Published in final edited form as:
Nature. 2021 June ; 594(7861): 77–81. doi:10.1038/s41586-021-03519-x.

A high-quality bonobo genome refines the analysis of hominid evolution

Yafei Mao[1,#], **Claudia R. Catacchio**[2,#], **LaDeana W. Hillier**[1], **David Porubsky**[1], **Ruiyang Li**[1], **Arvis Sulovari**[1], **Jason D. Fernandes**[3], **Francesco Montinaro**[2,4], **David S. Gordon**[1,5], **Jessica M. Storer**[6], **Marina Haukness**[3], **Ian T. Fiddes**[3], **Shwetha Canchi Murali**[1,5], **Philip C. Dishuck**[1], **PingHsun Hsieh**[1], **William T. Harvey**[1], **Peter A. Audano**[1], **Ludovica Mercuri**[2], **Ilaria Piccolo**[2], **Francesca Antonacci**[2], **Katherine M. Munson**[1], **Alexandra P. Lewis**[1], **Carl Baker**[1], **Jason G. Underwood**[7], **Kendra Hoekzema**[1], **Tzu-Hsueh Huang**[1], **Melanie Sorensen**[1], **Jerilyn A. Walker**[8], **Jinna Hoffman**[9], **Françoise Thibaud-Nissen**[9], **Sofie R. Salama**[3,10], **Andy WC Pang**[11], **Joyce Lee**[11], **Alex R. Hastie**[11], **Benedict Paten**[3], **Mark A. Batzer**[8], **Mark Diekhans**[3], **Mario Ventura**[2,*], **Evan E. Eichler**[1,5,*]

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

[2]Department of Biology, University of Bari, 70125, Bari, Italy.

[3]UC Santa Cruz Genomics Institute, University of California, Santa Cruz, Santa Cruz, CA 95064, USA.

[4]Estonian Biocentre, Institute of Genomics, Tartu, Riia 23b, 51010, Estonia.

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

[6]Institute for Systems Biology, Seattle, WA 98109, USA.

[7]Pacific Biosciences (PacBio) of California, Inc., Menlo Park, CA 94025, USA.

[8]Department of Biological Sciences, Louisiana State University, 202 Life Sciences Building, Baton Rouge, LA 70803 USA.

[9]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 38A, 8600 Rockville Pike, Bethesda, MD 20894, USA.

[*]Corresponding authors: Mario Ventura, Ph.D., Department of Biology, University of Bari, 70125 Bari, Italy, Tel. +39-080-544-3583, mario.ventura@uniba.it; Evan E. Eichler, Ph.D., Department of Genome Sciences, University of Washington School of Medicine, 3720 15th Ave NE, S413A, Seattle, WA 98195-5065, USA, Tel: +1-206-543-9526, eee@gs.washington.edu.
[#]these authors contributed equally to the paper and should be regarded as shared first authors

Author contributions: K.M.M. and A.P.L. generated long-read sequencing data; C.B. created the BAC library (VMRC74); K.H., M.S., P.A.A., D.S.G., L.W.H., S.C.M., M.D., C.R.C., L.M., I.P., M.V., E.E.E. and D.P. completed *de novo* assembly, its curation, and quality assessment; P.C.D., R.L., Y.M., W.H., T.H., D.S.G., M.V. and E.E.E. performed SD and gene family analyses; D.P. performed Strand-seq single-cell data analysis; C.R.C., I.P., L.M., F.A. and M.V. performed FISH analyses; A.W.C.P., J.L. and A.R.H. led the Bionano Genomics analyses; I.T.F., M.D., Y.M., M.H. and B.P. performed gene annotation and gene analyses; J.H. and F.TN. performed RefSeq annotation; J.G.U. performed Iso-Seq; F.M., P.H., and Y.M. performed population genetic analyses; J.D.F., J.M.S., S.R.S., Y.M., J.A.W. and M.A.B. performed MEI analyses; Y.M. performed SV analyses; Y.M., A.S. and P.H. performed ILS analyses; M.V. and E.E.E. supervised the project and finalized the manuscript. All authors read and approved the manuscript.

Code availability. Custom scripts used in this study are available at GitHub (https://github.com/EichlerLab and https://github.com/MaoYafei).

Competing interests: J.G.U. is an employee of Pacific Biosciences, Inc. All the other coauthors declare no competing financial and non-financial interests.

[10] Howard Hughes Medical Institute, University of California, Santa Cruz, Santa Cruz, CA 95064, USA.

[11] Bionano Genomics, San Diego, CA 92121, USA.

Summary

The divergence of chimpanzee and bonobo provides one of the few examples of recent hominid speciation[1,2]. We present a fully annotated, high-quality bonobo genome assembly, constructed without guidance from reference genomes, by applying a multiplatform genomics approach. We generate a bonobo genome assembly where >98% genes are completely annotated and 99% of the gaps are closed, including resolution of about half of the segmental duplications and almost all full-length mobile elements. We compare the bonobo genome to other apes[1,3-5] and identify >5,571 fixed structural variants that specifically distinguish the bonobo and chimpanzee lineages. We focus on genes that have been lost, changed in structure, or expanded in the last few million years of bonobo evolution. We produce a high-resolution map of incomplete lineage sorting (ILS) estimating that ~5.1% of the human genome is closer to chimpanzee/bonobo and >36.5% of the genome shows ILS if we consider a deeper phylogeny including gorilla and orangutan. Here, we show that 26% of human–chimpanzee/human–bonobo ILS segments are nonrandomly distributed and genes within these clustered segments show significantly accelerated amino acid evolution when compared to the rest of the genome.

Bonobo or pygmy chimpanzee (*Pan paniscus*) and the common chimpanzee (*Pan troglodytes*) are among the most recently diverged ape species (1.7 million years ago)[1,2]. Both species represent the closest living species to humans and, therefore, offer the potential to pinpoint genetic changes that are also unique to human. The first bonobo generated from short-read whole-genome sequence[1] produced a genome assembly (panpan1.1) with >108,000 gaps where the vast majority of segmental duplications (SDs) were not incorporated and few structural variants (SVs) were identified (Supplementary Table 1). As a result of the lower accuracy of early next-generation sequencing technology[6] and the fragmentary nature of the original chimpanzee genome[5], large fractions of ape genomes could not be compared[1,3-5] and gene models were often incomplete[7,8]. In the last few years, long-read genome sequencing technologies have significantly enhanced our ability to generate contiguous, high-quality genomes where most genes and common repeat elements are fully annotated[9]. Here, we apply a multi-platform approach to produce a highly contiguous, accurate bonobo reference genome. Our analysis highlights the extent and rapidity at which hominid genomes can differ and provides new insights with respect to ILS and its relevance to gene evolution and the genetic relationship among living hominids.

Sequence and assembly.

We sequenced DNA from a female bonobo (Mhudiblu, *Pan paniscus*) to 74-fold sequence coverage using the long-read PacBio RS II platform (Supplementary Tables 2 and 3, Supplementary Fig. 1). We generated a 3.0 Gbp assembly (contig N50 of 16.58 Mbp; Supplementary Table 4) and constructed a chromosomal-level AGP (a golden path) assembly (Mhudiblu_PPA_v0) using Bionano optical maps and a FISH BAC clone-order framework (Fig. 1)[10]. The Mhudiblu_PPA_v0 assembly assigns 74 Mbp of new sequence to chromosomes closing 99.5% of the original 108,390 gaps (Supplementary Table 5). This assembly has been annotated by NCBI and is available on the UCSC Genome Browser (panPan3, see Data access and Extended Data Figure 1 for assembly workflow schema). We estimate the sequence accuracy of the bonobo assembly to be 99.97-99.99% (Supplementary Table 6). The overall nucleotide divergence between chimpanzee and bonobo based on these new long-read assemblies is 0.421±0.086% for autosomes and 0.311±0.060% for the X chromosome (Supplementary Table 7). Using these new assemblies, we genotyped 27 previously sequenced ape genomes resulting in slight adjustments in mean effective population sizes for the apes (Extended data Fig. 2).

Gene annotation.

We predict 22,366 full-length protein-coding genes and 9,066 noncoding genes using the NCBI Eukaryotic Genome Annotation Pipeline. We also generated 857,000 full-length bonobo cDNA (Supplementary Table 8) and applied the Comparative Annotation Toolkit (CAT)[11] to identify 20,478 protein-coding and 36,880 noncoding bonobo gene models; 99.5% of the protein-encoding models show no frameshift errors[12] and 38.4% of protein-coding isoforms are now more complete. We identify 119 genes that have potential frameshifting indels disrupting the primary isoform relative to the human reference (GRCh38) (Supplementary Table 9). Respectively, 206 and 1,576 protein-coding genes are part of gene families that contracted or expanded in bonobo when compared to human (Supplementary Tables 10-12). We identify 65 putatively novel exons with support from full-length cDNA (Supplementary Tables 13-14), such as the novel protein-coding exon in *ANAPC2* found in bonobo but not in chimpanzee (Supplementary Fig. 2). Using other ape genomes[13,14] and a genome-wide analysis from 20 bonobo and chimpanzee samples, we identified genes showing an excess of amino acid replacement, balancing selection, and potential selective sweeps (Tajima's D)[15]. Most of the genes showing selective sweeps in bonobo (*DIRC2, GULP1, ERC2*; Supplementary Tables 15-18) or chimpanzee (*KIAA040*, *TM4SF4*, *FOXP2*; Supplementary Tables 19-22) are novel.

Mobile element insertions (MEIs).

