

# Complete sequencing of ape genomes

## Author list:

DongAhn Yoo<sup>1</sup>, Arang Rhie<sup>2</sup>, Prajna Hebbar<sup>3</sup>, Francesca Antonacci<sup>4</sup>, Glennis A. Logsdon<sup>1,5</sup>, Steven J. Solar<sup>2</sup>, Dmitry Antipov<sup>2</sup>, Brandon D. Pickett<sup>2</sup>, Yana Safonova<sup>6</sup>, Francesco Montinaro<sup>4,7</sup>, Yanting Luo<sup>8</sup>, Joanna Malukiewicz<sup>9</sup>, Jessica M. Storer<sup>10</sup>, Jiadong Lin<sup>1</sup>, Abigail N. Sequeira<sup>11</sup>, Riley J. Mangan<sup>12,13,14</sup>, Glenn Hickey<sup>3</sup>, Graciela Monfort Anez<sup>15</sup>, Parithi Balachandran<sup>16</sup>, Anton Bankevich<sup>6</sup>, Christine R. Beck<sup>10,16,17</sup>, Arjun Biddanda<sup>18</sup>, Matthew Borchers<sup>15</sup>, Gerard G. Bouffard<sup>19</sup>, Emry Brannan<sup>20</sup>, Shelise Y. Brooks<sup>19</sup>, Lucia Carbone<sup>21,22</sup>, Laura Carrel<sup>23</sup>, Agnes P. Chan<sup>24</sup>, Juyun Crawford<sup>19</sup>, Mark Diekhans<sup>3</sup>, Eric Engelbrecht<sup>25</sup>, Cedric Feschotte<sup>26</sup>, Giulio Formenti<sup>27</sup>, Gage H. Garcia<sup>1</sup>, Luciana de Gennaro<sup>4</sup>, David Gilbert<sup>28</sup>, Richard E. Green<sup>29</sup>, Andrea Guarracino<sup>30</sup>, Ishaan Gupta<sup>31</sup>, Diana Haddad<sup>32</sup>, Junmin Han<sup>33</sup>, Robert S. Harris<sup>11</sup>, Gabrielle A. Hartley<sup>10</sup>, William T. Harvey<sup>1</sup>, Michael Hiller<sup>34</sup>, Kendra Hoekzema<sup>1</sup>, Marlys L. Houck<sup>35</sup>, Hyeonsoo Jeong<sup>1,59</sup>, Kaivan Kamali<sup>11</sup>, Manolis Kellis<sup>12,13</sup>, Bryce Kille<sup>36</sup>, Chul Lee<sup>37</sup>, Youngho Lee<sup>38</sup>, William Lees<sup>25,39</sup>, Alexandra P. Lewis<sup>1</sup>, Qiuhui Li<sup>40</sup>, Mark Loftus<sup>41,42</sup>, Yong Hwee Eddie Loh<sup>43</sup>, Hailey Loucks<sup>3</sup>, Jian Ma<sup>44</sup>, Yafei Mao<sup>33,45,46</sup>, Juan F. I. Martinez<sup>6</sup>, Patrick Masterson<sup>32</sup>, Rajiv C. McCoy<sup>18</sup>, Barbara McGrath<sup>11</sup>, Sean McKinney<sup>15</sup>, Britta S. Meyer<sup>9</sup>, Karen H. Miga<sup>3</sup>, Saswat K. Mohanty<sup>11</sup>, Katherine M. Munson<sup>1</sup>, Karol Pal<sup>11</sup>, Matt Pennell<sup>47</sup>, Pavel A. Pevzner<sup>31</sup>, David Porubsky<sup>1</sup>, Tamara Potapova<sup>15</sup>, Francisca R. Ringeling<sup>48</sup>, Joana L. Rocha<sup>49</sup>, Oliver A. Ryder<sup>35</sup>, Samuel Sacco<sup>29</sup>, Swati Saha<sup>25</sup>, Takayo Sasaki<sup>28</sup>, Michael C. Schatz<sup>40</sup>, Nicholas J. Schork<sup>24</sup>, Cole Shanks<sup>3</sup>, Linnéa Smeds<sup>11</sup>, Dongmin R. Son<sup>50</sup>, Cynthia Steiner<sup>35</sup>, Alexander P. Sweeten<sup>2</sup>, Michael G. Tassia<sup>18</sup>, Françoise Thibaud-Nissen<sup>32</sup>, Edmundo Torres-González<sup>11</sup>, Mihir Trivedi<sup>1,59</sup>, Wenjie Wei<sup>51,52</sup>, Julie Wertz<sup>1</sup>, Muyu Yang<sup>44</sup>, Panpan Zhang<sup>26</sup>, Shilong Zhang<sup>33</sup>, Yang Zhang<sup>44</sup>, Zhenmiao Zhang<sup>31</sup>, Sarah A. Zhao<sup>12</sup>, Yixin Zhu<sup>47</sup>, Erich D. Jarvis<sup>37,53</sup>, Jennifer L. Gerton<sup>15</sup>, Iker Rivas-González<sup>54</sup>, Benedict Paten<sup>3</sup>, Zachary A. Szpiech<sup>11</sup>, Christian D. Huber<sup>11</sup>, Tobias L. Lenz<sup>9</sup>, Miriam K. Konkel<sup>41,42</sup>, Soojin V. Yi<sup>55</sup>, Stefan Canzar<sup>48</sup>, Corey T. Watson<sup>25</sup>, Peter H. Sudmant<sup>49,56</sup>, Erin Molloy<sup>57</sup>, Erik Garrison<sup>30</sup>, Craig B. Lowe<sup>8</sup>, Mario Ventura<sup>4</sup>, Rachel J. O'Neill<sup>10,17,58</sup>, Sergey Koren<sup>2</sup>, Kateryna D. Makova<sup>11,\*</sup>, Adam M. Phillippy<sup>2,\*</sup>, Evan E. Eichler<sup>1,59,\*</sup>

## Affiliations:

<sup>1</sup>Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA

<sup>2</sup>Genome Informatics Section, Center for Genomics and Data Science Research, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA

<sup>3</sup>UC Santa Cruz Genomics Institute, University of California Santa Cruz, Santa Cruz, CA 95060, USA

<sup>4</sup>Department of Biosciences, Biotechnology and Environment, University of Bari, Bari, 70124, Italy

- 40 <sup>5</sup>Department of Genetics, Epigenetics Institute, Perelman School of Medicine, University of  
41 Pennsylvania, Philadelphia, PA 19103, USA
- 42 <sup>6</sup>Computer Science and Engineering Department, Huck Institutes of Life Sciences, Pennsylvania  
43 State University, State College, PA 16801, USA
- 44 <sup>7</sup>Institute of Genomics, University of Tartu, Tartu, Estonia
- 45 <sup>8</sup>Department of Molecular Genetics and Microbiology, Duke University Medical Center,  
46 Durham, NC 27710, USA
- 47 <sup>9</sup>Research Unit for Evolutionary Immunogenomics, Department of Biology, University of  
48 Hamburg, 20146 Hamburg, Germany
- 49 <sup>10</sup>Institute for Systems Genomics, University of Connecticut, Storrs, CT 06269, USA
- 50 <sup>11</sup>Department of Biology, Penn State University, University Park, PA 16802, USA
- 51 <sup>12</sup>Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of  
52 Technology, Cambridge, MA 02139, USA
- 53 <sup>13</sup>The Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
- 54 <sup>14</sup>Genetics Training Program, Harvard Medical School, Boston, MA 02115, USA
- 55 <sup>15</sup>Stowers Institute for Medical Research, Kansas City, MO 64110, USA
- 56 <sup>16</sup>The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA
- 57 <sup>17</sup>Department of Genetics and Genome Sciences, University of Connecticut Health Center,  
58 Farmington, CT, USA
- 59 <sup>18</sup>Department of Biology, Johns Hopkins University, Baltimore, MD 21218, USA
- 60 <sup>19</sup>NIH Intramural Sequencing Center, National Human Genome Research Institute, National  
61 Institutes of Health, Bethesda, MD 20892, USA
- 62 <sup>20</sup>Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT, USA
- 63 <sup>21</sup>Department of Medicine, KCVI, Oregon Health Sciences University, Portland, OR, USA
- 64 <sup>22</sup>Division of Genetics, Oregon National Primate Research Center, Beaverton, OR, USA
- 65 <sup>23</sup>PSU Medical School, Penn State University School of Medicine, Hershey, PA, USA
- 66 <sup>24</sup>The Translational Genomics Research Institute, a part of the City of Hope National Medical  
67 Center, Phoenix, AZ, USA
- 68 <sup>25</sup>Department of Biochemistry and Molecular Genetics, School of Medicine, University of  
69 Louisville, Louisville, KY, USA
- 70 <sup>26</sup>Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853, USA
- 71 <sup>27</sup>Vertebrate Genome Laboratory, The Rockefeller University, New York, NY 10021, USA
- 72 <sup>28</sup>San Diego Biomedical Research Institute, San Diego, CA, USA
- 73 <sup>29</sup>University of California Santa Cruz, Santa Cruz, CA, USA
- 74 <sup>30</sup>Department of Genetics, Genomics and Informatics, University of Tennessee Health Science  
75 Center, Memphis, TN 38163, USA
- 76 <sup>31</sup>Department of Computer Science and Engineering, University of California San Diego, CA,  
77 USA
- 78 <sup>32</sup>National Center for Biotechnology Information, National Library of Medicine, National  
79 Institutes of Health, Bethesda, MD 20894, USA
- 80 <sup>33</sup>Bio-X Institutes, Key Laboratory for the Genetics of Developmental and Neuropsychiatric  
81 Disorders, Ministry of Education, Shanghai Jiao Tong University, Shanghai, China

