameters: windowsize = 10000; binMethod = "multi"; background = 0.1; peakTh = 0.25; trim = 10; zlim = 3.291; minReads = 20; min.mapq = 10. The inversion breakpoint resolution highly depends on the underlying genome architecture on each side of the inversion. Because short Strand-seq reads have difficulty mapping within SDs flanking the inversion, the breakpoint is typically placed within the SD range. In such regions, breakpoint prediction is less accurate and thus might not represent the exact breakpoint position. Details of the breakpoint resolution achieved by breakpointR have been discussed previously[24].

We curated every inversion breakpoint detected by breakpointR manually in the UCSC Genome Browser[65] and classified events as simple inversion calls without evidence for increased copy number (INV) and more complex inversions with increase in copy (whole-genome shotgun detection)[66] as inverted duplication (invDup). We assigned each inversion a genotype as either homozygous, where the vast majority of reads mapped in inverted orientation (Watson, minus strand) or heterozygous, where there is an approximately 1:1 ratio between reads in inverted (Watson, minus strand) and reference (Crick, plus strand) orientation.

Smaller changes in directionality not detected by breakpointR were included into the final callset only if they were supported either by the PacBio or Illumina callset. Because of limited coverage of Strand-seq data, we excluded simple inversions with <1 kb of unique sequence and inverted duplications smaller than 10 kb to ensure the high quality of our callset.

Long-read alignment parameters. Raw PacBio and isoform sequencing (Iso-Seq) reads were obtained from previous studies[12] (Supplementary Table 11). PacBio reads were aligned to GRCh38 using minimap2 v.2.14-r883. using the recommended minimap2 parameters by the PBSV pipeline (--MD -t 8 -x map-pb -a --eqx -L -O 5,56 -E 4,1 -B 5 --secondary=no -z 400,50 -r 2k -Y). Iso-Seq reads were mapped to GRCh38 with minimap2 using the following parameters: -ax splice -uf -C5 --secondary=no –eqx.

Inversion validations. Strand-seq inversion callsets were validated using multiple orthogonal datasets, such as PacBio, Illumina and Bionano optical maps (Bionano Genomics; Supplementary Table 1). We called inversions in PacBio data using PBSV v.2.0.2 and Sniffles v.1.0.10 with default parameters. We used DELLY v.0.7.9 to call inversions in short Illumina reads. Bionano inversion calls were obtained using an analysis pipeline provided by the vendor (Supplementary Note). Furthermore, we used previously published and validated NHP inversions[12,18,25]. We used the primatR v.0.1.0 package function getReciprocalOverlaps to find the best-matching inversion call from any of the orthogonal dataset for each Strand-seq inversion. Strand-seq inversions having ≥50% reciprocal overlap with any orthogonal dataset were deemed as validated. We attempted to validate the remaining unvalidated inversions by manual inspection of Bionano alignments and by projecting NHP de novo assemblies[12] (Supplementary Table 12) against GRCh38 using dotplot analysis. In the case of bonobo, we used long-read data generated from the Mhudiblu cell line examining local assemblies of the inversion breakpoints. Lastly, we attempted to validate a selected number of inversions using FISH (Supplementary Note).

Phylogenetic analyses. To identify human-specific inversions and eliminate reference artifacts, we repeated the Strand-seq analysis with data generated for the Yoruban individual NA19240 (ref. [4]) using the same parameters. Human-specific inversions were defined as regions that were homozygously inverted in all NHPs with regard to a human reference (homozygous reference orientation) and confirmed with NA19240. We also removed all previously reported misassemblies in the human reference (Supplementary Table 13) (ref. [4]). We used a 50% reciprocal overlap to delineate shared and lineage-specific inversions among great apes. We constructed a simple matrix where individuals (rows) who share any given loci (columns) based on 50% reciprocal overlap are assigned a value of 1; otherwise they are assigned a value of 0. Next, we computed the Hamming distance between all great apes, which was then used by hierarchical clustering to reconstruct the phylogeny purely based on the presence or absence of shared loci. In this analysis, we did not take heterozygosity into account.

Estimating inversion rates in the great ape lineage. We computed three different rate estimates: the mean fixation rate of simple inversions per million years; branch rate estimates; and the rate per inverted base per single-nucleotide substitution.