The number of full-length (retrotransposition competent), lineage-specific L1 elements in the bonobo genome (415 L1Pt) is similar to chimpanzee (383 L1Pt) and 15-25% greater than human (330 L1HS) (Supplementary Fig. 3-5). An analysis of Alu repeats leads to a refined subfamily classification and we find that the number of bonobo-specific elements is nearly identical to that of chimpanzee (n = 1,492). *Pan* lineages, thus, show among the lowest rates of Alu insertions when compared to humans (where the rate has doubled) and the rhesus genome (~10-fold increased rate) (Extended Data Fig. 3). While bonobo shows reduced single-nucleotide variant genetic diversity[16] when compared to chimpanzee, we find that bonobo SVA elements are more copy number polymorphic (45%) (Extended Data Fig. 3) when compared to chimpanzee (35%; $p < 6.5 \times 10^{-4}$). In contrast, the chimpanzee-specific endogenous retrovirus (PTERV1) shows an indistinguishable low rate of polymorphism for PtERV1 in both species (9%) suggesting relatively little activity since *Pan* divergence (Supplementary Data).

Segmental duplications (SDs).

We identify 87.4 Mbp of SDs (≥1 kbp and ≥90% identity) (Supplementary Fig. 6-7; Supplementary Table 23 and Extended Data Fig. 3), most of which was previously unassembled. SDs are interspersed with an excess of large (≥10 kbp) intrachromosomal duplications consistent with the burst of SDs that occurred at the root of the hominid lineage[17]. Despite the ~6-fold improvement, the largest and most identical duplications were still not initially resolved (~84 Mbp). Using the Segmental Duplication Assembler (SDA) algorithm[18,19], we successfully resolved an additional 56 Mbp (Supplementary Table 24) and used these to identify recent gene family expansions (Supplementary Tables 25-31 and Extended Data Fig. 4). We show, for example, that the Eukaryotic Translation Initiation Factor 4 Subunit A3 (*EIF4A3*) gene family has expanded in both chimpanzee and bonobo. There is evidence that five out of the six paralogs are expressed and encode a full-length open reading frame (ORF; Fig. 2 and Extended Data Fig. 5). We estimate that the initial *EIF4A3* gene duplication occurred in the ancestral lineage approximately 2.9 million years ago. It then subsequently expanded and experienced gene conversion events independently in the chimpanzee and bonobo lineages creating five and six copies of the *EIF4A3* gene family, respectively. Interestingly, some of the gene conversion signals correspond to a set of specific amino-acid changes in the basic ancestral structure that are now common to only chimpanzee and bonobo (Fig. 2 and Extended Data Fig. 5).

Structural variation and gene disruption.

As part of the assembly curation, we validated nine larger inversions that distinguish human and bonobo karyotypes and created a FISH-based chromosomal backbone (Fig. 1a and b) and used Strand-seq to assign orphan contigs to chromosomes (36 Mbp) (Mhudiblu_PPA_v1; Supplementary Tables 32-38). We identify 17 fixed inversions differentiating bonobo from chimpanzee of which 11 are bonobo specific (Supplementary Table 39) and 22 regions that likely represent bonobo inversion polymorphisms (Supplementary Table 40). Moreover, we assign 38 fixed inversions occurring in the common *Pan* ancestor (Supplementary Table 39). We annotated and validated the breakpoint intervals of each tested inversion (Supplementary Table 41) and found SDs or LINEs at the breakpoints of inversions in 82% and 86% of cases, respectively (Supplementary Table 40). We also compared the bonobo genome to that of human, chimpanzee, and gorilla to identify smaller (>50 bp) deletions and insertions. We classify 15,786 insertions and 7,082 deletions as bonobo-specific and genotyped these in a population of ape samples[16,20] to identify 3,604 fixed insertions and 1,965 fixed deletions of which only a small fraction (2.66% or 148/5,571) intersect genic functional elements (Supplementary Tables 42-45).

Bonobo-specific events deleting ENCODE regulatory elements (n = 381), for example, are enriched in membrane-associated genes with extracellular domains while chimpanzee-specific events (n = 185) are associated with cadherin-related genes (Supplementary Table 46). Deletions (n = 1,040) shared between chimpanzee and bonobo show an enrichment for the loss of putative regulatory elements associated with post-synaptic genes (3.32 enrichment; p = 1.2 x 10-7) and pleckstrin homology-like domains (6.15 enrichment; p = 1.20 x 10-9). We validate 110 events that disrupt protein-coding genes by generating high-fidelity genomic sequencing for each of the ape reference genomes and restricting to those events that could be genotyped in a population of genomes (Supplementary Data). As expected, many fixed gene-loss events occurred in genes tolerant to mutation, redundant duplicated genes, or genes where the event simply altered the structure of the protein. For example, we validate a 25.7 kbp gene loss of one of the keratin-associated genes (*KRTAP19-16*) associated with hair production in the ancestral lineage of chimpanzee and bonobo (Supplementary Fig. 8). In the bonobo lineage, we identify five fixed SVs affecting protein-coding genes (Supplementary Table 47), but only two of which completely ablate the gene. For example, *LYPD8*, which encodes a secreted protein that prevents gram-negative bacteria invasion of colonic epithelium, has been totally deleted by a 24.3 kbp bonobo-specific deletion. Similarly, *SAMD9* (SAMD Family Member 9) is a fixed gene loss in bonobo as a result of a 41.46 kbp bonobo-specific deletion. The other three bonobo-specific fixed SV events in protein-coding regions all maintain the ORF, including a 49-amino acid deletion of *ADAR1*, a gene critical for RNA editing and implicated in human disease (Extended Data Fig. 6)[21-23].

Incomplete lineage sorting (ILS).

The higher quality and more contiguous nature of the bonobo genome provide an opportunity to generate a higher-resolution ILS map. Compared to the original bonobo assembly where only ~800 Mbp (27%) could be analyzed, it is now possible to align ~76% of the genome in a four-way ape genome alignment (2,357 Mbp within 10 kbp windows; Supplementary Table 48) due to long-read genome assemblies[14]. We performed a genome-wide phylogenetic window-based analysis to systematically identify regions inconsistent with the species tree and classified these as human–bonobo and human–chimpanzee ILS topologies (Fig. 3). We predict that 5.07% of the human genome is closer to chimpanzee/bonobo (Table 1); 2.52% of the human genome is more closely related to the bonobo genome (human–bonobo (H-B) ILS segments) than the chimpanzee genome while 2.55% of the human genome is more closely related to the chimpanzee genome (Human-chimpanzee (H-C) ILS) than the bonobo genome (Fig 3a). This proportion of ILS nearly doubles earlier estimates (3.1%)[1] (Supplementary Table 1). Consistent with previous observations[1], the largest ILS segments are biased (1.8-fold) to intergenic regions, depleted for genes (>35%) and are particularly enriched in L1 content. Of note, the distribution of ILS segments is highly nonrandom based on simulation experiments. We specifically measured the distance between ILS segments (see below) and identified a subset (~26%) of sites that are significantly more clustered than expected by chance (Fig. 3b).

We focused specifically on protein-coding exons based on human RefSeq annotation[24] and identified 1,446 exons mapping to ILS regions (713 exons to H-B and 733 exons to a H-C topology; Supplementary Table 49). As a whole, genes corresponding to these ILS exons are significantly enriched in both glycoprotein function ($p = 1.30 \times 10^{-14}$ for H-B and $p = 5.60 \times 10^{-11}$ for H-C) and calcium-binding epidermal growth factor (EGF) domain function ($p = 4.40 \times 10^{-12}$ for H-B and $p = 9.40 \times 10^{-7}$ for H-C) (Supplementary Table 50). We considered multiple occurrences in the same gene and identified 84 genes with at least two exons under ILS (Supplementary Table 51) with some enrichment in photoreceptor activity ($p = 1.6 \times 10^{-4}$, Supplementary Table 51 and Supplementary Fig. 9) as well as EGF-like (p =

1.9 x 10^{-6}) and transmembrane (p = 2.4 x 10^{-3}) function. Overall, we observe a significant excess of amino acid replacement (dN/dS) for all 1,446 ILS exons when compared to non-ILS exons consistent with either the action of relaxed selection or positive selection (Fig. 3c and Extended Data Fig. 7). Exons mapping to the clustered ILS segments show greater dN/dS with respect to exons in the non-clustered ILS segments suggesting that they are contributing disproportionately to accelerated amino acid evolution in the hominid genome.

We extended the ILS analysis (Supplementary Data) across 15 million years of hominid evolution by inclusion of orangutan and gorilla ape genome data. As expected, ILS estimates for the human genome increase to >36.5% (Supplementary Table 52, Extended Data Fig. 7) similar to (albeit still greater than) earlier estimates[3,14]. We measured the inter-ILS distance and observed a consistent nonrandom pattern of clustered ILS for these deeper topologies with more ancient ILS showing an even greater proportion of clustered sites (Fig. 3). Once again, we observe a significant elevated mean dN/dS in clustered H-C and H-B (p < 2.2e-16, mean = 0.366) as well as clustered Orangutan-Human and Orangutan-Gorilla-Human topologies (p < 2.2e-16, mean = 0.316) when compared to the null distribution (Supplementary Fig. 10). A GO analysis[25] of the genes intersecting these combined data confirm the most significant signals for immunity (e.g., glycoprotein (p = 1.3E-25), immunoglobulin-like fold/FN3 (p = 2.4E-20)), but also genes related to EGF signaling (p = 1.4E-18), solute transporter function (e.g., transmembrane region (p = 1.3E-25)), and specifically calcium transport (p = 3.7E-8) (Supplementary Table 53). While ILS regions, in general, show single-nucleotide polymorphism diversity patterns consistent with balancing selection, it is noteworthy that both clustered and non-clustered ILS exons show a significant excess of polymorphic gene-disruptive events consistent with the action of relaxed or balancing selection (Supplementary Fig. 11). An examination of these gene-rich clustered ILS regions reveals a complex pattern of diverse ILS topologies suggesting deep coalescence operating across specific regions of the human genome as has been reported for major histocompatibility complex (Extended Data Fig. 8).