- 82 <sup>34</sup>LOEWE Centre for Translational Biodiversity Genomics, Senckenberg Research Institute,  
83 Goethe University, Frankfurt, Germany
- 84 <sup>35</sup>San Diego Zoo Wildlife Alliance, Escondido, CA, 92027-7000, USA
- 85 <sup>36</sup>Department of Computer Science, Rice University, Houston, TX 77005, USA
- 86 <sup>37</sup>Laboratory of Neurogenetics of Language, The Rockefeller University, New York, NY, USA
- 87 <sup>38</sup>Laboratory of bioinformatics and population genetics, Interdisciplinary program in  
88 bioinformatics, Seoul National University, Republic of Korea
- 89 <sup>39</sup>Bioengineering Program, Faculty of Engineering, Bar-Ilan University, Ramat Gan, Israel
- 90 <sup>40</sup>Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA
- 91 <sup>41</sup>Department of Genetics & Biochemistry, Clemson University, Clemson, SC, USA
- 92 <sup>42</sup>Center for Human Genetics, Clemson University, Greenwood, SC, USA
- 93 <sup>43</sup>Neuroscience Research Institute, University of California, Santa Barbara, CA, USA
- 94 <sup>44</sup>Ray and Stephanie Lane Computational Biology Department, School of Computer Science,  
95 Carnegie Mellon University, PA, USA
- 96 <sup>45</sup>Center for Genomic Research, International Institutes of Medicine, Fourth Affiliated Hospital,  
97 Zhejiang University, Yiwu, Zhejiang, China
- 98 <sup>46</sup>Shanghai Jiao Tong University Chongqing Research Institute, Chongqing, China
- 99 <sup>47</sup>Department of Quantitative and Computational Biology, University of Southern California, Los  
100 Angeles, CA, USA
- 101 <sup>48</sup>Faculty of Informatics and Data Science, University of Regensburg, 93053 Regensburg,  
102 Germany
- 103 <sup>49</sup>Department of Integrative Biology, University of California, Berkeley, Berkeley, USA
- 104 <sup>50</sup>Department of Ecology, Evolution and Marine Biology, Neuroscience Research Institute,  
105 University of California, Santa Barbara, CA, USA
- 106 <sup>51</sup>School of Life Sciences, Westlake University, Hangzhou 310024, China
- 107 <sup>52</sup>National Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, 430070,  
108 Wuhan, China
- 109 <sup>53</sup>Howard Hughes Medical Institute, Chevy Chase, MD, USA
- 110 <sup>54</sup>Department of Primate Behavior and Evolution, Max Planck Institute for Evolutionary  
111 Anthropology, Leipzig, Germany
- 112 <sup>55</sup>Department of Ecology, Evolution and Marine Biology, Department of Molecular, Cellular and  
113 Developmental Biology, Neuroscience Research Institute, University of California, Santa  
114 Barbara, CA, USA
- 115 <sup>56</sup>Center for Computational Biology, University of California, Berkeley, Berkeley, USA
- 116 <sup>57</sup>Department of Computer Science, University of Maryland, College Park, MD 20742, USA
- 117 <sup>58</sup>Departments of Molecular and Cell Biology, UConn Storrs, CT, USA
- 118 <sup>59</sup>Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA
- 119 \*Correspondence to: [ee3@uw.edu](mailto:ee3@uw.edu), [adam.phillippy@nih.gov](mailto:adam.phillippy@nih.gov), [kdm16@psu.edu](mailto:kdm16@psu.edu)

120	<b>Table of Contents</b>	
121	<b>ABSTRACT</b> .....	<b>5</b>
122	<b>INTRODUCTION</b> .....	<b>5</b>
123	<b>RESULTS</b> .....	<b>6</b>
124	<b>Section I: Ape genome assembly and a pangenome resource</b> .....	<b>6</b>
125	<b>Section II. Resource improvement highlights</b> .....	<b>9</b>
126	<b>Sequence divergence..</b> .....	<b>9</b>
127	<b>Speciation time and incomplete lineage sorting (ILS).....</b>	<b>10</b>
128	<b>Gene annotation..</b> .....	<b>10</b>
129	<b>Repeat annotation and mobile element insertion (MEI) identification.</b> .....	<b>11</b>
130	<b>Selection and diversity.</b> .....	<b>12</b>
131	<b>Immunoglobulin and major histocompatibility complex (MHC) loci.</b> .....	<b>14</b>
132	<b>Epigenetic features.</b> .....	<b>17</b>
133	<b>Evolutionary rearrangements and serial ape inversions.</b> .....	<b>17</b>
134	<b>Structurally divergent and accelerated regions of mutation.</b> .....	<b>19</b>
135	<b>Section III. New genomic regions</b> .....	<b>21</b>
136	<b>Acrocentric chromosomes and nucleolar organizer regions.....</b>	<b>22</b>
137	<b>Centromere satellite evolution.....</b>	<b>25</b>
138	<b>Subterminal heterochromatin.</b> .....	<b>28</b>
139	<b>Lineage-specific segmental duplications and gene families.</b> .....	<b>29</b>
140	<b>DISCUSSION</b> .....	<b>32</b>
141	<b>DATA AVAILABILITY</b> .....	<b>34</b>
142	<b>CODE AVAILABILITY</b> .....	<b>35</b>
143	<b>ACRONYMS &amp; ABBREVIATIONS</b> .....	<b>35</b>
144	<b>COMPETING INTERESTS</b> .....	<b>36</b>
145	<b>ACKNOWLEDGMENTS</b> .....	<b>36</b>
146	<b>AUTHOR CONTRIBUTIONS</b> .....	<b>37</b>
147	<b>REFERENCES</b> .....	<b>39</b>
148		
149		

## 150 **ABSTRACT**

151 We present haplotype-resolved reference genomes and comparative analyses of six ape species,  
152 namely: chimpanzee, bonobo, gorilla, Bornean orangutan, Sumatran orangutan, and siamang. We  
153 achieve chromosome-level contiguity with unparalleled sequence accuracy (<1 error in 500,000  
154 base pairs), completely sequencing 215 gapless chromosomes telomere-to-telomere. We resolve  
155 challenging regions, such as the major histocompatibility complex and immunoglobulin loci,  
156 providing more in-depth evolutionary insights. Comparative analyses, including human, allow us  
157 to investigate the evolution and diversity of regions previously uncharacterized or incompletely  
158 studied without bias from mapping to the human reference. This includes newly minted gene  
159 families within lineage-specific segmental duplications, centromeric DNA, acrocentric  
160 chromosomes, and subterminal heterochromatin. This resource should serve as a definitive  
161 baseline for all future evolutionary studies of humans and our closest living ape relatives.

162

## 163 **INTRODUCTION**

164 High-quality sequencing of ape genomes has been a high priority of the human genetics and  
165 genomics community since the initial sequencing of the human genome in 2001<sup>1,2</sup>. Sequencing  
166 of these genomes is critical for reconstructing the evolutionary history of every base pair of the  
167 human genome—one of the grand challenges put forward to the genomics community after the  
168 release of the first draft of the Human Genome Project<sup>3</sup>. As a result, there have been numerous  
169 publications ranging from initial draft genomes to significant updates over the last two decades<sup>4-7</sup>.  
170 Due to the repetitive nature of ape genomes, however, complete assemblies have not been  
171 achieved. Current references lack sequence resolution of some of the most dynamic genomic  
172 regions, including regions corresponding to lineage-specific gene families.

173 Advances in long-read sequencing and new assembly algorithms were needed to overcome the  
174 challenge of repeats and finish the first complete, telomere-to-telomere (T2T) assembly of a  
175 human genome<sup>8,9</sup>. Using these same methods, we recently published six additional pairs of  
176 complete sex chromosomes from distinct branches of the ape phylogeny<sup>10</sup>. Although these initial  
177 projects targeted haploid chromosomes and required substantial manual curation, improved  
178 assembly methods now enable the complete assembly of diploid chromosomes<sup>11,12</sup>. Using these  
179 methods, we present here the complete, phased, diploid genomes of six ape species making all  
180 data and curated assemblies freely available to the scientific community. We organize the  
181 manuscript into three sections focused primarily on 1) finishing the genomes and the  
182 development of an ape pangenome, 2) the added value for standard evolutionary analyses, and  
183 3) providing new evolutionary insights into the previously unassembled regions. Although the  
184 interior of the ribosomal DNA (rDNA) arrays as well as some small portions of the largest  
185 centromeres remain unresolved, these genomes represent an order of magnitude improvement in  
186 quality over the prior ape references and are of equivalent quality to the T2T-CHM13 human  
187 reference. We propose that these assemblies will serve as the definitive references for all future  
188 studies involving human/ape genome evolution.

## 189 RESULTS

### 190 Section I: Ape genome assembly and a pangenome resource

191 Unlike previous reference genomes that selected female individuals for improved representation  
192 of the X chromosome<sup>4-7</sup>, we focused on male samples (**Table 1**) in order to fully represent both  
193 sex chromosomes<sup>10</sup> and provide a complete chromosomal complement for each species. Samples  
194 from two of the species, bonobo and gorilla, originated from parent–child trios (**Supplementary**  
195 **Note I**) facilitating phasing of parental haplotypes. For other samples where parental data were  
196 not available, deeper Hi-C datasets were used (**Table 1**) to achieve chromosome-scale phasing.  
197 For all samples, we prepared high-molecular-weight DNA and generated deep PacBio HiFi  
198 (high-fidelity; mean=90-fold sequence coverage) and ONT (Oxford Nanopore Technologies;  
199 mean=136.4-fold sequence coverage) sequence data (**Table Assembly S1**). For the latter, we  
200 specifically focused on producing at least 30-fold of ultra-long (UL > 100 kbp) ONT sequence  
201 data to scaffold assemblies across larger repetitive regions, including centromeres and segmental  
202 duplications (SDs). We applied Verkko<sup>11</sup> (v. 2.0)—a hybrid assembler that leverages the  
203 accuracy of HiFi data for generating the backbone of the assembly (**Methods, Supplementary**  
204 **Note II**); UL-ONT sequencing for repeat resolution, local phasing, and scaffolding; and Hi-C or  
205 trio data for chromosome-scale phasing of haplotypes into a fully diploid assembly. To serve as a  
206 reference genome, a haploid “primary” assembly was selected from the diploid assembly of each  
207 species. For the trios, the most accurate and complete haplotype was selected as the primary  
208 assembly (maternal vs. paternal), and for the non-trios, the most accurate and complete  
209 chromosome was selected for each chromosome pair. With convention, both sex chromosomes  
210 and the mitochondrial genome were also included in the primary assembly.

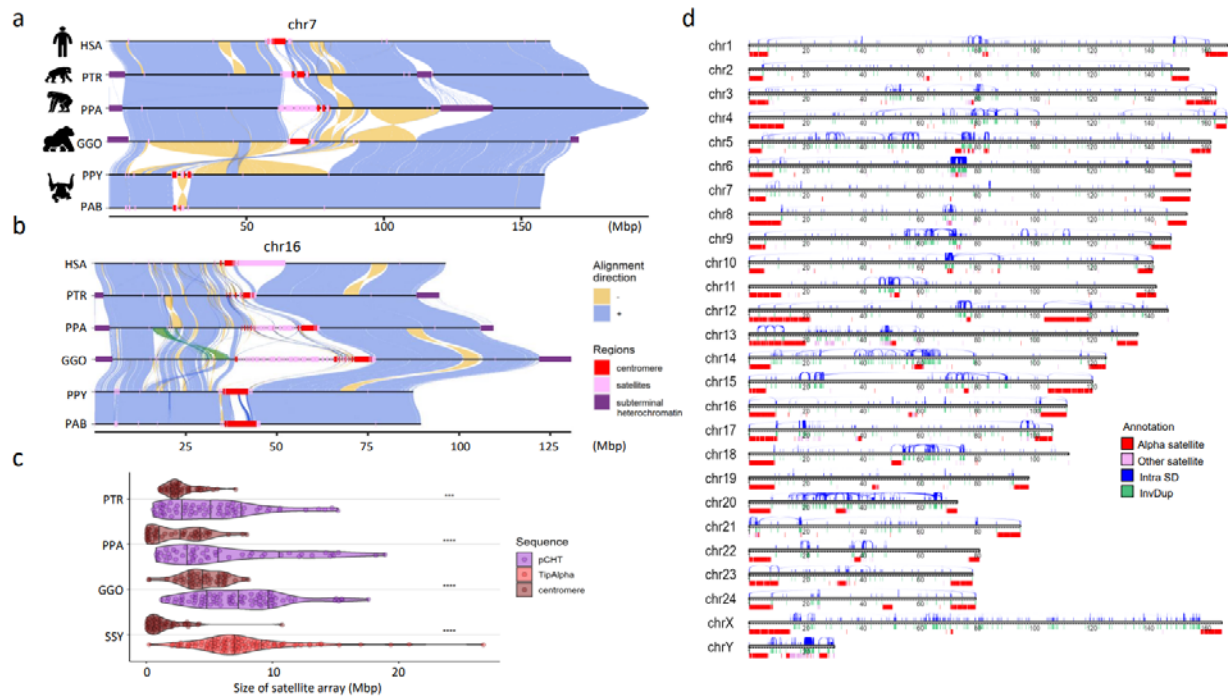
211 Considering the diploid genomes of each species, 74% (215/290) of all chromosomes are T2T  
212 assembled (gapless with telomere on both ends) and at least 80.8% of chromosomes are T2T in  
213 at least one haplotype (**Fig. 1, Table 1 & AssemblyS2**). Overall, there are an average of six gaps  
214 or breaks in assembly contiguity per haplotype (range=1-12), typically localized to the rDNA  
215 array, reducing to an average of 1.6 gaps if those acrocentric chromosomes are excluded. All  
216 assemblies were curated to extend partially into each rDNA array from both sides, ensuring that  
217 no non-rDNA sequence was missed. In addition to gaps, we searched specifically for collapses  
218 and misassemblies using dedicated methods (**Table AssemblyS3, Methods**). We estimate, on  
219 average, 1–2 Mbp of collapse per haplotype assembly and a wider range of potential  
220 misassemblies with an average of 0.2 to 11 Mbp flagged per haplotype assembly (**Table**  
221 **AssemblyS3**). Comparison with Illumina data<sup>13</sup> from the same samples provided a lower-bound  
222 accuracy of QV=49.3, limited by Illumina coverage loss in high-GC regions, while comparisons  
223 including HiFi data suggest even higher accuracy (QV=61.7; **Table AssemblyS1**). Overall, we  
224 estimate 99.2–99.9% of each genome is completely and accurately assembled, including  
225 heterochromatin. This is consistent with the T2T-CHM13v1.1 assembly, for which 0.3% of the  
226 genome was covered by known issues<sup>13</sup>. In short, these ape diploid genome assemblies represent  
227 an advance by at least one order of magnitude in terms of sequence accuracy and contiguity with  
228 respect to all prior ape genome assemblies<sup>4-7</sup>. For the first time, the centromeric regions, large  
229 blocks of SDs, and subterminal heterochromatin have been fully sequenced and assembled in