The mean fixation rate of simple inversions per million years assumes a clockwise inversion rate across the great ape phylogeny; thus, it is defined as the total number of simple species-specific inversions divided by the sum of divergence times (in million years) among species. To infer the branch-specific rates and the phylogenetic relationships among primates of interest using the 358 autosomal inversion calls, we performed the Lewis–Markov k model[67] implemented in Bayesian phylogeny-based (BEAST v.2.5.0) analyses. We modeled the evolution of individual inversions as changes in separate discrete traits, where each trait has three states: homozygous human reference; heterozygous inverted; and homozygous inverted orientations. To run BEAST, we used Lewis–Markov k model, with GAMMA Category Count = 3 for the site model and a random local clock for the clock model parameter to explicitly test mutation rate on individual branches in the tree. For tree priors, we used the birth–death model with default parameters but added a prior for the calibration of human–gorilla divergence using a log-normal distribution ($M = 2.1$, $S = 0.085$). We performed five independent runs to infer the phylogeny using a chain length of 10,000,000 samples and recorded every 1,000 samples. We used the accompanying program Tracer v.1.7.1 to determine the quality of each run and used the first 10% as burn-in. All phylogenetic trees were plotted using FigTree v.1.4.3 and DensiTree v.2.2.6.

Finally, to estimate the rate of simple inversions relative to that of single-nucleotide variants (SNVs) on each branch of the inferred phylogeny as proposed by Sudmant et al.[31], we computed the rate of inverted bases per substitution for each branch with the following formula: number of inverted bases per substitution of a branch = (number of total inverted bases on the branch/2.87×10^9)/the substitution rate of inversion, where 2.87×10^9 is the genome size after excluding simple repeats and the rate of inversion is estimated by BEAST as listed in Supplementary Table 4.

Inverted duplication analysis. Besides inverted duplications, we listed the number of direct duplications in the NHP genomes. We did this by scanning Strand-seq composite files in the UCSC Genome Browser and reporting regions of increased read depth (based on the WSSD track) and reads mapped preferentially in the reference orientation (Supplementary Table 7). Furthermore, we genotyped all lineage-specific duplications detected previously[31]. Of all 11,260 lineage-specific duplications, we retained only regions ≥10 kb that did not appear in humans. Note that a strength of Strand-seq is that it distinguishes the directionality of intrachromosomal duplications in most cases, including clustered duplications. In such cases, seeing a mixture of direct and inverted reads mapping over the duplicated loci is evidence that at least one copy of this locus is in the inverted orientation (Fig. 1b). However, the directionality of interchromosomal duplications is more difficult to assess reliably using Strand-seq. Because the strand state of chromosomes where a corresponding duplication resides might differ within a single Strand-seq library based on assortment, read directionality of these duplicated copies will reflect the strand state of the chromosome they reside in. To avoid evaluating low-confidence inverted duplication sites, we removed regions that overlapped with interchromosomal links predicted by the PacBio split-read mappings. This left us with 504 nonredundant regions. Of these, 1 region failed to lift from GRCh36 to GRCh38 coordinates. Next, we genotyped these regions as heterozygous, homozygous inverted or homozygous reference (using the primatR genotypeRegions function with min.reads = 5, alpha = 0.05). Lastly, we calculated the proportions between inverted and direct duplications for each NHP and

established the significance of this difference using a chi-squared test. The resulting *P* values were Bonferroni-corrected for multiple testing. The same significance tests were repeated with the 387 inverted and 88 direct duplications reported in this study (Supplementary Fig. 25c).

We attempted to validate inverted duplications using discordantly mapped BAC-end sequences with signatures of an inversion, deletion or insertion. Inverted duplications that overlapped with at least 5% of discordantly mapped BAC ends and at least two discordantly mapped BAC ends in total were marked as supported by BAC-end mapping. In addition, we attempted to validate inverted duplications using de novo assembly and segmental duplication assembly[35] for each NHP. We aligned assembled contigs against the human reference using minimap2 v.2.17 to obtain the genomic locations where given contigs mapped. Next, we used nucmer v.3.1 to align all contigs against specific loci in the human reference with the following parameters: --mumreference -c 100 -g 1000 -l 5. Such alignments were visualized as dotplots; regions showing clear inversion patterns were marked as supported by de novo assembly (Supplementary Table 7).

Mapping inverted duplication loci. To identify the putative integration sites of lineage-specific duplications, we constructed a pseudo-mate-pair read using PacBio reads that extended over the inverted duplication breakpoints. Specifically, we split individual long PacBio reads using a *k*-mer size of 2 kb and a step size of 1 kb. For instance, a 12-kb PacBio read was cut such that we initially created a 2-kb portion on the left and left the rest of the PacBio read (10 kb) on the right. Then, we moved the cut site by 1 kb to the right, creating the left portion of the PacBio read of 4 kb and leaving the remaining 8-kb portion on the right. We iterated this procedure until the left mate read equalled 2 kb (Supplementary Fig. 27a). The resulting pseudo-mate-pairs were mapped to the human reference genome (GRCh38) in a paired-end fashion using the Burrows–Wheeler Aligner-MEM v.0.7.15-r1140 with the '-x pacbio' parameter. Discordant read pairs that mapped to different chromosomal locations pointed to the sites where duplicated sequences integrated in the genome. We required a minimum of ten unique PacBio reads to support such interchromosomal connections.