Discussion

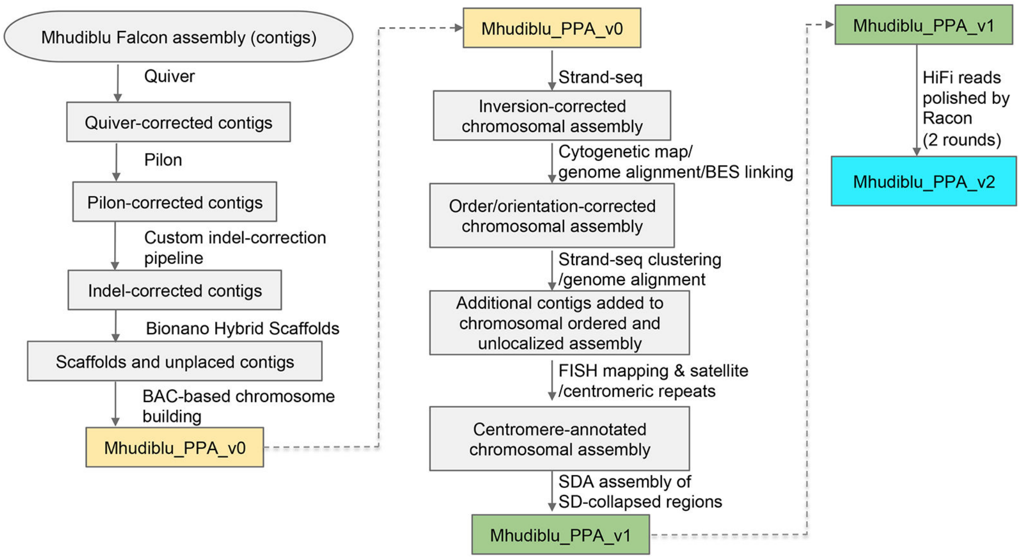
High-quality hominid genomes are a critical resource for understanding the genetic differences that make us human as well as the diversification of the *Pan* lineage over the last two million years of evolution. The bonobo represents one the last of the great ape genomes to be sequenced using long-read sequencing technology. Its sequence will facilitate more systematic genetic comparisons between human, chimpanzee, gorilla, and orangutan without the limitations of technological differences in sequencing and assembly of the original reference[1,3-5,14]. As a result, we now predict that a significantly greater fraction (~5.1%) of the human genome is closer to chimpanzee/bonobo when compared to previous studies (3.1%)[1,3]. We estimate that >36.5% of the hominid genome shows ILS if we consider a deeper phylogeny including gorilla and orangutan. Remarkably, 26% of the ILS regions are clustered and exons underlying these clustered ILS signals show elevated rates of amino acid replacement. These findings support a previous study in gorilla that showed a subtler correlation where genes with higher dN/dS values are enriched in ILS segments[3]. In that study, however, they explained the observation as a result of stronger purifying selection in non-ILS sites or background selection reducing the effective population size and as a result a depletion of ILS. Our genome-wide exon analyses specifically show that only a subset of clustered ILS exons are driving this effect and that these genes are enriched for glycoprotein and EGF-lke calcium signaling functions due to the action of either relaxed or positive selection for genes in these pathways (Supplementary Information).

METHODS SUMMARY

We sequenced and assembled the genome of a single female bonobo (Mhudiblu a.k.a. Mhudibluy obtained from the San Diego Zoo ISIS 601152 born 15 April 2001 and later transferred to the Wuppertal Zoo in Germany where she was referred to as Muhdeblu) using long-read PacBio RS II sequencing chemistry and the Falcon genome assembler. The assembly was error corrected using Quiver[26], Pilon[27], and an in-house FreeBayes-based[28] indel correction pipeline optimized to improve CLR assemblies[14]. We also generated Illumina WGS data using the Illumina TruSeq PCR-Free library preparation kit. Genome assembly contigs were ordered and oriented into scaffolds using Bionano optical maps (Supplementary Table 54) (HybridScaffolds suite, Bionano Genomics Saphyr platform) and four-color FISH of 324 BAC clones. Cell lines from chimpanzee, bonobo, gorilla and orangutan were obtained from Coriell (S006007) or from a collection developed by Professor Mariano Rocchi; no approval from ethics committees were required for use of these established lines. We assigned each contig/scaffold into unique groups corresponding to individual chromosomal homologues using SaaRclust[29,30] while applying Strand-seq to detect inversions, assign orphan contig and orient contigs[31,32]. To estimate genome-wide sequence accuracy, we applied Merqury[33] using Illumina WGS data. We also generated a bonobo large-insert BAC library (VMRC74) and selected at random 17 clones for complete PacBio insert sequencing[34]. CAT[11] was utilized for genome annotation using human GENCODE V33 and RNA-seq data. We also generated >850,000 full-length non-chimeric transcripts from Iso-Seq data generated from iPSC and derived neuronal progenitor cell lines[35] from bonobo sample AG05253 and we searched for gene structures split over multiple contigs (Supplementary Table 55). Repeat content of the assembled genome was analyzed using RepeatMasker (RepeatMasker-Open-4.1.0) and the Dfam3 repeat library. We assigned lineage-specific Alu and full-length LINE, SVA_D and PtERV elements to subfamilies by applying COSEG (www.repeatmasker.org/COSEGDownload.html) to determine the lineage-specific subfamily composition. For cross-species analysis of MEIs, we performed liftOver based on chains built from the Cactus whole-genome alignments generated during CAT annotation. For cross-assembly analyses of bonobo MEI insertions and a specific subset of other analyses (Supplementary Data), we used bowtie2 to map MEI flanking sequences between genomes. We estimated the duplication content in the bonobo assembly, applying the whole-genome analysis comparison (WGAC) method[36] and targeted collapsed duplications for assembly using SDA[19]. Insertions and deletions were detected in bonobo, chimpanzee, and gorilla using PBSV, Sniffles[37] and Smartie-sv[14] and genotyped using Paragraph[38] against a panel of 27 Illumina WGS genomes. We searched for evidence of ILS among the chimpanzee, gorilla, and human lineages applying Prank (v.140110) to construct multiple sequence alignments and using ete3 module to identify segments under ILS (Supplementary Table 56). For consistency, NCBI reference genome nomenclature has been used throughout the manuscript and corresponds to the following UCSC IDs (NCBI/

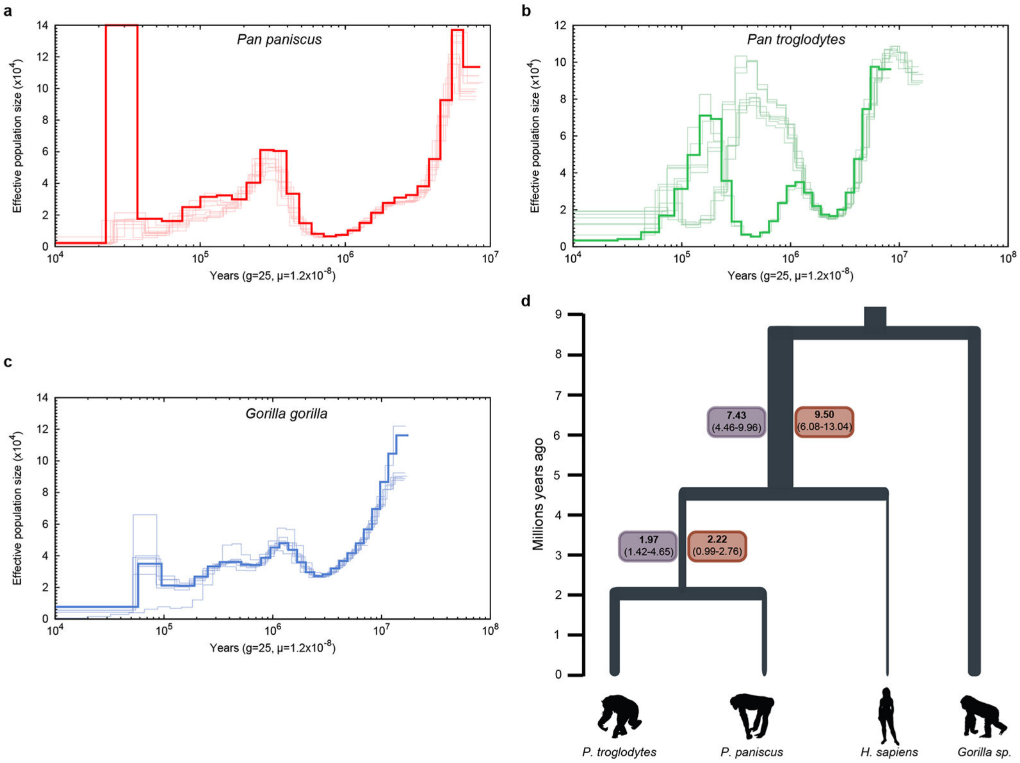
UCSC): panpan1.1/panPan2, Mhudiblu_PPA_v0/panPan3, Clint_PTRv2/panTro6, Kamilah_GGO_v0/gorGor6, Susie_PABv2/ponAbe3, GRCh38/hg38 (methods details are in the Supplementary Data).