230 both haplotypes as well as more subtle improvements genome-wide. For example, a comparison  
 231 with previous genome assemblies for the same species shows an enrichment in sequence motifs  
 232 capable of forming non-canonical (non-B) DNA (A-phased, direct, mirror, inverted, and short  
 233 tandem repeats in particular) in newly gained regions of the new assemblies (**Table**  
 234 **AssemblyS10; Supplementary Note III**); such motifs have been shown to be difficult  
 235 sequencing targets<sup>14</sup> but are resolved here. Each genome assembly was annotated by NCBI and  
 236 has been adopted as the main reference in RefSeq, replacing the previous short- or long-read-  
 237 based, less complete versions of the genomes and updating the sex chromosomes with the newly  
 238 assembled and polished versions.

239 **Table 1: Summary of ape genome assemblies**

Sample information				Assembly stats (v2.0) Hap1 (Hap2) or mat (pat)					
Common name	Scientific name	Tissue	Sex	Accession	Total bases (Gbp)	config N50 (Mbp)	Number of T2T contigs	Number of non-rDNA gaps or missing telomeres	QV
Chimpanzee (PTR)	<i>Pan troglodytes</i>	Lymphoblastoid	M	PRJNA916736	3.136 (3.030)	146.29 (140.84)	19 (17)	0 (2)	66.0
Bonobo (PPA)*	<i>Pan paniscus</i>	Fibroblast/ Lymphoblastoid	M	PRJNA942951	3.206 (3.072)	147.03 (147.48)	20 (19)	0 (0)	62.7
Gorilla (GGO)*	<i>Gorilla gorilla</i>	Fibroblast	M	PRJNA942267	3.553 (3.350)	151.43 (150.80)	19 (22)	4 (0)	61.7
Bornean orangutan (PPY)	<i>Pongo pygmaeus</i>	Fibroblast	M	PRJNA916742	3.164 (3.048)	140.59 (137.91)	15 (13)	1 (2)	65.8
Sumatran orangutan (PAB)	<i>Pongo abelii</i>	Fibroblast	M	PRJNA916743	3.168 (3.077)	146.20 (140.60)	14 (13)	2 (3)	63.3
Siamang (SSY)	<i>Symphalangus syndactylus</i>	Lymphoblastoid	M	PRJNA916729	3.235 (3.122)	146.71 (144.67)	24 (20)	0 (5)	66.4
Average				-	3.244 (3.117)	146.38 (143.72)	18.5 (17.3)	1.2 (2)	64.3

240 \*Sample with parental sequence data. QV represents the score from Illumina/HiFi hybrid-based approach.



241  
 242 **Figure 1. Chromosomal-level assembly of complete great ape genomes. a)** A comparative ape  
 243 alignment of human (HSA) chromosome 7 with chimpanzee (PTR), bonobo (PPA), gorilla (GGO),  
 244 Bornean and Sumatran orangutans (PPY and PAB) shows a simple pericentric inversion in the *Pongo*  
 245 lineage (PPY and PAB) and **b)** HSA chromosome 16 harboring complex inversions. Each chromosome is  
 246 compared to the chromosome below in this stacked representation using the tool SVbyEye  
 247 (<https://github.com/daewoooo/SVbyEye>). Regions of collinearity and synteny (+/blue) are contrasted with  
 248 inverted regions (-/yellow) and regions beyond the sensitivity of minimap2 (homology gaps), including  
 249 centromeres (red), subterminal/interstitial heterochromatin (purple), or other regions of satellite expansion  
 250 (pink). A single transposition (green in panel b) relocates ~4.8 Mbp of gene-rich sequence in gorilla from  
 251 human chromosome 16p13.11 to human chromosome 16p11.2. **c)** Distribution of assembled satellite  
 252 blocks for centromere (alpha) and subterminal heterochromatin including, African great ape's pChT or  
 253 siamang's (SSY)  $\alpha$ -satellite, shows that subterminal heterochromatin are significantly longer in ape  
 254 species possessing both heterochromatin types (One-sided Wilcoxon ranked sum test; \*\*\*\*  $p < 0.0001$ ;  
 255 \*\*\*  $p < 0.001$ ). **d)** Schematic of the T2T siamang genome highlighting segmental duplications (Intra SDs;  
 256 blue), inverted duplications (InvDup; green), centromeric, subterminal and interstitial  $\alpha$ -satellites (red),  
 257 and other satellites (pink).

258

259 Human and nonhuman primate (NHP) genome assemblies are now comparable in quality,  
 260 helping to mitigate reference biases in alignment and variant discovery. We employed  
 261 Progressive Cactus<sup>15</sup> to construct 7-way (six primary and T2T-CHM13) 8-way (six primary ape  
 262 and two human haplotypes), and 16-way (diploid ape genomes including and four human  
 263 haplotypes) reference-free multiple genome alignments (**Supplementary Note IV; Data  
 264 Availability**). The more complete sequence and representation facilitates ancestral state  
 265 reconstruction for more genomic regions. For example, we annotated the human-primate  
 266 ancestral state of the GRCh38 reference genome by applying the parsimony-like method used by



267 the 1000 Genomes Project and Ensembl<sup>16</sup>. We observed a genome-wide increase of 6.25% in the  
268 total ancestrally annotated base pairs over the existing Ensembl annotation (release 112), with  
269 the greatest autosomal increase for chromosome 19 (21.48%; **Fig. 2a**). We annotated over 18  
270 million base pairs for chromosome Y, which is 4.67 times the annotated base pairs in the  
271 Ensembl annotation. Additionally, we find that the T2T annotation has more high-confidence  
272 bases in regions where the two annotations disagree most (**Fig. AncestralallelesS1**). We also  
273 constructed an interspecies 10-way pangenome representation of the ape genomes by Minigraph-  
274 Cactus<sup>17</sup>, using the ape and four human haplotypes (**Supplementary Note IV**). Compared to the  
275 recently released human pangenome from 47 individuals<sup>18</sup>, the resulting interspecies graph  
276 increases by ~3-fold the number of edges and nodes, resulting in a 3.38 Gbp ape “pan” genome.

277 As a second approach, we also applied pangenome graph builder (PGGB)<sup>19</sup> to construct all-to-all  
278 pairwise alignments for all 12 human primate haplotypes along with three T2T human  
279 haplotypes (T2T-CHM13v2.0 and T2T-HG002v1.0). We used these pairwise alignment data to  
280 construct an implicit graph (**Methods**) of all six species and computed a conservation score for  
281 every base pair in the genome (**Fig. PanGenomeS1; Methods**). The approach is transitive  
282 without a reference bias and considers both assembled haplotypes for each genome, as well as  
283 unique and repetitive regions, identifying the most rapidly evolving regions in each primate  
284 lineage (**Fig. PanGenomeS1**). We highlight the performance of this implicit graph in some of  
285 the most structurally diverse and dynamic regions of our genome, including the major  
286 histocompatibility complex (MHC) and the chromosome 8p23.1 inversion (**Fig. PanGenomeS1**).

287

## 288 **Section II. Resource improvement highlights**

289 **Sequence divergence.** The oft-quoted statistic of ~99% sequence identity between chimpanzee  
290 and human holds for most of the genome when considering single-nucleotide variants (SNVs)  
291 (**Fig. 2b**). However, comparisons of T2T genomes suggest a much more nuanced estimate.  
292 Examining the distribution of 1 Mbp aligned windows shows that the tail of that distribution is  
293 much longer with 12.5–27.3% of the genome failing to align or inconsistent with a simple 1-to-1  
294 alignment, especially within centromeres, telomeres, acrocentric regions, and SDs (**Figs. 1 & 2b**).  
295 We, therefore, considered SNV divergence separately from “gap” divergence, which considers  
296 poorly aligned sequences (**Methods**). Both parameters scale linearly with evolutionary time  
297 except for an inflated gorilla gap divergence (both between and within species comparisons) (**Fig.**  
298 **SeqDiv S1 & 2**). Gap divergence shows a 5- to 15-fold difference in the number of affected Mbp  
299 when compared to SNVs due to rapidly evolving and structural variant regions of the genome—  
300 most of which can now be fully accessed but not reliably aligned. As part of this effort, we also  
301 sequenced and assembled two pairs of closely related, congeneric ape species. For example, the  
302 Sumatran and Bornean orangutan species (the latter genome has not been sequenced previously)  
303 are the most closely related ape species, estimated to have diverged ~0.5–2 million years ago  
304 (mya)<sup>20-22</sup>. The autosome sequence identity of alignable bases between these two closely related  
305 orangutan genomes was 99.5% while the gap divergence was ~8.9% (autosomes). These

306 numbers are highly consistent with analyses performed using alternative alignment approaches  
307 (**Table SeqDiv. S1 & S2, Table OrangSeqDivS3; Supplementary Note V**).