Human inversion callset and overlap. We compared the NHP inversion data to a set of 150 human polymorphic inversions identified and phased from three 1000 Genomes Project trios of Han Chinese, Puerto Rican and Yoruba Ibadan ancestry[4]. To detect inverted loci shared between NHPs and humans, we constructed a nonredundant dataset of NHP simple inversions. The set of human polymorphic inversions[4] was filtered for events with ≥1 kb of unique sequences. We detected shared inversions between the Human Genome Structural Variation Consortium (HGSVC) and NHP callsets based on 50% reciprocal overlap. Next, we regenotyped shared NHP inversions based on Strand-seq composite files and reported the inverted loci frequency for both HGSVC and NHP individuals based on the number of inverted loci (homozygous, 2 inverted loci; heterozygous, 1 inverted locus; and reference, 0 inverted loci). To see how many of these regions were flanked by known human SDs, we downloaded a UCSC Genome Browser track of known human SDs and calculated the distance of each inversion breakpoint to the closest SD. We set inversions where both breakpoints were no further than 5 kb away from the closest SD as being flanked by SDs.

Overlap between simple inversions and pathogenic CNVs. The list of human pathogenic CNVs was obtained from a previous study that identified regions showing an excess of large deletions and duplications in cases of pediatric developmental delay when compared to normal population controls[34]. We searched for 50% reciprocal overlap between pathogenic CNVs ($n = 36$) and simple inversions ($n = 682$) using the primatR getReciprocalOverlaps function. Those pathogenic CNVs that overlapped with simple inversions were regenotyped (primatR genotypeRegions function) in all NHPs to compute the frequency of inverted loci in these regions (homozygous, 2 inverted loci; heterozygous, 1 inverted locus; and reference, 0 inverted loci). To estimate the level of enrichment of pathogenic CNVs in NHP simple inversions, we randomly shuffled these pathogenic CNVs 100 times and evaluated the 50% reciprocal overlap each time. This randomization was performed using the primatR randomizeRanges function. Each pathogenic CNV was shuffled within its chromosome of origin and we excluded assembly gaps and centromeres from the randomization process.

Assigning human and NHP inversion to haplotypes. To assign all inversions (homozygous and heterozygous) to their corresponding haplotypes, we used the phasing information embedded in the Strand-seq data[64]. We used the RTG tool[68] (RTG Core Non-Commercial v.3.9.1) to call SNVs in the Strand-seq data merged in a single BAM file. We used the following RTG parameters: --min-mapq 10 --min-base-quality 10 --snps-only --no-calibration --machine-errors illumina --max-coverage 30. After obtaining the set of heterozygous SNVs, we used StrandPhaseR v.1.0.0 to phase single-cell haplotypes and split all Strand-seq reads into their respective haplotypes[69]. Next, we used the read depth profile of haplotype-specific reads to assign the inverted and reference alleles, in heterozygous conformation, into their respective haplotypes (Supplementary Fig. 19). We visualized the order and orientation of inverted regions using the Comprehensive R Archive Network package gggenes v.0.4.0 (https://cran.r-project.org/web/packages/gggenes/).

Fusion gene detection. To detect putative fusion genes, we used a tool called cDNA_Cupcake[70] (https://github.com/Magdoll/cDNA_Cupcake/) and its function fusion_finder.py to perform fusion gene prediction based on the recommended settings at https://github.com/Magdoll/cDNA_Cupcake/wiki/. To remove excess false positive calls, we narrowed down the gene fusions predicted initially to only those that were in the vicinity (±1 kb) of predicted simple inversion and inverted duplication breakpoints. We further investigated the split-read mapping signatures of Iso-Seq (FLNC) reads that mapped to different chromosomes of the human reference genome. To reduce the level of false positive calls, we further removed fusion predictions that did not overlap with known genes from the GENCODE database v.29 at both donor and acceptor sites, as well as sites that overlapped with known SD regions on either donor or acceptor sites. Lastly, we attempted to validate these fusion gene predictions based on the PacBio split-read mapping described in 'Mapping inverted duplication loci'.