Extended Data


Mhudiblu Falcon assembly (contigs)
Quiver
Quiver-corrected contigs
Pilon
Pilon-corrected contigs
Custom indel-correction pipeline
Indel-corrected contigs
Bionano Hybrid Scaffolds
Scaffolds and unplaced contigs
BAC-based chromosome building
Mhudiblu_PPA_v0
Mhudiblu_PPA_v0
Strand-seq
Inversion-corrected chromosomal assembly
Cytogenetic map/ genome alignment/BES linking
Order/orientation-corrected chromosomal assembly
Strand-seq clustering /genome alignment
Additional contigs added to chromosomal ordered and unlocalized assembly
FISH mapping & satellite /centromeric repeats
Centromere-annotated chromosomal assembly
SDA assembly of SD-collapsed regions
Mhudiblu_PPA_v1
Mhudiblu_PPA_v1
HiFi reads polished by Racon (2 rounds)
Mhudiblu_PPA_v2

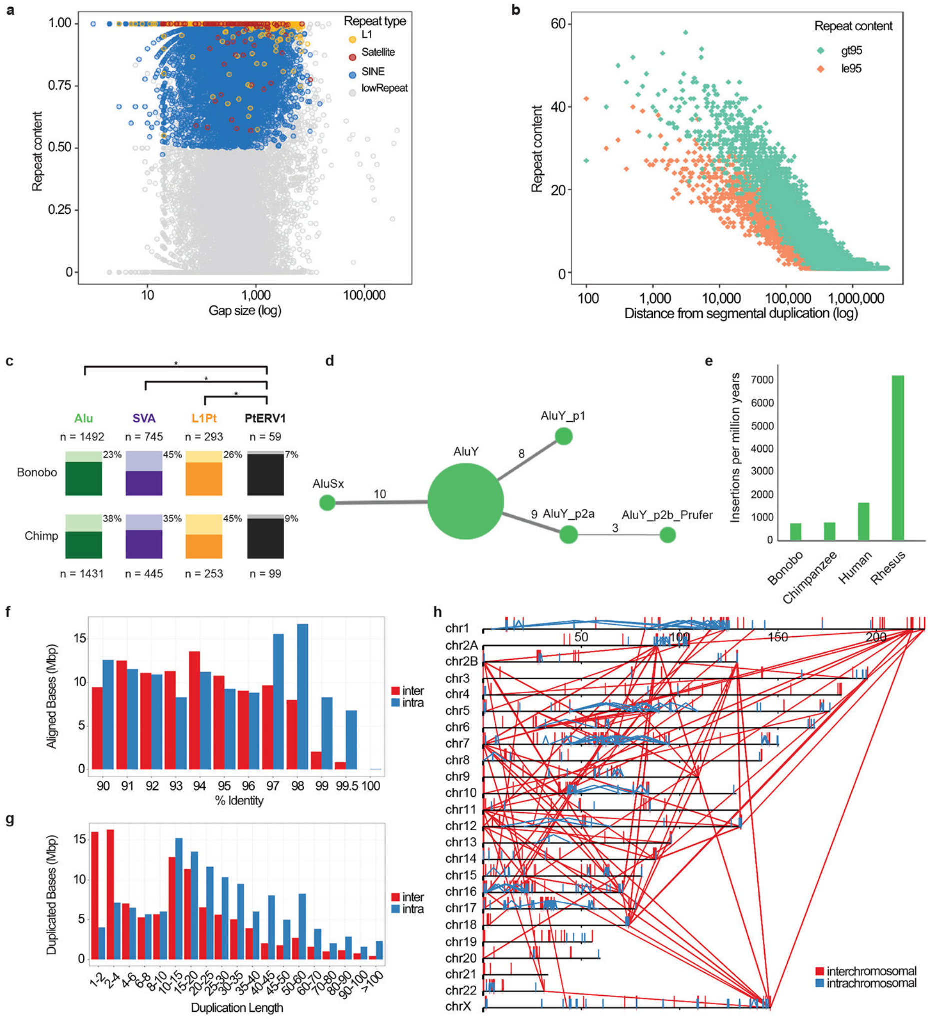
Extended Data Figure 1.

Processing steps to create the reference sequences Mhudiblu_PPA_v0, Mhudiblu_PPA_v1, and Mhudiblu_PPA_v2.



Extended Data Figure 2. PSMC analysis and estimates of effective populations size (Ne) in demes predating divergence in Homo and Pan.

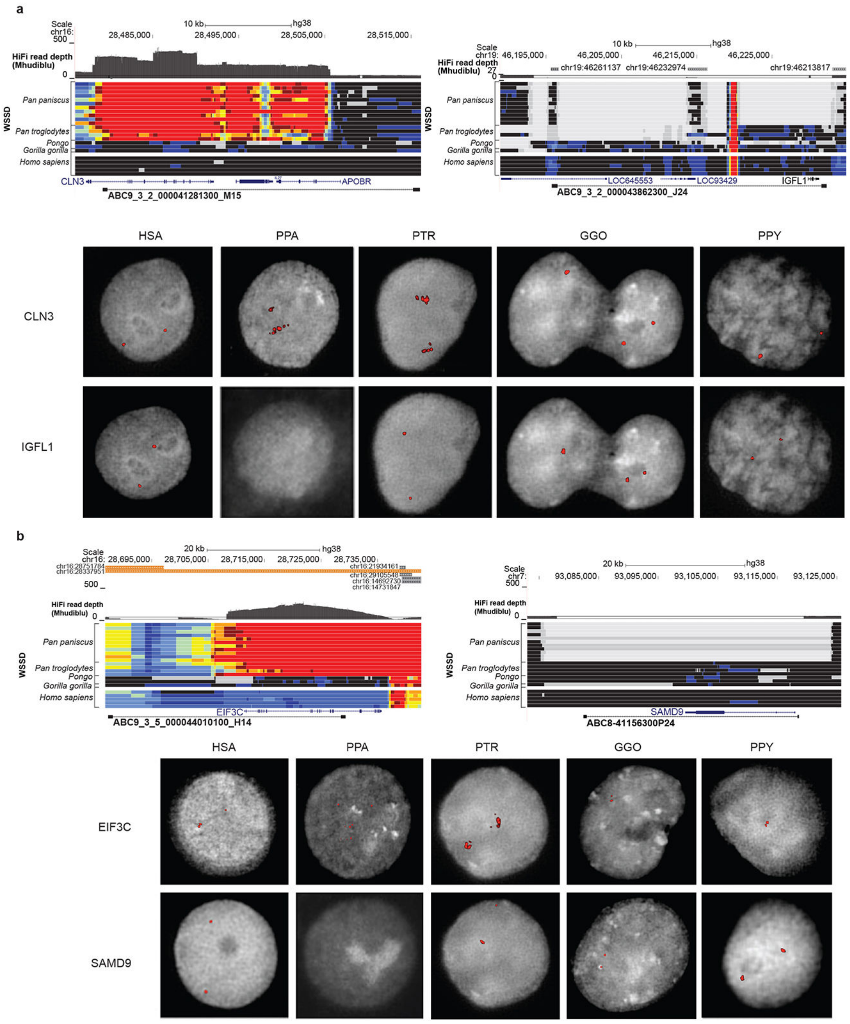
PSMC plots based on an analysis of Illumina WGS genomes (**a**) 10 bonobos (red), (**b**) 10 chimpanzees (green), and (**c**) 7 gorillas (blue). The y-axis represents the Ne (x 10^4) inferred by the PSMC and the x-axis represents the time in years. Ne and time are scaled with generation time g = 25 years and the mutation rate per generation μ = 0.5 x 10^{-9} mut (bp x year)[16]; (**d**) values in boxes refer to Ne (x 10^4) inferred through PSMC analysis considering bonobo (red boxes) and chimpanzee (purple). We extracted size estimates from time intervals between 4-7 million years ago for the (human, pan) Ne and 1-2.5 million years ago for (*P. paniscus*, *P. troglodytes*), considering μ= 0.5 x 10^{-9} mut (bp x year) and generation time of 25 years. Values using μ= 1 x 10^{-9} mut (bp x year) are reported in Supplemntary Note Table S23.



Extended Data Figure 3. Sequence and assembly of the bonobo genome and bonobo genome repeat structure.

a, The size (log scale x-axis) and repeat content of gaps filled in the new bonobo assembly as compared to panpan1.1[1]. Gaps composed of more than 50% repeat content for any particular class of repeat are colored. **b,** Distance from filled gaps to the nearest segmental duplication (SD). The number of base pairs (counts in bins of 100 base pairs) mapping near annotated SDs are shown for high-identity (>95%, brown) and low-identity (≤95%, green) SDs revealing preferential closure of larger gaps mapping adjacent to higher-identity duplications. Note: An additional 2,600 and 1,755 filled gaps map directly within SD sites ≤95% and >95% identity, respectively. **c,** Polymorphism rates for lineage-specific mobile element insertions (MEIs). Alu, SVA, L1Pt, and PTERV1 insertions that do not “liftover” between chimpanzee and bonobo reference genomes were identified and genotyped for deletions using data from 10 bonobos and 9 chimpanzees. Light bars and percentages represent the fraction of instances of the MEI type that display support for polymorphism; solid bars represent the fraction of fixed insertions in these populations. PTERV1 displays a significantly less polymorphic fraction than Alu ($p = 2.6 \times 10^{-74}$, chimpanzee; $p = 6.9 \times 10^{-35}$, bonobo; chi-squared test, Bonferroni correction), SVA ($p = 3.8 \times 10^{-19}$; $p = 1.9 \times$