308 **Speciation time and incomplete lineage sorting (ILS).** To jointly estimate speciation times (the  
309 minimum time at which two sequences can coalesce) and ancestral effective population sizes  
310 ( $N_e$ ), we modeled ILS across the ape species tree (**Table.ILS.S1**)<sup>23</sup>. Among the great apes  
311 (human, chimpanzee, gorilla and orangutan), our analyses date the human–chimpanzee split at  
312 5.5–6.3 mya, the African ape split at 10.6–10.9 mya, and the orangutan split at 18.2–19.6 mya  
313 (**Fig. 2e**). We infer ILS for an average of 39.5% of the autosomal genome and 24% of the X  
314 chromosome, representing an increase of approximately 7.5% compared to recent reports from  
315 less complete genomes<sup>24</sup> in part due to inclusion of more repetitive DNA (**Fig.ILS.S1**). We  
316 estimate that the human–chimpanzee–bonobo ancestral population (average  $N_e=198,000$ ) is  
317 larger than that of the human–chimpanzee–gorilla ancestor ( $N_e=132,000$ ), suggesting an increase  
318 of the ancestral population 6–10 mya. In contrast, the effective population sizes of more terminal  
319 branches are estimated to be smaller. For example, we estimate it is much smaller ( $N_e=46,800$ )  
320 in the *Pan* ancestor at 1.7 mya and *Pongo* ancestor ( $N_e=63,000$ ) at 0.93 mya, though these  
321 estimates should be taken with caution. For each terminal species branch, we infer the population  
322 size to range from 13,500 (*Pan troglodytes*) to 41,200 (*Symphalangus syndactylus*) (**Methods**).  
323 Compared to the autosomes, we find reduced X chromosome diversity for the African ape  
324 ancestor ( $N_e=115,600$ , X-to-A ratio of 0.87), and, more strongly, for the human–chimpanzee–  
325 bonobo ancestor ( $N_e=76,700$ , X-to-A ratio of 0.42). We additionally reconstruct a high-  
326 resolution, time-resolved ILS map (**Fig.ILS.S2**). T2T genomes support relatively high ILS  
327 estimates in previously inaccessible genomic regions, such as those encompassing the HLA  
328 genes (**Fig.ILS.S3**). Furthermore, multiple haplotypes for several species can also reveal cases of  
329 ancient polymorphism that have been sustained for thousands of years until present-day genomes,  
330 reflected in genomic regions with differential ILS patterns that depend on which combination of  
331 haplotypes are analyzed (**Fig.ILS.S3**).

332 **Gene annotation.** We applied two gene annotation pipelines (CAT and NCBI) to identify both  
333 protein-coding and noncoding RNA (ncRNA) genes for the primary assembly for each NHP. We  
334 complemented the annotation pipelines by direct mapping of Iso-Seq (50 Gbp of full-length non-  
335 chimeric [FLNC] cDNA) generated from each sample and searching for multi-exon transcripts.  
336 The number of protein-coding genes is very similar among different apes ( $N_e=22,114–23,735$ )  
337 with a little over a thousand genes predicted to be gained/duplicated or lost specifically per  
338 lineage (**Table. GeneS1**). Using the UCSC gene set, based on GENCODE<sup>25</sup>, we estimate that  
339 99.0–99.6% of corresponding human genes are now represented with >90% of genes being full-  
340 length. We identify a fraction (3.3–6.4%) of protein-coding genes present in the NHP T2T  
341 genomes that are absent in the human annotation set used. This includes 770–1,482 novel gene  
342 copies corresponding to 315–528 families in the NHPs with ~68.6% corresponding to lineage-  
343 specific SDs, all supported by Iso-Seq transcripts (**Table. GeneS1, S2**). In addition, 2.1%–5.2%  
344 of transcripts show novel NHP splice forms once again supported by Iso-Seq data (**Table.**  
345 **GeneS1**). We provide a unique resource in the form of a curated consensus protein-coding gene  
346 annotation set by integrating both the NCBI and CAT pipelines (**Methods**). Finally, we analyzed

347 FLNC reads obtained from testis from a second individual<sup>10</sup> to quantify the potential impact  
348 genome-wide on gene annotation and observed improvements in mappability, completeness, and  
349 accuracy (**Fig. Gene.S3 and Supplementary Note VII**). In gorilla, for example, we mapped  
350 28,925 (0.7%) additional reads to the T2T assembly in contrast to only 171 additional reads to  
351 the previous long-read assembly<sup>5</sup>. Similarly, we observed 33,032 (0.7%) soft-clipped reads  
352 (>200 bp) in the gorilla T2T assembly in contrast to 89,498 (2%) soft-clipped reads when  
353 mapping to the previous assembly<sup>5</sup>. These improvements in mappability are non-uniformly  
354 distributed with loci at centromeric, telomeric, and SD regions, leading to increased copy number  
355 counts when compared to previous genome assemblies (**Fig. Gene.S3e-g**).

356 **Repeat annotation and mobile element insertion (MEI) identification.** Based on  
357 RepeatMasker annotations (Dfam 3.7) and extensive manual curation<sup>26</sup> (**Methods;**  
358 **Supplementary Note VIII**), we generated a near-complete census of all high-copy repeats and  
359 their distribution across the ape genomes (**Table Repeat.S1; Extended data Table 2**). We now  
360 estimate that the autosomes of the ape genomes contain 53.21–57.99% detectable repeats, which  
361 include transposable elements (TEs), various classes of satellite DNA, variable number tandem  
362 repeats (VNTRs), and other repeats (**Fig. 2c**), significantly lower than the sex chromosomes (X  
363 [61.79–66.31%] and Y [71.14–85.94%])<sup>10</sup>. Gorilla, chimpanzee, bonobo, and siamang genomes  
364 show substantially higher satellite content driven in large part by the accumulation of  
365 subterminal heterochromatin through lineage-specific satellite and VNTR expansions (**Fig. 1,**  
366 **Fig. 2d, Extended data Table 2**). Satellites account for the largest repeat variation (**Extended**  
367 **data Table 2**), ranging from 4.94% satellite content in Bornean orangutan (159.2 Mbp total) to  
368 13.04% in gorilla (462.50 Mbp total). Analyzing gaps in exon and repeat annotations led to the  
369 identification of 159 previously unknown satellite monomers (**Table Repeat S2-S9**), ranging  
370 from 0.474 to 7.1 Mbp in additional base pairs classified per genome (**Fig. 2d**). Of note, 3.8 Mbp  
371 of sequence in the gorilla genome consists of a ~36 bp repeat, herein named VNTR\_148,  
372 accounting for only 841.9 kbp and 55.9 kbp in bonobo and chimpanzee, respectively (**Fig. 2d**).  
373 This repeat displays a pattern of expansion similar to that of the unrelated repeat  
374 pCht/subterminal satellite (StSat)<sup>10</sup>, suggesting it may have undergone expansion via a similar  
375 mechanism.

376 We find that 40.74% (gorilla) to 45.81% (Bornean orangutan) of genomes correspond to TEs  
377 (**Extended data Table 2; Table Repeat S1**). Leveraging the unaligned sequences in a 7-way  
378 Cactus alignment, we define a comprehensive set of both truncated and full-length, species-  
379 specific LINE, *Alu*, ERV, and SVA insertions for each ape species (**Table Repeat S12**).  
380 Orangutans appear to have the highest L1 mobilization rate based both on absolute number of  
381 insertions and the number of full-length elements with intact open reading frames (ORFs), while  
382 the African apes (gorilla, chimpanzee, bonobo, and human) show a higher accumulation of *Alu*  
383 insertions (**Fig. 2f, Supp Fig. Species Specific MEI 1**). The number of L1s with intact ORFs  
384 varies by a factor of 5.83, with chimpanzee having the lowest (95) and orangutans having the  
385 highest with at least 2.5 times more L1s with intact ORFs (more than 500 in both orangutan  
386 species compared to 203 in gorilla). Humans and gorillas fall in between this spectrum. The  
387 overall number and high percentage of full-length L1 elements with intact ORFs in orangutans

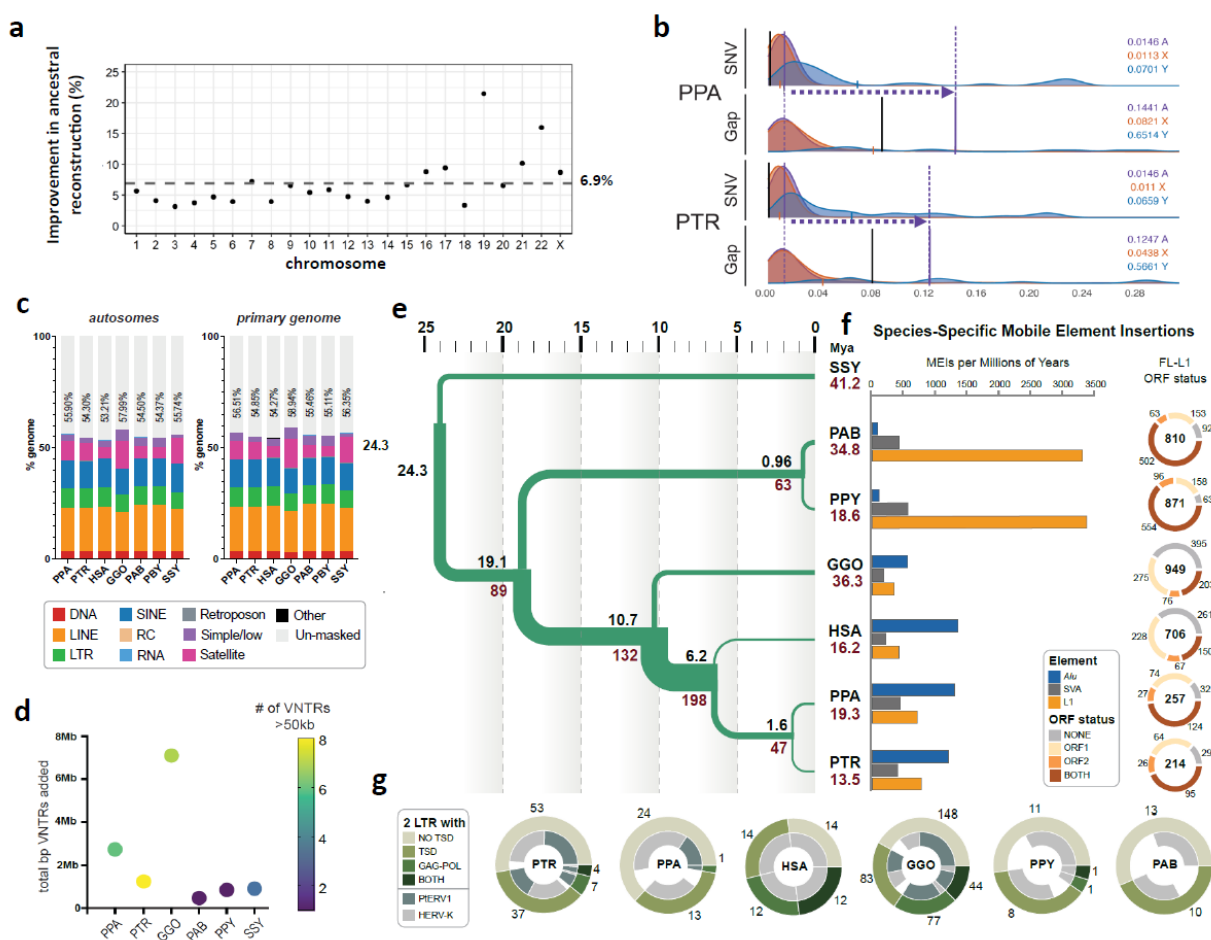
388 suggests recent high L1 activity. *Alu* activity is shown to be quiescent in the orangutans,  
 389 consistent with previous reports<sup>27</sup>, suggesting a genome environment where L1s out-compete *Alu*  
 390 retrotransposons. When considering only full-length ERV elements with both target site  
 391 duplications, and *gag* (capsid) and *pol* (reverse transcriptase and integrase) coding domains, a  
 392 striking difference is observed with higher full-length, species-specific ERV content in gorilla  
 393 (44), followed by human with 12, and chimpanzee with only three. PtERV and HERVK account  
 394 for the ERVs with both target site duplications and protein domains, along with more degraded  
 395 ERVs in gorilla, human, chimpanzee, and bonobo (**Fig. 2g; Table Repeat S13**). In addition to  
 396 MEIs, we also characterized the distribution of integrated NUMTs (nuclear sequences of  
 397 mitochondrial DNA origin) in ape genomes (**Methods**). We observe a substantial gain in the  
 398 number (3.7-10.5%) and total length of NUMTs (6.2-30%) (**Table Repeat S10**) over non-T2T  
 399 assemblies, with the largest gain observed for bonobo; Sumatran and Bornean orangutan species  
 400 differ in NUMT content by 73,990 bp despite their recent divergence.