Defining inverted breakpoint clusters. To detect regions of clustered inversion breakpoints, we merged together nonredundant HGSVC and NHP inversion callsets. We extracted inversion breakpoints for each inversion and submitted a sorted list of inversion breakpoints to the primatR hotspotter function[24] (parameters: bw = 2,000,000, pval = 5×10^{-10}). This function searches for regions of increased density of inversion breakpoints around the genome by using the density function to perform a kernel density estimation. A *P* value was calculated by comparing the density profile of the genomic events with the density profile of a randomly subsampled set of genomic events (bootstrapping).

Analysis of TAD-disrupting inversions. A set of human-specific TAD boundaries was obtained from the study by Dixon et al.[40]. The coordinates of these boundaries were translated into the GRCh38 reference assembly using the liftOver tool available from the UCSC Genome Browser. All but one TAD were successfully mapped to the new reference genome (GRCh38). We measured the distance of TAD boundaries to the breakpoints of simple inversions (nonredundant set, $n = 388$) separately for various inversion sizes (<100 kb, >100 kb, <10 Mb, >10 Mb). (We excluded inverted duplications since we did not want copy number changes to affect the differential expression analyses.) The distribution of distances to the closest TAD boundaries for each inversion size category was drawn as a kernel density estimation-fitted curve. TADs were further marked as 'disrupted' in a scenario when only one breakpoint of a given inversion was positioned within the TAD (Supplementary Fig. 28), otherwise the TAD was classified as 'intact'. The rates of disrupted TADs for different inversion size categories were examined as follows: the number of disrupted TADs per inversion category was counted and compared to values after the inversion positions were randomized within each chromosome (excluding gaps and centromeric regions, and preserving inversion lengths and their relative distances) 100 times using the regioneR[63] v.1.16.2 circularRandomizeRegions function. This resulted in an estimate for the fold enrichment of broken TADs compared to randomly expected levels.

We further reported genes whose differential expression was probably caused by an inversion that disrupted the predicted gene–enhancer interaction. A gene2enhancer interaction was considered disturbed if one but not both inversion breakpoints fell between a gene and its associated enhancer. For this analysis, we used gene–enhancer interactions obtained from the GeneHancer v.4.8 (ref. [48]) track from the UCSC Genome Browser. Only so-called 'double elite' gene–enhancer interactions derived from more than one experimental or computational method have been considered.

Differential gene expression analysis. Our differential expression considered 16,524 1:1:1:1:1 orthologs provided by ENSEMBL v.91. We excluded genes that were not expressed consistently across all samples (fragments per kilobase million <1 across all samples and tissues). We also excluded a list of 91 genes escaping X inactivation (obtained from Tukiainen et al.[71] due to expected sex-specific expression bias), which left us with 15,117 genes. The level of differential expression per gene was calculated using DESeq2 (ref. [72]) v.1.24.0, with information about sex included as a cofactor. Gene-wise read counts derived from RNA-seq data were obtained from Brawand et al.[41]. All NHPs were tested separately against human, resulting in a list of differentially expressed genes for each species. We consistently performed between-species differential expression analyses for matched tissues (for example, human brain versus chimpanzee brain, human brain versus bonobo brain, human kidney versus orangutan kidney). There were no data available for orangutan testis; accordingly, we performed 23 differential expression comparisons overall (4 species × 6 tissues, minus orangutan testis). Genes with an absolute shrunken fold change >2 and an adjusted Shannon information value (also known as 'surprisal (s) value', a standard feature of DESeq2) below 0.005 were considered as differentially expressed. Supplementary Fig. 29 depicts differential expression in the brain as an example. Overall differential expression levels per ape genome were consistent with NHP phylogeny and species divergence (Supplementary Fig. 29b).

Differential expression in broken versus intact TADs. Genes were assigned to two groups based on whether or not they fell into a broken TAD (mediated by a balanced inversion); the ratio of differentially expressed genes over total genes was calculated for each group separately. All genes were counted once for each tissue and species, resulting in $15{,}117 \times 23 = 347{,}691$ tests. A permutation test was used to test for statistical significance of the enrichment of differentially expressed genes in broken TADs. In 50,000 repetitions, genes were randomly assigned to the two groups (preserving the number of genes in both) and differential expression ratios were calculated after each permutation. *P* values were derived from the percentile of the observed versus randomized differential expression ratio. The distances of differentially expressed genes to the closest inversion breakpoint were obtained across all genes and for all 23 differential expression comparisons; randomization was pursued by shuffling each inversion randomly on the chromosome that the inversion had been observed in (shuffling was pursued 1,000 times).