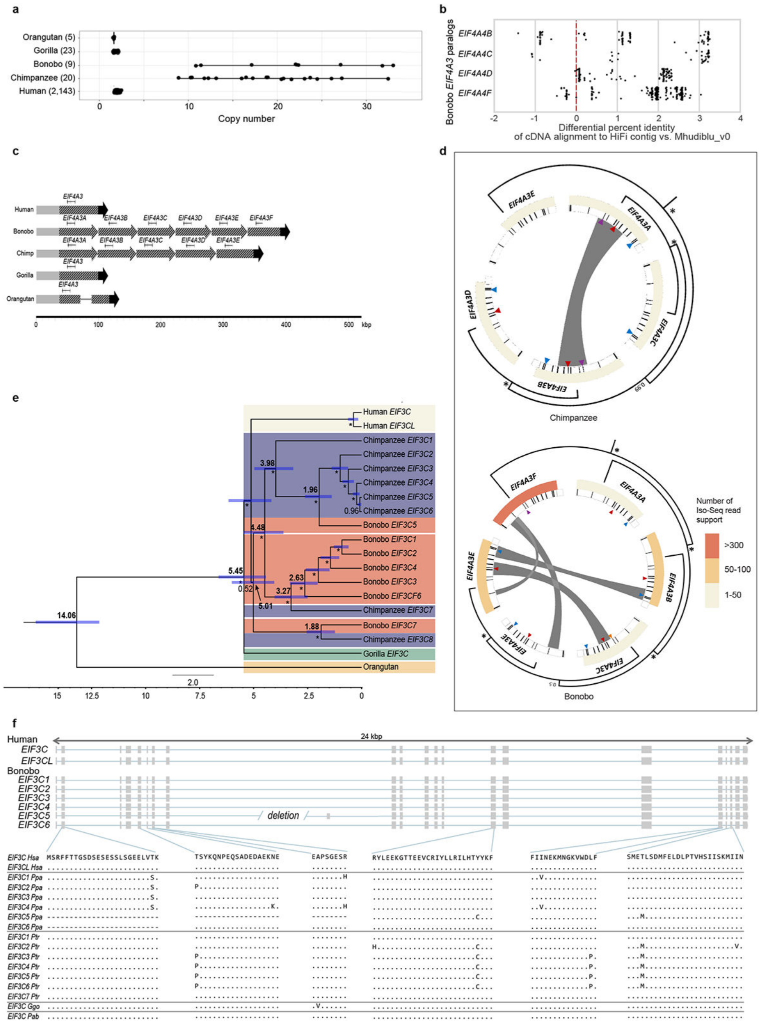
10^{-62}), or L1Pt ($p = 2.2 \times 10^{-18}$; $p = 1.3 \times 10^{-8}$), reflecting its lack of activity since *Pan* divergence. SVA is the only MEI type with a greater polymorphism rate in bonobo. **d,** A COSEG network of bonobo-specific Alu subfamilies indicating the relative number of elements (size of the node) and number of mutations (line thickness) distinguishing subfamilies. **e,** A comparison of the retrotransposition rate per million years based on lineage-specific Alu insertions from a select panel of primate genomes. The (**f**) percent (%) identity distribution and (**g**) length distribution of SDs (≥90% identify, ≥1 kbp, and no unplaced contigs) are shown as well as the (**h**) pattern of the largest and most identical (≥10 kbp and ≥98%) intrachromosomal (blue) and interchromosomal (red) SDs in the bonobo genome.



Extended Data Figure 4. *Pan*-specific duplications and bonobo-specific deletions.

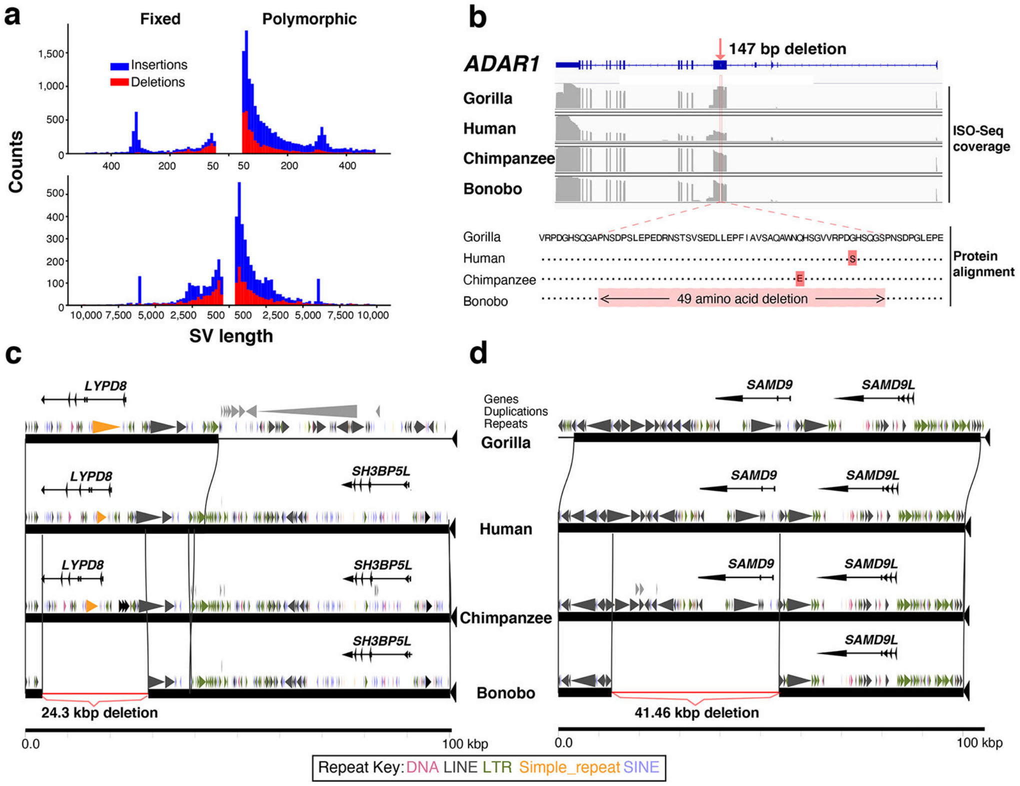
a, *Pan*-specific duplication of *CLN3* locus and bonobo-specific deletion of *IGFL1*. HiFi read depth and whole-genome shotgun detection (WSSD) of bonobo, chimpanzee, orangutan, gorilla, and human individuals relative to GRCh38 detect these events (above), which are validated by interphase FISH of each species using fosmid clones spanning the region (below). **b,** *Pan*-specific duplication of *EIF3C* locus and bonobo-specific deletion of

SAMD9. HiFi read depth and WSSD of bonobo, chimpanzee, orangutan, gorilla, and human individuals* relative to GRCh38 detect these events (above), which are validated by interphase FISH of each species using fosmid clones spanning the region (below). *In order from the top line: Pan_paniscus_A915_Kosana, A927_Salonga, A922_Catherine, A917_Dzeeta, A918_Hermien, A924_Chipita, A926_Natalie, A928_Kumbuka, A914_Hortense, A919_Desmond, A925_Bono; Pan_troglodytes_troglodytes_A958_Doris, A957_Vaillant, A960_Clara, Pan_troglodytes_verus_Clint, Pongo_abelii_A950_Babu, Pongo_pygmaeus_A944_Napoleon, Gorilla_gorilla_gorilla_KB4986_Katie, AFR_Aari_ETAR005_F, AMR_Nahua_Mex20_M, EA_Mongola_HGDP01228_M, SA_Kalash_HGDP00328_M, WEA_FinlandFIN_HG00360_M.



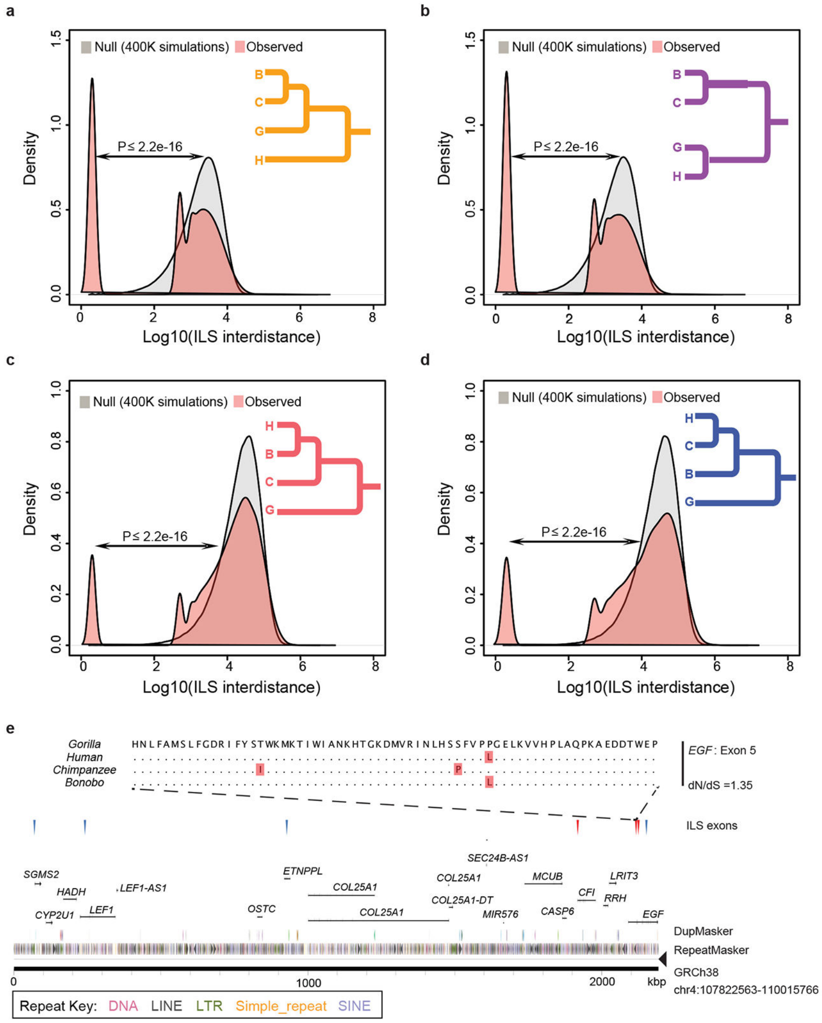
Extended Data Figure 5. *EIF4A3 and EIF3C* gene family expansion and sequence resolution.