401 **Extended data Table 2: Overview of repeat content (Mbp and percentage) in the ape genomes**

	Human		Chimpanzee		Bonobo		Gorilla		Sumatran orangutan		Bornean orangutan		Siamang	
	Mbp	%	Mbp	%	Mbp	%	Mbp	%	Mbp	%	Mbp	%	Mbp	%
<b>DNA</b>	108.59	3.48	110.60	3.48	110.74	3.47	111.82	3.15	111.20	3.41	110.83	3.44	106.76	3.27
<b>LINE</b>	633.55	20.33	636.62	20.04	638.42	19.99	649.75	18.33	686.42	21.06	683.56	21.22	631.83	19.37
<b>PLE</b>	0.06	0.00	0.07	0.00	0.07	0.00	0.07	0.00	0.07	0.00	0.07	0.00	0.07	0.00
<b>LTR</b>	272.91	8.75	274.90	8.65	275.67	8.63	277.85	7.84	279.36	8.57	278.28	8.64	256.49	7.86
<b>SINE</b>	393.51	12.62	391.01	12.30	393.08	12.3	397.19	11.20	396.85	12.17	395.47	12.28	402.73	12.34
<b>RC</b>	0.45	0.01	0.46	0.01	0.46	0.01	0.46	0.01	0.46	0.01	0.46	0.01	0.44	0.01
<b>Retroposon</b>	4.31	0.14	4.78	0.15	4.90	0.15	5.21	0.15	3.87	0.12	4.39	0.14	6.66	0.20
<b>Satellite</b>	161.77	5.19	251.68	7.92	270.19	8.46	462.50	13.04	191.45	5.87	159.20	4.94	376.70	11.55
<b>Simple/low</b>	108.38	3.48	69.17	2.18	107.68	3.37	181.36	5.11	134.09	4.11	138.98	4.31	54.29	1.66
<b>Other</b>	5.34	0.17	2.21	0.07	2.61	0.08	2.15	0.06	2.23	0.07	2.35	0.07	1.34	0.04
<b>RNA</b>	2.78	0.09	1.55	0.05	1.53	0.05	1.38	0.04	1.89	0.06	1.59	0.05	1.36	0.04
<b>Un-masked</b>	1,434.70	46.02	1,434.70	45.15	1,389.24	43.49	1,456.09	41.06	1,451.97	44.54	1,445.75	44.89	1,424.20	43.65
<b>Total masked</b>	1,691.65	54.27	1,743.04	54.85	1,805.35	56.51	2,089.74	58.94	1,807.88	55.46	1,775.19	55.11	1,838.69	56.35
<b>NUMT</b>	0.64	0.021	0.79	0.025	0.87	0.027	0.63	0.018	0.76	0.024	0.84	0.026	0.52	0.016

402  
 403 **Selection and diversity.** Using short-read whole-genome sequencing data generated from the  
 404 great ape genetic diversity project<sup>28</sup> (**Supplementary Note IX**) and mapped to the T2T genomes,  
 405 we searched for signatures of adaptation by identifying regions of hard<sup>29</sup> and soft (partial)<sup>30</sup>  
 406 selective sweeps in 10 great ape subspecies (**Methods**). Across all taxa, we identify 143 and 86  
 407 candidate regions for hard and partial selective sweeps, respectively, with only two overlapping  
 408 (**Table Selection S1**). Approximately 50% of hard (75/143) and 80% of partial selective (70/86)  
 409 sweeps are novel and a total of 43 regions overlap with sweeps previously found in humans<sup>31</sup>. As  
 410 expected, pathways related to diet (sensory perception for bitter taste, lipid metabolism, and iron  
 411 transport), immune function (antigen/peptide processing, MHC-I binding—strongest signal for

412 balancing selection), cellular activity, and oxidoreductase activity were enriched among bonobos,  
 413 central and eastern chimpanzees, and western lowland gorillas. While some of these findings are  
 414 confirmatory, the updated analysis provides remarkable precision. For example, among the well-  
 415 documented bitter taste receptor targets of selection<sup>32</sup>, we detect significant enrichment in  
 416 selection signals for such genes in bonobos (*TAS2R3*, *TAS2R4*, and *TAS2R5*) and western  
 417 lowland gorillas (*TAS2R14*, *TAS2R20*, and *TAS2R50*), as well as identified a bitter taste receptor  
 418 gene (*TAS2R42*) within a sweep region in eastern chimpanzees. Within the chimpanzee lineage,  
 419 it is notable that hard sweep regions in both eastern and central chimpanzees show significantly  
 420 greater differentiation ( $F_{ST} = 0.21$  and  $0.15$ , Mann-Whitney  $p < 0.001$ ) when compared to the  
 421 genome-wide average ( $F_{ST} = 0.09$ ). One of these regions was enriched (**Table Selection S3**) for  
 422 genes associated with epidermal differentiation (*KDF1* and *SFN*).



423  
 424 **Figure 2. Genome resource improvements.** **a**) Improvement in the ancestral allele inference by Cactus  
 425 alignment over the Ensembl/EPO alignment of the T2T ape genomes. **b**) Genome-wide distribution of  
 426 1 Mbp single-nucleotide variant (SNV)/gap divergence between human and bonobo (PPA)/chimpanzee  
 427 (PTR) genomes. The purple vertical lines represent the median divergence observed. The horizontal  
 428 dotted arrows highlight the difference in SNV vs. gap divergence. The black vertical lines represent the  
 429 median of allelic divergence within species. **c**) Total repeat content of ape autosomes and the primary  
 430 genome including chrX and Y. **d**) Total base pairs of previously unannotated VNTR satellite annotations

431 added per species. The color of each dot indicates the number of newly annotated satellites, out of 159,  
432 which account for more than 50 kbp in each assembly. (**Table Repeat S2**). **e**) Demographic inference.  
433 Black and red values refers to speciation times and effective population size ( $N_e$ ), respectively. For  $N_e$ ,  
434 values in inner branches refer to TRAILS estimates, while that of terminal nodes is predicted via msmc2,  
435 considering the harmonic mean of the effective population size after the last inferred split. **f**) (Left)  
436 Species-specific *Alu*, SVA and L1 MEI counts normalized by millions of years (using speciation times  
437 from (2e)). (Right) Species-specific Full-length (FL) L1 ORF status. The inner number within each circle  
438 represents the absolute count of species-specific FL L1s. **g**) Species-specific ERV comparison shows that  
439 the ERV increase in gorilla and chimpanzee lineages is due primarily to PTERV1 expansions.

440

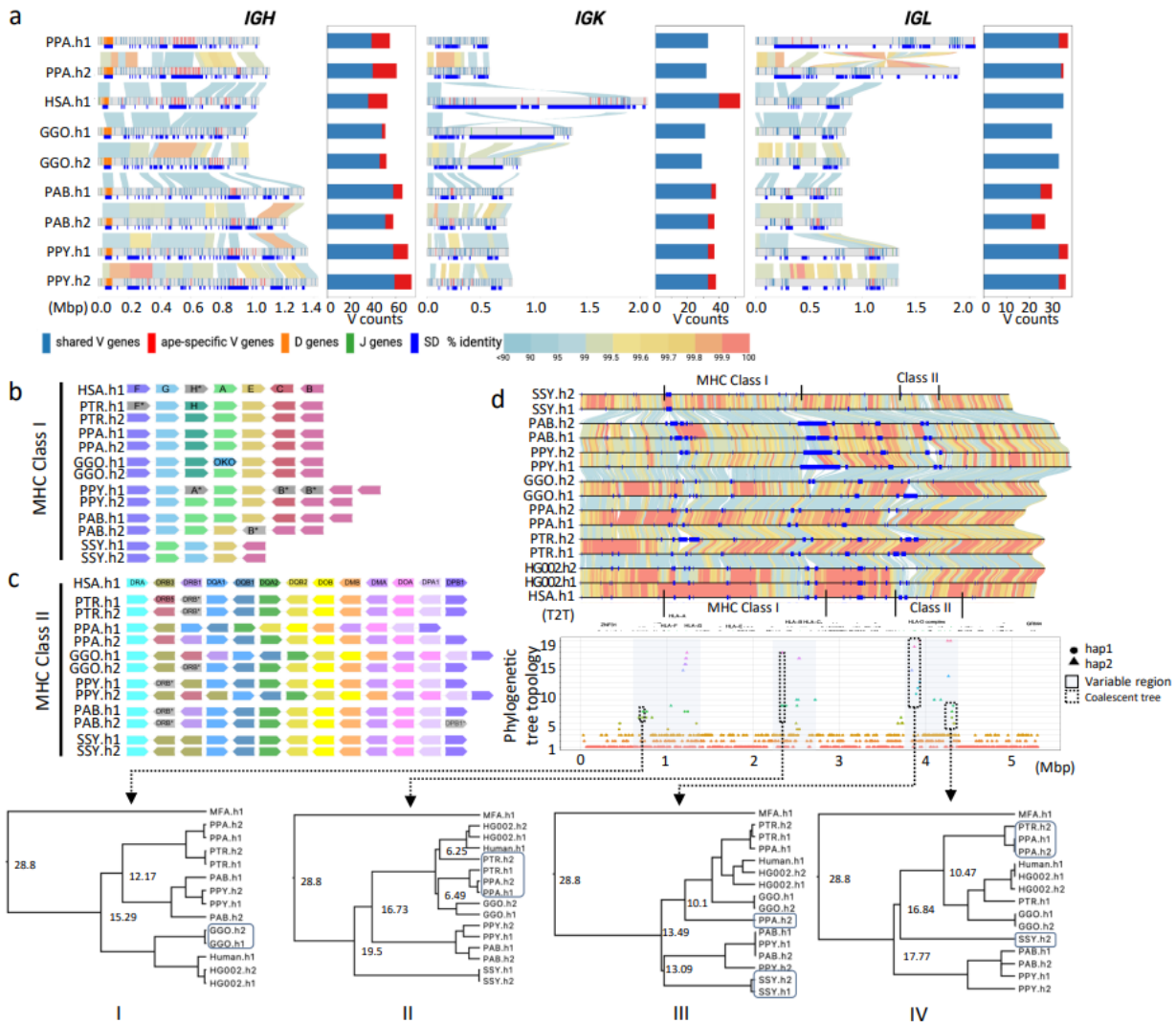
441 **Immunoglobulin and major histocompatibility complex (MHC) loci.** Complete ape genomes  
442 make it possible to investigate more thoroughly structurally complex regions known to have a  
443 high biomedical relevance, especially with respect to human disease. Importantly, four of the  
444 primate genomes sequenced and assembled here are derived from fibroblast (bonobo, gorilla and  
445 two orangutans as well as the human T2T reference) instead of lymphoblastoid cell lines. The  
446 latter, in particular, has been the most common source of most previous ape genome assemblies  
447 limiting characterization of loci subject to somatic rearrangement (e.g., VDJ genes)<sup>33</sup>. Thus, we  
448 specifically focused on nine regions associated with the immune response or antigen presentation  
449 that are subjected to complex mutational processes or selective forces.