External datasets. The following external datasets were used: set of TADs in human[40]; bulk RNA-seq data for all NHPs[41]; set of X inactivation escape genes[71]; brain organoids sequencing data obtained from the Gene Expression Omnibus under accession no. GSE124299 and database of Genotypes and Phenotypes under accession no. phs000989.v3.p1 (ref. [43]); raw Strand-seq data for NA19240, which can be accessed at the European Nucleotide Archive under accession no. PRJEB12849; raw 10X Genomics data available at the National Center for Biotechnology Information under BioProject no. PRJNA593056).

Reporting Summary. Further information on research design is available in the Nature Research Life Sciences Reporting Summary linked to this article.

Data availability

Strand-seq data aligned to GRCh38 and ape-specific composite files are available at zenodo, (https://doi.org/10.5281/zenodo.3818043); the PacBio and Bionano datasets are reported in Supplementary Tables 11 and 14; Supplementary data are available at GitHub (https://github.com/daewoooo/ApeInversion_paper); the PacBio and Bionano inversion callset are available at GitHub (https://github.com/daewoooo/ApeInversion_paper/tree/master/Supplementary_datasets).

Code availability

The primatR package is available at GitHub (https://github.com/daewoooo/primatR); the breakpointR package is available at GitHub (https://github.com/daewoooo/breakpointR) (devel branch); custom scripts are available at GitHub (https://github.com/daewoooo/ApeInversion_paper/tree/master/Custom_scripts); software releases at the publication date are available at Zenodo (https://doi.org/10.5281/zenodo.3556774).

References

64. Porubský, D. et al. Direct chromosome-length haplotyping by single-cell sequencing. *Genome Res.* **26**, 1565–1574 (2016).
65. Kent, W. J. et al. The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
66. Sudmant, P. H. et al. Diversity of human copy number variation and multicopy genes. *Science* **330**, 641–646 (2010).
67. Lewis, P. O. A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst. Biol.* **50**, 913–925 (2001).
68. Cleary, J. G. et al. Joint variant and de novo mutation identification on pedigrees from high-throughput sequencing data. *J. Comput. Biol.* **21**, 405–419 (2014).
69. Porubsky, D. et al. Dense and accurate whole-chromosome haplotyping of individual genomes. *Nat. Commun.* **8**, 1293 (2017).
70. Weirather, J. L. et al. Characterization of fusion genes and the significantly expressed fusion isoforms in breast cancer by hybrid sequencing. *Nucleic Acids Res.* **43**, e116 (2015).
71. Tukiainen, T. et al. Landscape of X chromosome inactivation across human tissues. *Nature* **550**, 244–248 (2017).
72. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

Acknowledgements

We thank T. Brown for assistance in editing this manuscript. In addition, we thank S. Pääbo for generously providing the bonobo (Ulindi) and chimpanzee (Dorien) cell lines used in this study, along with H. Kaessmann and E. Leushkin for access to the ape RNA-seq data. We acknowledge the technical assistance provided by A. Pang and A. Hastie, who provided the Bionano inversion calls for NHPs. We also thank the European Molecular Biology Laboratory Genomics Core facility, particularly V. Benes and J. Zimmermann, for assistance with automating the Strand-seq library generation. This work was supported, in part, by grants from the National Institutes of Health (NIH; grant nos. HG002385 and HG010169 to E.E.E.). A.D.S. was supported by an Alexander von Humboldt Foundation Research Fellowship. P.H. was supported by the NIH Pathway to Independence Award (National Human Genome Research Institute, no. K99HG011041). A.S. was supported by the NIH Genome Training Grant (T32, no. HG000035-23). J.O.K. was supported by a European Research Council Consolidator grant (no. 773026). S.C. was supported by a National Health and Medical Research Council CJ Martin Biomedical Fellowship (no. 1073726). E.E.E. is an investigator of the Howard Hughes Medical Institute.

Author contributions

D.P., A.D.S. and E.E.E. designed the study, analyzed and interpreted the data, produced the figures and wrote the manuscript. A.D.S. and J.O.K. generated the Strand-seq libraries. W.H. analyzed the TADs and differential gene expression. P.H. and A.S. reconstructed the NHP phylogeny and helped with the statistical analysis. R.L., M.S., S.C., L.M., M.V. and F.A. provided validation of the inversion calls. S.C.M. and D.G. processed the PacBio data. T.M. and A.A.P. supported data analysis and interpretation.

Competing interests

E.E.E. is on the scientific advisory board of DNAnexus.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-020-0646-x.

Correspondence and requests for materials should be addressed to E.E.E.

Reprints and permissions information is available at www.nature.com/reprints.