a, A comparison of *EIF4A3* copy number among apes based on a sequence read-depth analysis confirms a variable copy number expansion in the bonobo and chimpanzee lineages (9-33 diploid copies). This recent duplication was not fully resolved initially in the bonobo reference genome (Mhudiblu_PPA_v0) because high-identity duplicated sequences were collapsed. **b,** Bonobo Iso-Seq full-length transcript reads map with higher identity to four of the paralogs when compared to Mhudiblu_PPA_v0. **c,** Contigs that encompass *EIF4A3* expansions and 100 kbp of the flanking regions were assembled using bonobo and chimpanzee PacBio HiFi data. The 12 kbp genomic sequence of human *EIF4A3* mapped onto the assembled contigs. Six tandem copies of *EIF4A3* spanning 310 kbp in bonobo and five tandem copies spanning 262 kbp in chimpanzee are recovered. Schematics show structural differences of *EIF4A3* in primate genomes. Gray, black, and striped arrows show different alignment blocks across the samples. A solid line connecting alignment blocks indicates an insertion event. **d,** Paralogs are expressed and show evidence of gene conversion in both bonobo and chimpanzee lineages. Analysis of bonobo Iso-Seq data confirms that five of the six *EIF4A3* copies are expressed and maintain an open reading frame (heatmap indicates the number of Iso-Seq transcripts supporting each copy; minimap2 -ax splice -G 3000 -f 1000 --sam-hit-only --secondary=no --eqx -K 100M -t 20 --cs −2 | samtools view -F 260). GENECONV software shows significant signals ($p \leq 0.05$ after multiple test correction) of gene conversion for 16/67 kbp of the paralogous locus (gray bars) (MSA was performed using MAFFT version 7.453 (command: mafft -adjustdirection [input.fasta] > [output.msa_fasta]; GENECONV version 1.81a)). A subset of gene conversion events overlap with sites of amino-acid specific to the *Pan* lineage. Triangles indicate the sites of amino acid change in each of the primate genomes compared to GRCh38. Different colors mark different changes: purple marks phenylalanine to leucine; yellow marks arginine to cysteine; red marks serine to arginine; teal marks tyrosine to serine. Same phylogenetic tree from Figure 2 is reshaped to show the inferred evolutionary relationships among the paralogs. Nodes with >99% Bayesian posterior probabilities are indicated by asterisks; otherwise the actual number is shown. **e,** A phylogenetic tree was constructed from 16 kbp noncoding *IF3C* paralogs using Bayesian phylogenetic inference. This analysis was conducted using BEAST2 software. Bolded numbers on each major node denote estimated divergence time. Regular numbers indicate posterior probability. The blue error bar on each node indicates 95% confidence interval of the age estimation. Bootstrap supports are reported using asterisks for nodes with posterior probability >99%. **f,** Gene models for transcribed loci based on Iso-Seq data (above). Human *EIF3C* and *EIF3CL* are compared to predicted open reading frames for bonobo paralogs and Liftoff gene predictions for chimpanzee, orangutan, and gorilla paralogs from contigs assembled from HiFi reads.



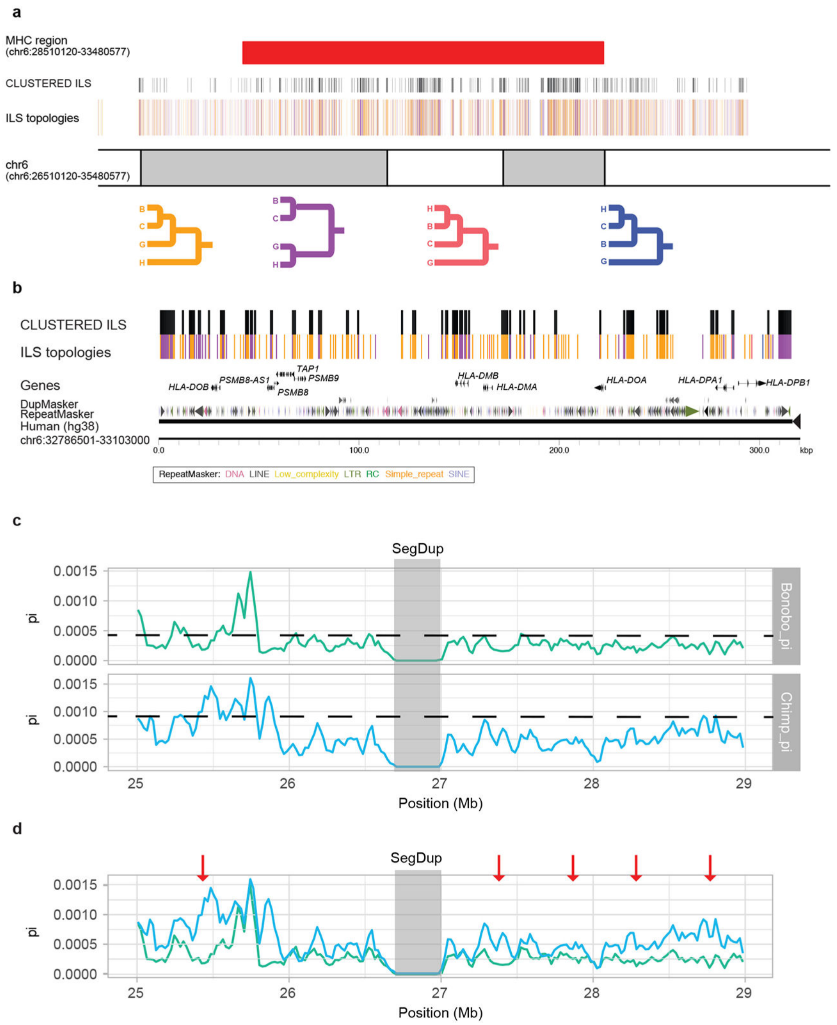
Extended Data Figure 6. Bonobo structural variants (SVs) and gene deletions.

a, Size distribution of fixed (left) and polymorphic (right) SV insertions and deletions in the bonobo genome. Top panel: 50-1000 bp and bottom panel: >1000 bp in length. Events are deemed to be specific to the bonobo lineage based on copy number genotyping against a panel of 26 ape genomes and a threshold of Fst > 0.8 to define fixed events in bonobo. Modes are observed corresponding to full-length L1 (6 kbp) and Alu (300 bp) mobile elements and are predominantly insertions reflecting the homoplasy-free nature of this class of mutation. **b,** A small fixed deletion predicts a 49 amino acid deletion in *ADAR1* in the bonobo lineage. RefSeq *ADAR1* structure is shown (top) compared to Iso-Seq coverage of gorilla, human, chimpanzee, and bonobo (middle). Protein alignment (bottom) shows that an in-frame deletion is created. **c,** A 24.3 kbp fixed deletion results in the complete loss of *LYPD8* in bonobo. Gene structure, duplication, and repeat annotations are shown with respect to gorilla, human, chimpanzee, and bonobo genomes. Note: a lineage-specific duplication adjacent to *LYPD8* is present in the gorilla genome (large gray triangles). **d,** A 41.5 kbp fixed deletion mediated by directly orientated L1 repeats ablates *SAMD9* leaving only *SAMD9L* in the bonobo lineage. **e,** Short-read WSSD genotyping shows *LYPD8* was lost in the bonobo lineage. **f,** Short-read WSSD genotyping shows *SAMD9* was lost in the bonobo lineage.



Extended Data Figure 7. Hominid incomplete lineage sorting (ILS).

The distance between adjacent ILS segments (inter-ILS) (500 bp resolution) was calculated and the distribution was compared to a simulated expectation based on a random distribution. The analysis reveals a bimodal (and possibly an emerging trimodal) pattern where a distinct subset of ILS are clustered (i.e., clustered ILS sites). Four different topologies were considered: **a,** (orangutan,(((bonobo,chimp),gorilla),human)) ILS topology where 31.58% of inter-ILS are clustered; **b,** (orangutan,((bonobo,chimp),(gorilla,human))) ILS topology where 33.5% are clustered; **c,** (orangutan,(((bonobo,human),chimp),gorilla)) ILS topology (8.14%); and **d,** (orangutan,((bonobo,(chimp,human)),gorilla)) ILS topology (9.89% of sites) and **e**, An example of a cluster of human–bonobo (red triangles) and human–chimpanzee (blue triangles) ILS corresponding to a group of genes. A four-species alignment of one exon from *EGF* (exon 5) is shown with a nominal signal of positive selection.