450 *Immunoglobulin and T-cell receptor loci.* Antibodies, B-cell receptors, and T-cell receptors  
451 mediate interactions with both foreign and self-antigens and are encoded by large, expanded  
452 gene families that undergo rapid diversification both within and between species<sup>34,35</sup>. We  
453 conducted a comparative analyses of the immunoglobulin heavy chain (IGH), light chain kappa  
454 (IGK), and lambda (IGL) as well as T-cell receptor alpha (TRA), beta (TRB), gamma (TRG),  
455 and delta (TRD) loci in four ape species (**Supplementary Note X**) for which two complete intact  
456 haplotypes were constructed (**Fig. 3a, Fig.IG.S1a**). With respect to genes, we identify an  
457 average of 60 (IGHV), 36 (IGKV), 33 (IGLV), 46 (TRAV/TRDV), 54 (TRBV), and 8 (TRGV)  
458 putatively functional IG/TR V genes per parental haplotype per species across the seven loci (**Fig.**  
459 **3a, Fig.IG.S1a**); and provide an expanded set of curated IG/TR V, D, and J sequences for each  
460 species, including ORF genes (**Table.IG.S1 and Table.IG.S2**). The ape IG genes cluster into  
461 phylogenetic subfamilies similar to human (**Fig. IG.S1b**) but there are large structural  
462 differences between haplotypes within and between species, accounting for as much as 33% of  
463 inter-haplotype length differences in IG and up to 10% in the TR loci (**Fig.IG.S2ab**). IG loci  
464 show the most pronounced differences, including large structural changes and a 1.4 Mbp  
465 inversion distinguishing the two IGL haplotypes of bonobo (**Fig. 3a; Fig. IG.S2cd**). These large-  
466 scale differences frequently correspond to ape-specific genes (those that comprise  
467 phylogenetically distinct clades exclusive of human genes) (**Fig. 3a; Fig. IG.S2e**). We observe  
468 the greatest number of ape-specific genes within IGH (**Fig. 3a; Fig.IG.S2f**), where we note a  
469 greater density of SDs longer than 10 kbp relative to the other six loci (**Fig. IG.S2.g**).

470 *MHC loci.* We also completely assembled and annotated 12 ape haplotypes corresponding to the  
471 4–5 Mbp MHC region (**Supplementary Note XI**). The loci encode diverse cell surface proteins

472 crucial for antigen presentation and adaptive immunity<sup>36</sup>, are highly polymorphic among  
473 mammals<sup>37</sup>, and are strongly implicated in human disease via genome-wide association<sup>38</sup>.  
474 Comparative sequence analyses confirm extraordinary sequence divergence and structural  
475 variation (an average of 328 kbp deletions and 422 kbp insertions in apes compared to human),  
476 including duplications ranging from 99.3 kbp in siamang to 701 kbp in the Sumatran orangutan  
477 h2 (**Table MHC.S1-2**), as well as contractions and expansions associated with specific MHC  
478 genes (**Fig. 3b-c**). Overall, MHC class I genes show greater structural variation within and  
479 among the apes than MHC class II genes (**Fig. 3b-c**) with threefold greater average duplication  
480 sequences per haplotype (171 kbp vs. 62 kbp). Particularly high divergence in this region is seen  
481 in the siamang, which lacks a Syy-C locus and exhibits an inversion between the MHC-G and  
482 MHC-A loci compared to the great apes (**Fig. 3b, Fig. MHC.S1-S8**). While MHC I gene content  
483 and organization is nearly identical in human, bonobo, and chimpanzee, other apes show much  
484 more variation, including additional genes such as Gogo-OKO, related but distinct from Gogo-A  
485 (**Fig. 3b**)<sup>37</sup>. We observe expansion of MHC-A and MHC-B genes in both orangutan species (**Fig.**  
486 **3b, Fig. MHC.S6-S7**), with MHC-B being duplicated in both haplotypes of the two orangutan  
487 species while the MHC-A locus is only duplicated on one haplotype of each species. Similarly,  
488 both orangutan species show copy-number-variation of MHC-C, lacking on one haplotype but  
489 retaining it on the other (**Fig. 3b, Fig. MHC.S6-S7, Table MHC.S1**). All apes have a nearly  
490 identical set of MHC II loci with the exception of the *DRB* locus, which is known to exhibit  
491 copy-number-variation in humans<sup>39</sup>, and here shows the same pattern among the apes (**Fig. 3c,**  
492 **Fig. MHC.S1-S8, Table MHC.S2**). We also observe two cases where an MHC locus is present  
493 as a functional gene on one haplotype and as a pseudogene on the other haplotype (e.g., Gogo-  
494 DQA2 locus in gorilla and the Poab-DPB1 locus in Sumatran orangutan). Overall, this observed  
495 variation in MHC gene organization is consistent with long-term balancing selection<sup>39</sup>.

496 Given the deep coalescence of the HLA locus<sup>40</sup>, we performed a phylogenetic analysis with the  
497 complete ape sequences. We successfully constructed 1,906 trees encompassing 76% of the  
498 MHC region from the six ape species (**Fig. 3d**). We identify 19 distinct topologies (**Methods**)  
499 with three representing 96% (1,830/1,906) of the region and generally consistent with the species  
500 tree and predominant ILS patterns. The remaining 4% are discordant topologies that cluster  
501 within 200–500 kbp regions (**Table.MHC.S1**) corresponding to MHC I and II genes. We  
502 estimate coalescence times of these exceptional regions ranging from 10–24 mya (**Fig. 3d**).  
503 Finally, we performed genome-wide tests of selection as described above. We find that selection  
504 signatures and nucleotide diversity in the MHC region are among the top 0.1% genome-wide.  
505 These signatures confirm long-term balancing selection on MHC in multiple great ape lineages,  
506 including central and eastern chimpanzees, as well as at least two regions in MHC consistent  
507 with positive selection in bonobos and western chimpanzees<sup>40</sup>.



508

**Figure 3. IG and MHC genome organization in apes.** a) Annotated haplotypes of *IGH*, *IGK* and *IGL* loci across four primate species and one human haplotype (HSA.h1 or T2T-CHM13). Each haplotype is shown as a line in the genome diagram where the top part shows positions of shared V genes (blue), ape-specific V genes (red), D genes (orange), and J genes (green) and the bottom part shows segmental duplications (SDs) that were computed for a haplotype pair of the same species and depicted as dark blue rectangles. Human SDs were computed with respect to the GRCh38.p14 reference. Alignments between pairs in haplotypes are shown as links colored according to their percent identity values: from blue (<90%) through yellow (99.5%) to red (100%). The bar plot on the right from each genome diagram shows counts of shared and ape-specific V genes in each haplotype. b and c) show schematic representation of MHC locus organization for MHC-I and MHC-II genes, respectively, across the six ape haplotypes (PTR.h1/h2, PPA.h1/h2, GGO.h1/h2, PPY.h1/h2, PAB.h1/h2, SSY.h1/2) and human (HSA.h1). Only orthologs of functional human HLA genes are shown. Loci naming in apes follows human HLA gene names (HSA.h1), and orthologs are represented in unique colors across haplotypes and species. Orthologous genes that lack a functional coding sequence are grayed out and their name marked with an asterisk. One human HLA class I pseudogene (HLA-H) is shown, because functional orthologs of this gene were identified in some apes. d) Pairwise alignment of the 5.31 Mbp MHC region in the genome, with human gene annotations

16



525 and MHC-I and MHC-II clusters. Below is the variation in phylogenetic tree topologies according to the  
526 position in the alignment. The x-axis is the relative coordinate for the MHC region and the y-axis shows  
527 topology categories for the trees constructed. The three prominent sub-regions with highly discordant  
528 topologies are shown through shaded boxes. Four sub-regions (1-4) used to calculate coalescence times  
529 are shown with dashed boxes.

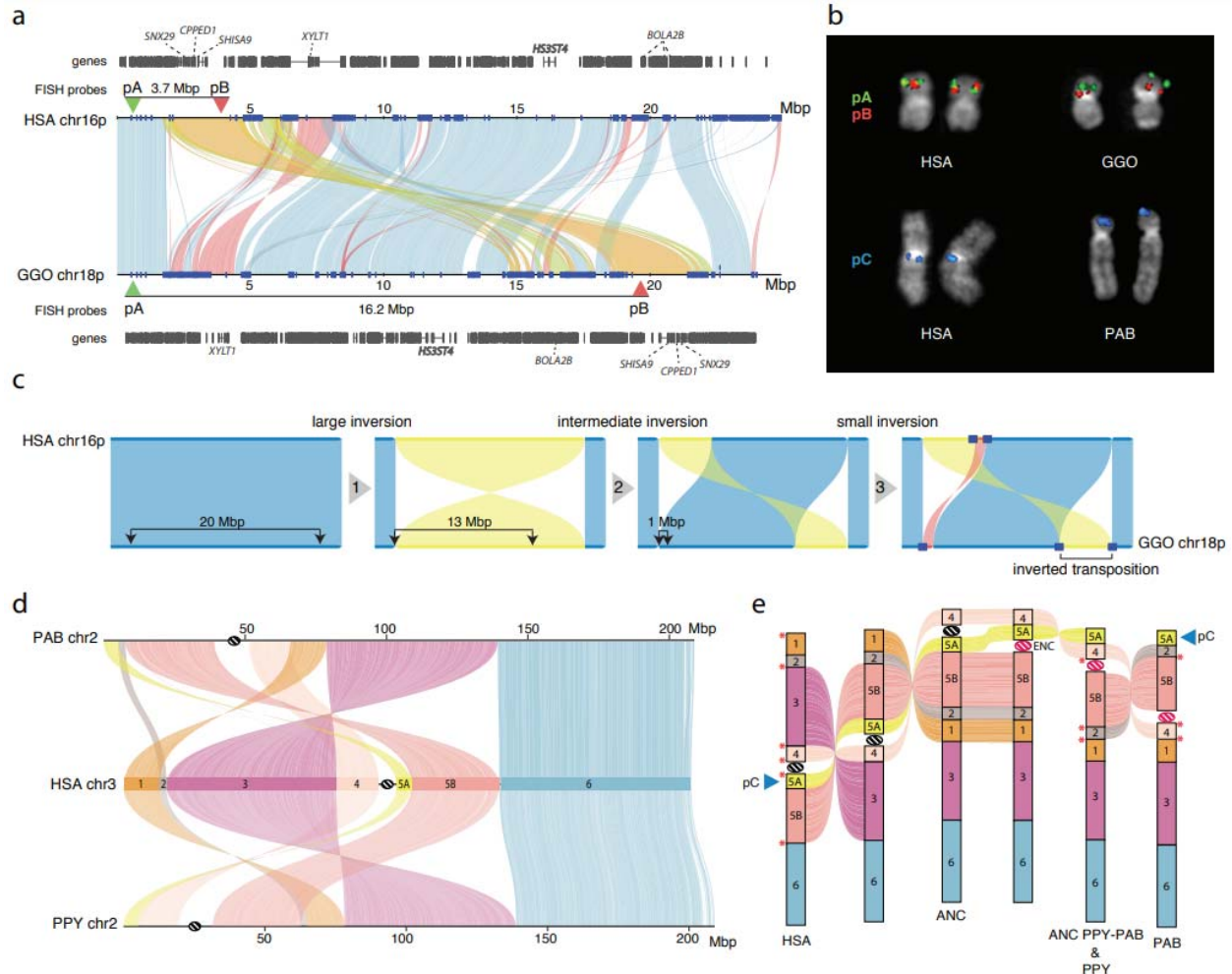
530

531 **Epigenetic features.** Using the T2T genomes, we also created a first-generation, multiscale  
532 epigenomic map of the apes, including DNA methylation, 3D chromatin organization, and DNA  
533 replication timing (**Supplementary Note XII-XIII**). The long-read sequencing data from  
534 individual ape species, for example, allowed us to construct a comparative map of 5-  
535 methylcytosine (5mC) DNA signatures for each ape genome sequenced here. We distinguish  
536 hypomethylated and hypermethylated promoters associated with gene expression and  
537 demonstrate that in each cell type, the majority (~83%) of promoters are consistently methylated  
538 (8,174 orthologous ape genes assessed) (**Table. MET.S1-2**). Specifically, we identify 1,997  
539 differentially methylated promoters (1,382 for fibroblast and 1,381 lymphoblast cell lines  
540 samples) as candidates for gene expression differences among the species (**Table. MET.S1-2**).  
541 Consistently methylated promoters were more lowly methylated, more highly expressed, and had  
542 a higher density of CpG sites compared to variably methylated promoters ( $P < 10^{-16}$  two-sided  
543 Mann–Whitney U test, **Fig. MET.S1**). These results highlight the interactions between sequence  
544 evolution and DNA methylome evolution with consequences on gene expression in ape  
545 genomes<sup>41,42</sup>. Additionally, we mapped Repli-seq, including previously collected NHP datasets<sup>43</sup>,  
546 to investigate evolutionary patterns of replication timing. We identified 20 states with different  
547 patterns of replication timing. Overall, the replication timing program is largely conserved, with  
548 53.1% of the genome showing conserved early and late replication timing across primates, while  
549 the remaining regions exhibit lineage-specific patterns (**Fig. RT.S1**) such as the very late  
550 replication pattern associated with heterochromatic caps in gorilla and chimpanzee. We inspected  
551 replication timing of SDs and found unique patterns for each type of lineage-specific SD, as  
552 shown in **Fig. RT.S2**.

553 **Evolutionary rearrangements and serial ape inversions.** Yunis and Prakash (1982)<sup>44</sup>  
554 originally identified 26 large-scale chromosomal rearrangements distinguishing humans and  
555 other great ape karyotypes, including translocation of gorilla chromosomes 4 and 19,  
556 chromosome 2 fusion in human, and 24 peri- and paracentric inversions. We completely  
557 sequenced and analyzed 43 of the breakpoints associated with these chromosomal  
558 rearrangements, with variable length resolutions (average=350 kbp with a maximum of ~700 kbp  
559 from previous cytogenetic mapping; **Table INV.S1 & Fig. INV.S1; Supplementary Note XIV**).  
560 These include six cases where the boundaries involving the centromere and/or the telomere are  
561 now fully resolved and additional cases where a more complex series of structural changes are  
562 suggested (**Fig. 4; Fig. INV.S1**). As an example, for human chromosome 3<sup>45</sup> we discovered an  
563 additional evolutionary rearrangement and inferred the occurrence of an evolutionary new  
564 centromere in the orangutan lineage (**Fig. 4e**). This increases the number of new centromere  
565 seedings to 10, making this chromosome a hotspot for neocentromeres in humans (15

566 documented cases in humans<sup>46,47</sup>). The next highest is chromosome 11, which has only four such  
567 events<sup>48</sup>.

568 During the finishing of the ape genomes, particularly the SDs flanking chromosomal  
569 evolutionary rearrangements, we noted several hundred smaller inversions and performed a  
570 detailed manual curation to catalog both homozygous and heterozygous events. Focusing on  
571 events larger than 10 kbp, we curate 1,140 interspecific inversions—522 are newly  
572 discovered<sup>7,20,44,48-68</sup> (**Table INV.S2**); 632 of the events are homozygous (found in both the  
573 assembled ape haplotypes) with remainder present in only one of the two ape haplotypes and,  
574 thus, likely polymorphisms. We also refine the breakpoints of 85/618 known inversions and  
575 identify several events that appeared to be the result of serial inversion events. In particular, we  
576 identify a 4.8 Mbp fixed inverted transposition on chromosome 18 in gorilla (**Fig. 4a-c**) that was  
577 incorrectly classified as a simple inversion but more likely to be explained by three consecutive  
578 inversions specific to the gorilla lineage transposing this gene-rich segment to 12.5 Mbp  
579 downstream (**Fig. 4a-c**). Similarly, the complex organization of orangutan chr2 can be explained  
580 through a model of serial inversions requiring three inversions and one centromere repositioning  
581 event (evolutionary neocentromere; ENC) to create PPY chromosome 2, and four inversions and  
582 one ENC for PAB (**Fig. 4d,e**). SDs map to seven out of eight inversion breakpoints. In total, 416  
583 inversions have an annotated gene mapping at least one of the breakpoints with 724 apparently  
584 devoid of protein-coding genes (**Table INV.S2**). Of these inversions, 63.5% (724/1140) have  
585 annotated human SDs at one or both ends of the inversion representing a significant 4.1-fold  
586 enrichment ( $p < 0.001$ ). The strongest predictable signal was for inverted SDs mapping to the  
587 breakpoints (6.2-fold;  $p < 0.001$ ) suggesting non-allelic homologous recombination driving many  
588 of these events. We also observed significant enrichment of novel transcripts (**Table Gene.S2**) at  
589 the breakpoints of the inversions of African great apes ( $p < 0.036$ ). Finally, we assigned  
590 parsimoniously >64% of homozygous inversions to the ape phylogeny (**Fig. INV.S3**) with the  
591 remaining inversions predicted to be recurrent. The number of inversions generally correlates  
592 with evolutionary distance ( $r^2=0.77$ ) with the greatest number assigned to the siamang lineage  
593 ( $n=44$ ). However, the human lineage shows fivefold less than that expected based on branch  
594 length and the trend still holds when using the Bornean orangutan as a reference instead of  
595 human.



596

597 **Figure 4. Great ape inversions and evolutionary rearrangements.** **a)** Alignment plot of gorilla chr18p  
 598 and human chr16p shows a 4.8 Mbp inverted transposition (yellow). SDs are shown with blue rectangles.

599 **b)** Experimental validation of the gorilla chr18 inverted transposition using FISH with probes pA  
 600 (CH276-36H14) and pB (CH276-520C10), which are overlapping in human metaphase chromosomes.

601 The transposition moves the red pB probe further away from the green pA probe in gorilla, resulting in  
 602 two distinct signals. FISH on metaphase chromosomes using probe pC (RP11-481M14) confirms the

603 location of a novel inversion to the p-ter of PAB chr2. **c)** An evolutionary model for the generation of the  
 604 inverted transposition by a series of inversions mediated by SDs. **d)** Alignment plot of orangutan

605 chromosome 2 homologs to human chromosome 3 highlights a more complex organization than  
 606 previously known by cytogenetics<sup>45</sup>: a novel inversion of block 5A is mapping at the p-ter of both chr2 in

607 PAB and PPY. **e)** A model of serial inversions requires three inversions and one centromere repositioning  
 608 event (evolutionary neocentromere; ENC) to create PPY chromosome 2, and four inversions and one

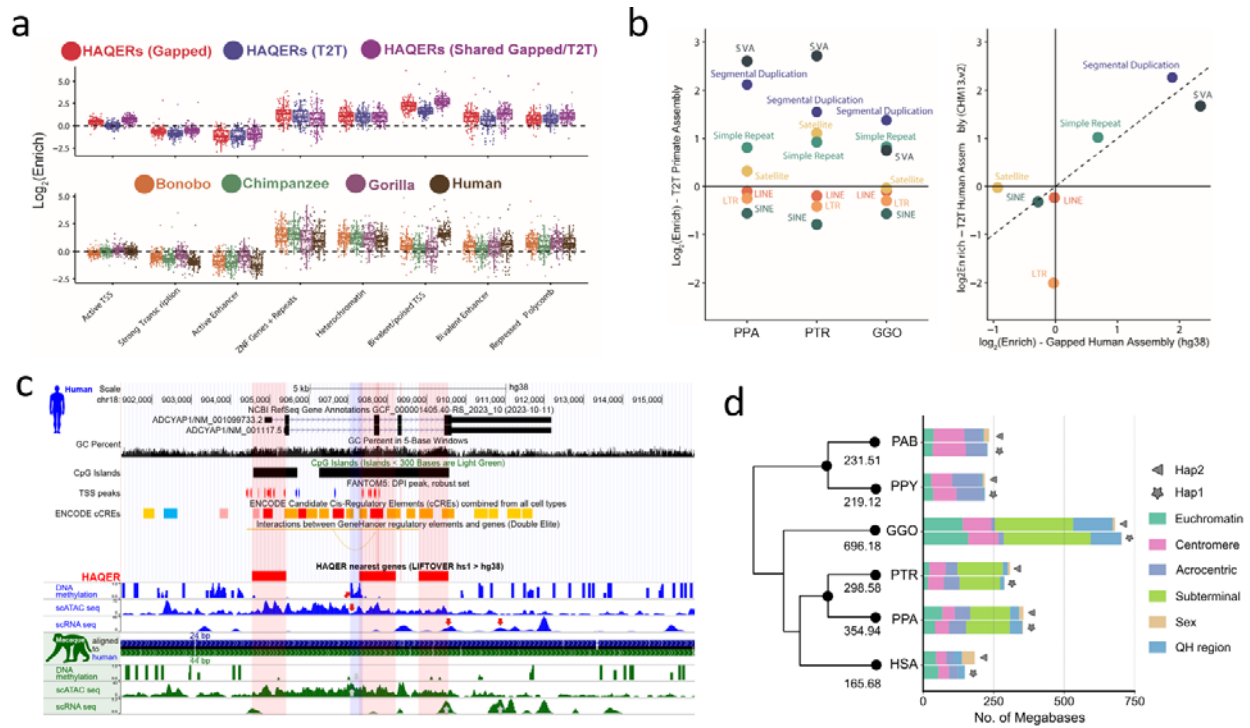
609 ENC for PAB. Red asterisks show the location of SDs mapping at the seven out of eight inversion  
 610 breakpoints.