Extended Data Figure 8. Ideogram of the MHC region with ILS annotations.

a, The four main ILS topologies are color-coded below. The four color lines representing ILS segments are shown above the chromosome coordinate (hg38). The clustered ILS are shown above the four color lines (black). The MHC region (red bar) corresponds to genomic coordinates chr6:28510120-33480577. **b,** A zoomed-in view of the MHC region (chr6:32786501-33103000) depicts clustered ILS nearby *HLA* genes. **c,** Nucleotide diversity of bonobo (green) and chimpanzee (blue) are shown based on human genomic coordinates (hg38, chr6: 25000000-29000000). The mean (dashed line) is shown for bonobo (mean = 4.45e-4) and chimpanzee (mean = 9.35e-4). A region of reduced diversity (gray) is shown but corresponds to an SD where single-nucleotide polymorphisms (SNPs) were excluded due to potential mismapping. **d,** Same as (c) but merged onto the same scale and highlighting five regions (red arrows) where diversity is reduced in bonobo when compared to chimpanzee. Three of these correspond to regions identified by Prufer et al.[1]; however, they are not among the top 1% of genome candidates showing positive selection by Tajima's D and SweepFinder2[15]. Overall SNP diversity is reduced across the region in bonobo when compared to chimpanzee.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements:

We thank Lucia Carbone and Oliver Ryder for providing the Mhudiblu bonobo cell line; Rusty Gage and Carol Marchetto for preparation of RNA sequence and access to bonobo iPSC lines; Arang Rhie for assistance with Merqury analysis; and Tonia Brown for manuscript proofreading and editing. The silhouettes of the bonobo, gorilla, and human in figure 3 and in the extended data figure 2 are downloaded from phylopic.org under the Public Domain Dedication 1.0 license. The silhouette of the chimpanzee in figure 3 and in the extended data figure 2 is downloaded from phylopic.org under the Creative Commons Attribution 3.0 Unported license (https://creativecommons.org/licenses/by/3.0/, credit to T. Michael Keesey and Tony Hisgett). This work was supported, in part, by the National Institutes of Health (NIH) grants HG002385 and 1U24HG009081 to E.E.E.; Futuro in Ricerca 2010-RBFR103CE3 to M.V.; R01 GM59290 to M.A.B.; 2U41HG007234 to M.D. and B.P.; U01HG010961, U41HG010972, R01HG010485, U01HL137183, and 5U54HG007990 to B.P.; Arian Smit's NHGRI grant RO1 HG002939 to J.M.S.; 5T32HG008345-04 to B.P; R01HG010329-01 to S.R.S; and European Regional Development Fund 2014-2020.4.01.16-0030 to F.M. E.E.E. is an investigator of the Howard Hughes Medical Institute. The work of Jinna Hoffman and Françoise Thibaud-Nissen was supported by the Intramural Research Program of the National Library of Medicine, National Institutes of Health. P.H. is supported by an NIH Pathway to Independence Award (NHGRI, K99HG011041).

Data availability.

The Mhudiblu_PPA_v0 (GCA_013052645.1), Mhudiblu_PPA_v1 (GCA_013052645.2) and Mhudiblu_PPA_v2 (GCA_013052645.3) assemblies are deposited at NCBI with BioProject accession PRJNA526933. The raw PacBio CLR, StrandSeq, Illumina, and Iso-Seq data of bonobo are deposited at NCBI with SRA accession SRP188441. The BioNano map of bonobo Mhudiblu is deposited at NCBI with BioProject accession PRJNA526933. The raw PacBio HiFi data of bonobo Mhudiblu and gorilla Kamilah are deposited at NCBI with accession SRP301932 in BioProject accession PRJNA691628. The BACs used in this study are listed in Supplementary Table 57 at NCBI with BioProject accession PRJNA634395.

References

1. Prufer K et al. The bonobo genome compared with the chimpanzee and human genomes. Nature 486, 527–531, doi:10.1038/nature11128 (2012). [PubMed: 22722832]
2. Takemoto H, Kawamoto Y & Furuichi T How did bonobos come to range south of the congo river? Reconsideration of the divergence of Pan paniscus from other Pan populations. Evol Anthropol 24, 170–184, doi:10.1002/evan.21456 (2015). [PubMed: 26478139]
3. Scally A et al. Insights into hominid evolution from the gorilla genome sequence. Nature 483, 169–175, doi:10.1038/nature10842 (2012). [PubMed: 22398555]
4. Locke DP et al. Comparative and demographic analysis of orang-utan genomes. Nature 469, 529–533, doi:10.1038/nature09687 (2011). [PubMed: 21270892]
5. Consortium, T. C. S. a. A. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437, 69–87, doi:nature04072 [pii] 10.1038/nature04072 (2005). [PubMed: 16136131]
6. Luo C, Tsementzi D, Kyrpides N, Read T & Konstantinidis KT Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. PLoS One 7, e30087, doi:10.1371/journal.pone.0030087 (2012). [PubMed: 22347999]
7. Prado-Martinez J et al. Great ape genetic diversity and population history. Nature 499, 471–475, doi:10.1038/nature12228 (2013). [PubMed: 23823723]
8. Sudmant PH et al. Global diversity, population stratification, and selection of human copy-number variation. Science 349, aab3761, doi:10.1126/science.aab3761 (2015). [PubMed: 26249230]

9. Logsdon GA, Vollger MR & Eichler EE Long-read human genome sequencing and its applications. Nat Rev Genet, doi:10.1038/s41576-020-0236-x (2020).

10. Ventura M et al. Gorilla genome structural variation reveals evolutionary parallelisms with chimpanzee. Genome Res 21, 1640–1649, doi:10.1101/gr.124461.111 (2011). [PubMed: 21685127]

11. Fiddes IT et al. Comparative Annotation Toolkit (CAT)-simultaneous clade and personal genome annotation. Genome Res 28, 1029–1038, doi:10.1101/gr.233460.117 (2018). [PubMed: 29884752]

12. Stanke M, Diekhans M, Baertsch R & Haussler D Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics 24, 637–644, doi:10.1093/bioinformatics/btn013 (2008). [PubMed: 18218656]

13. Gordon D et al. Long-read sequence assembly of the gorilla genome. Science 352, aae0344, doi:10.1126/science.aae0344 (2016). [PubMed: 27034376]

14. Kronenberg ZN et al. High-resolution comparative analysis of great ape genomes. Science 360, doi:10.1126/science.aar6343 (2018).

15. Pavlidis P & Alachiotis N A survey of methods and tools to detect recent and strong positive selection. J Biol Res (Thessalon) 24, 7, doi:10.1186/s40709-017-0064-0 (2017). [PubMed: 28405579]

16. de Manuel M et al. Chimpanzee genomic diversity reveals ancient admixture with bonobos. Science 354, 477–481, doi:10.1126/science.aag2602 (2016). [PubMed: 27789843]

17. Marques-Bonet T et al. A burst of segmental duplications in the genome of the African great ape ancestor. Nature 457, 877–881 (2009). [PubMed: 19212409]

18. Vollger MR et al. Improved assembly and variant detection of a haploid human genome using single-molecule, high-fidelity long reads. Ann Hum Genet, doi:10.1111/ahg.12364 (2019).

19. Vollger MR et al. Long-read sequence and assembly of segmental duplications. Nat Methods 16, 88–94, doi:10.1038/s41592-018-0236-3 (2019). [PubMed: 30559433]

20. Sudmant PH et al. Evolution and diversity of copy number variation in the great ape lineage. Genome Res 23, 1373–1382, doi:10.1101/gr.158543.113 (2013). [PubMed: 23825009]

21. Rice GI et al. Mutations in ADAR1 cause Aicardi-Goutières syndrome associated with a type I interferon signature. Nat Genet 44, 1243–1248, doi:10.1038/ng.2414 (2012). [PubMed: 23001123]

22. Savva YA, Rieder LE & Reenan RA The ADAR protein family. Genome Biol 13, 252, doi:10.1186/gb-2012-13-12-252 (2012). [PubMed: 23273215]

23. Gallo A, Vukic D, Michalík D, O'Connell MA & Keegan LP ADAR RNA editing in human disease; more to it than meets the I. Hum Genet 136, 1265–1278, doi:10.1007/s00439-017-1837-0 (2017). [PubMed: 28913566]

24. O'Leary NA et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44, D733–745, doi:10.1093/nar/gkv1189 (2016). [PubMed: 26553804]

25. Huang d. W., Sherman BT & Lempicki RA Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4, 44–57, doi:10.1038/nprot.2008.211 (2009). [PubMed: 19131956]

26. Chin CS et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods 10, 563–569, doi:10.1038/nmeth.2474 (2013). [PubMed: 23644548]

27. Walker BJ et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9, e112963, doi:10.1371/journal.pone.0112963 (2014). [PubMed: 25409509]

28. Haplotype-based variant detection from short-read sequencing (arXiv, 2012).

29. Ghareghani M et al. Strand-seq enables reliable separation of long reads by chromosome via expectation maximization. Bioinformatics 34, i115–i123, doi:10.1093/bioinformatics/bty290 (2018). [PubMed: 29949971]

30. Porubsky D et al. A fully phased accurate assembly of an individual human genome. bioRxiv, 855049, doi:10.1101/855049 (2019).

31. Falconer E et al. DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. Nat Methods 9, 1107–1112, doi:10.1038/nmeth.2206 (2012). [PubMed: 23042453]

32. Sanders AD et al. Characterizing polymorphic inversions in human genomes by single-cell sequencing. Genome Res 26, 1575–1587, doi:10.1101/gr.201160.115 (2016). [PubMed: 27472961]

33. Rhie A, Walenz BP, Koren S & Phillippy AM Merqury: reference-free quality and phasing assessment for genome assemblies. bioRxiv, 2020.2003.2015.992941, doi:10.1101/2020.03.15.992941 (2020).

34. Huddleston J et al. Reconstructing complex regions of genomes using long-read sequencing technology. Genome Res 24, 688–696, doi:10.1101/gr.168450.113 (2014). [PubMed: 24418700]

35. Marchetto MC et al. Species-specific maturation profiles of human, chimpanzee and bonobo neural cells. Elife 8, doi:10.7554/eLife.37527 (2019).

36. Bailey JA, Yavor AM, Massa HF, Trask BJ & Eichler EE Segmental duplications: organization and impact within the current human genome project assembly. Genome Res 11, 1005–1017, doi:10.1101/gr.187101 (2001). [PubMed: 11381028]

37. Sedlazeck FJ et al. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods 15, 461–468, doi:10.1038/s41592-018-0001-7 (2018). [PubMed: 29713083]

38. Chen S et al. Paragraph: a graph-based structural variant genotyper for short-read sequence data. Genome Biol 20, 291, doi:10.1186/s13059-019-1909-7 (2019). [PubMed: 31856913]

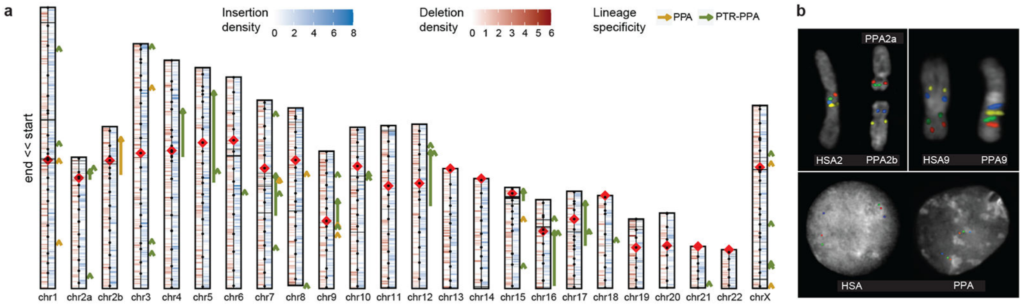


Figure 1. Sequence and assembly of the bonobo genome.

a, Schematic of the Mhudiblu_PPA_v0 assembly depicting centromere location (red rhombus), FISH probes used to create assembly backbone (black dots), fixed bonobo-specific insertions (blue) and deletions (red) (Supplementary Data), remaining gaps (black horizontal lines), and large-scale inversions (arrows). We distinguish bonobo-specific inversions (dark orange) from *Pan*-specific inversions (dark green). **b,** FISH validation of the bonobo chromosome 2A and 2B fusion and the 2B pericentric inversion (probes: RP11-519H15 in red, RP11-67L14 in green, RP11-1146A22 in blue, RP11-350P7 in yellow); the chromosome 9 pericentric inversion (probes: RP11-1006E22 in red, RP11-419G16 in green, RP11-876N18 in blue, RP11-791A8 in yellow); and the inversion Strand-seq_chr7_inv4a (probes: RP11-118D11 in green, WI2-3210F8 in red, RP11-351B3 in blue).

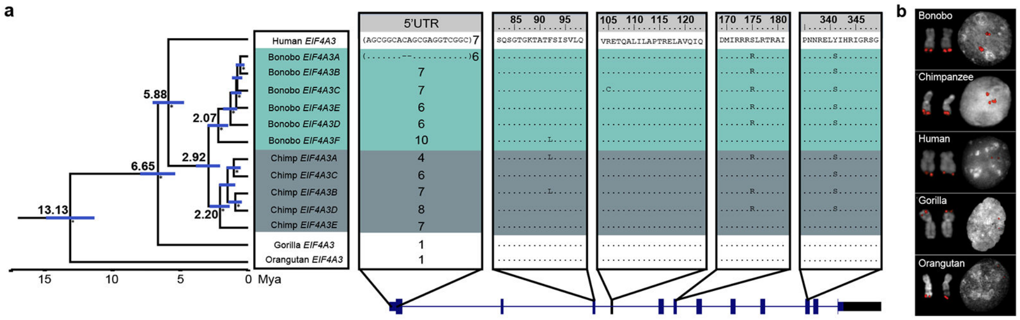


Figure 2. *EIF4A3* gene family expansion and sequence resolution.

a, Multiple sequence alignment shows *EIF4A3* amino acid differences between the human, Mhudiblu_PPA and chimpanzee assembled paralogs, and other great apes. A polymorphic 18 bp motif VNTR is located at the 5' UTR of nonhuman primate *EIF4A3* and accounts for most of the differences between various isoforms. A phylogenetic tree is built from neutral sequences of *EIF4A3* paralogs using Bayesian phylogenetic inference. This analysis is conducted using BEAST2 software. Numbers on each major node denote estimated divergence time (Mya: million years ago). The blue error bar on each node indicates 95% confidence interval of the age estimation. Bayesian posterior probabilities are reported using asterisks for nodes with posterior probability >99%. **b,** FISH on metaphase chromosomes and interphase nuclei with human probe WI2-3271P14 confirms an *EIF4A3* subtelomeric expansion of chromosome 17 in bonobo and chimpanzee relative to human, gorilla, and orangutan.

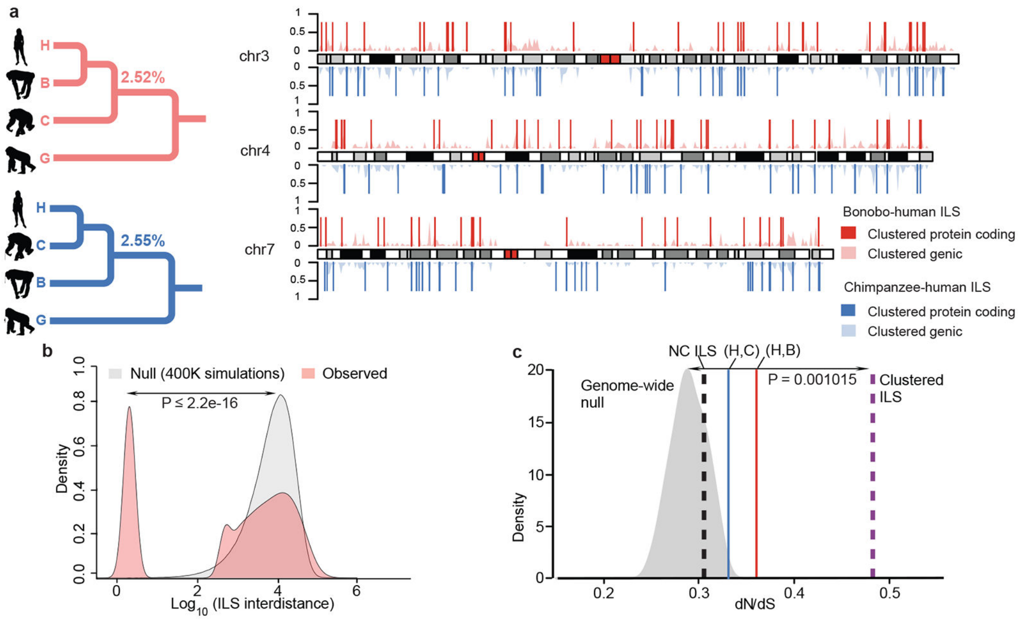


Figure 3. Hominid incomplete lineage sorting (ILS).

a, A schematic map of clustered ILS segments (500 bp resolution) for bonobo–human (red) and chimpanzee–human (blue) for chromosomes 3, 4 and 7. The lighter density plot represents the clustered ILS events (i.e., interdistance <100 bp) mapping to genes, while vertical lines represent the subset that overlap with protein-coding exons. **b,** Distribution of distances between ILS segments (inter-ILS) (500 bp resolution) compared to a simulated expectation reveals a bimodal pattern with a subset (26%) clustered and significantly nonrandomly distributed. A two-samples Wilcoxon test was used to calculate the p-value in R. **c,** Exons show a significant excess of amino-acid replacement (dN/dS) for both human–bonobo (red line, p-value= 0.004778) and human–chimpanzee ILS (blue line, p-value = 0.03924). In particular, mapping to the clustered segments (b) shows the most significant excess of amino acid replacements dN/dS (dotted purple line, p-value = 0.001015) when compared to the genome-wide null distribution (gray density plot). This shift is not observed for the non-clustered ILS segments (NC ILS, dotted black line, p-value = 0.3161). Significance performed using the one-sample t test in R.

Table 1.

Hominid genome-wide ILS estimates

Window size	Number of ILS segments		Percentage of ILS		Total ILS*	Genomic properties				
	(G,((B,H),C))	(G,((H,C),B))	(G,((B,H),C))	(G,((H,C),B))		GC*	Intergenic/Intragenic	Alu*	L1*	Exon*
20 kb	218	218	0.19	0.19	0.38	37.7	1.79	6.37	31.44	0.49
10 kb	1,143	1,138	0.49	0.48	0.97	38.39	1.73	7.35	27.08	0.47
5 kb	4,314	4,373	0.91	0.92	1.83	38.95	1.64	7.85	24.67	0.58
2 kb	18,218	18,334	1.52	1.53	3.05	39.58	1.49	8.71	21.51	0.72
1 kb	46,584	46,938	2.06	2.07	4.13	40.06	1.37	9.8	19.85	0.8
500 bp	102,197	103,338	2.52	2.55	5.07	40.54	1.33	11.24	18.66	0.75
Genome average						40.89	1.15	10.17	17.42	1.17

G, gorilla; B, bonobo; H, human; C, chimpanzee. (G,((B,H),C)) and (G,((H,C),B)) represent two different ILS topologies.

[*] content in (%); GC, Alu, L1 and Exon contents are based on the GRCh38 genome.