611

612 **Structurally divergent and accelerated regions of mutation.** Previous studies have pinpointed  
 613 rapidly evolving regions associated with genes under positive selection<sup>69</sup> or cis-regulatory

614 elements (CREs) undergoing functional changes<sup>70</sup>. We utilized three strategies to systematically  
615 assess regions of accelerated mutation. First, was a bottom-up mutation-counting approach that  
616 identifies windows of ancestor quickly evolved regions (AQERs) based on sequence  
617 divergence<sup>70</sup> (**Methods; Supplementary Note XV**). We identified 14,210 AQER sites (**Table**  
618 **AQER**) across four primate lineages, including 3,268 on the human branch (i.e., HAQERs). Our  
619 analysis more than doubles the number of HAQERs identified from previous gapped primate  
620 assemblies ( $n=1,581$ ) (**Fig. 5a, Fig. AQER.S1**). Such elements are highly enriched in repetitive  
621 DNA, though not universally. With respect to MEIs, AQERs are depleted in SINEs, but enriched  
622 within the VNTRs of hominin-specific SVA elements (**Fig. 5b**). Additionally, HAQERs also  
623 exhibit a significant enrichment for bivalent chromatin states (repressing and activating  
624 epigenetic marks) across diverse tissues, with the strongest enrichment being for the bivalent  
625 promoter state ( $p < 1e-35$ ) (**Fig. 5a; Table AQER.S1**)—a signal not observed among other apes  
626 likely due to the chromatin states being called from human cells and tissues (**Supplementary**  
627 **Note**). An example of a human-specific HAQER change includes an exon and a potential CRE in  
628 the gene *ADCYAP1*, in the layer 5 extratelencephalic neurons of primary motor cortex. This gene  
629 shows convergent downregulation in human speech motor cortex and the analogous songbird  
630 vocal learning layer 5 type extratelencephalic neurons necessary for speech and song  
631 production<sup>71,72</sup>. We find here downregulation in layer 5 neurons of humans relative to macaques  
632 (RNA-seq) and an associated unique human epigenetic signature (hyper methylation and  
633 decreased ATAC-Seq) in the middle HAQER region of the gene that is not observed in the same  
634 type of neurons of macaque, marmoset, or mouse (**Fig. 5c, Fig. AQER.S2-S3**).

635 The second approach applied a top-down method that leveraged primate genome-wide all-by-all  
636 alignments to identify larger structurally divergent regions (SDRs) flanked by syntenic regions  
637 (**Methods; Supplemental Note XVI**) (Mao et al, 2024). We identified an average of 327 Mbp of  
638 SDRs per great ape lineage (**Fig. 5d**). SDRs delineate known sites of rapid divergence, including  
639 centromeres and subterminal heterochromatic caps but also numerous gene-rich SD regions  
640 enriched at the breakpoints of large-scale rearrangements (**Fig. SDR.S1**). The third approach  
641 used a gene-based analysis (TOGA—Tool to infer Orthologs from Genome Alignments) that  
642 focuses on the loss or gain of orthologous sequences in the human lineage (**Supplementary Note**  
643 **XVII**)<sup>73</sup>. TOGA identified six candidate genes from a set of 19,244 primate genes as largely  
644 restricted to humans (absent in >80% of the other apes; **Table TOGA.S1**). Among the candidate  
645 genes is a processed gene, *FOXO3B*, (present in humans and gorillas) whose paralog, *FOXO3A*,  
646 has been implicated in human longevity<sup>74</sup>. While the *FOXO3B* is expressed, its study has been  
647 challenging because it is embedded in a large, highly identical SD mediating Smith-Magenis  
648 deletion syndrome (**Fig. TOGA.S1**). While extensive functional studies will be required to  
649 characterize the hundreds of candidates we identified, we generated an integrated genomic  
650 (**Table SDR.S1**) and genic (**Table SDR.S2**) callset of accelerated regions for future investigation.



651  
 652 **Figure 5. Divergent regions of the ape genomes.** **a)** HAQER (human ancestor quickly evolved region)  
 653 sets identified in gapped (GRCh38) and T2T assemblies show enrichments for bivalent gene regulatory  
 654 elements across 127 cell types and tissues, with the strongest enrichment observed in the set of HAQERs  
 655 shared between the two analyses (top). The tendency for HAQERs to occur in bivalent regulatory  
 656 elements (defined using human cells and tissues) is not present in the sets of bonobo, chimpanzee, or  
 657 gorilla AQERs (ancestor quickly evolved regions; bottom). **b)** AQERs are enriched in SVAs, simple  
 658 repeats, and SDs, but not across the general classes of SINEs, LINEs, and LTRs (left). With T2T genomes,  
 659 the set of HAQERs defined using gapped genome assemblies became even more enriched for simple  
 660 repeats and SDs (right). **c)** HAQERs in a vocal learning-associated gene, *ADCYAP1* (adenylate cyclase  
 661 activating polypeptide 1), are marked as containing alternative promoters (TSS peaks of the FANTOM5  
 662 CAGE analysis), candidate cis-regulatory elements (ENCODE), and enhancers (ATAC-Seq peaks). For  
 663 the latter, humans have a unique methylated region in layer 5 extra-telencephalic neurons of the primary  
 664 motor cortex. Tracks are modified from the UCSC Genome Browser<sup>75</sup> above the HAQER annotations and  
 665 the comparative epigenome browser<sup>76</sup> below the HAQER annotations. **d)** Lineage-specific structurally  
 666 divergent regions (SDRs). SDRs are detected on two haplotypes and classified by different genomic  
 667 content. The average number of total bases was assigned to the phylogenetic tree.

668

### 669 Section III. New genomic regions

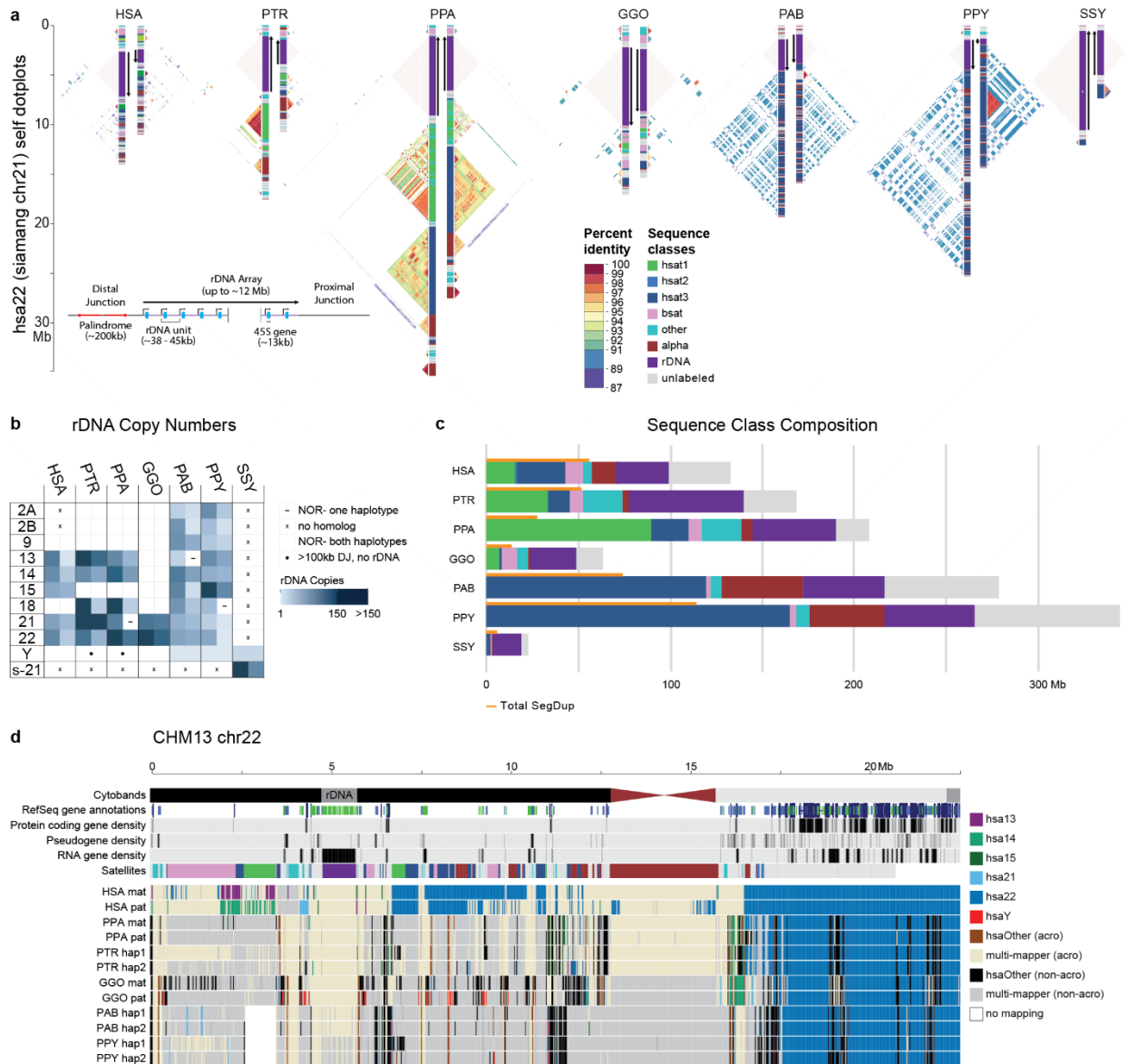
670 In addition to these improved insights into genes, repeats, and diversity, the contiguity afforded  
 671 by the complete genomes allowed regions typically excluded from both reference genomes and  
 672 evolutionary analyses to be investigated more systematically. We highlight four of the most  
 673 notable: acrocentric, centromeres, subterminal heterochromatic caps, and lineage-specific SDs.

674 **Acrocentric chromosomes and nucleolar organizer regions.** The human acrocentric  
675 chromosomes (13, 14, 15, 21, 22) are the home of nucleolar organizer regions (NORs) and  
676 encode ribosomal RNA (rRNA) components of the 60S and 40S subunits. The precise sequence  
677 of the human NORs and the surrounding heterochromatin of the short arms was only recently  
678 elucidated in the first T2T human genome<sup>8</sup>. Human acrocentric chromosomes typically contain a  
679 single NOR with a head-to-tail rDNA array that is uniformly transcribed in the direction of the  
680 centromere. Each NOR is preceded by a distal junction (DJ) region extending approximately 400  
681 kbp upstream of the rDNA array and including a 230 kbp palindrome (**Fig. Acro S1;**  
682 **Supplemental Note XVIII**) that encodes a long ncRNA associated with nucleolar function<sup>77</sup>. A  
683 variable patchwork of satellites and SDs flank the NOR, where heterologous recombination is  
684 thought to occur, as well as within the rDNA array itself, to maintain NOR homology through  
685 the action of concerted evolution<sup>78</sup>.

686 One conspicuous observation confirmed by our assemblies is that the ape NORs exist on  
687 different chromosomes for each species (**Fig. 6b, Fig. Acro.S2**). For example, HSA15 is NOR-  
688 bearing (NOR+) in human but not in chimpanzee or bonobo (NOR-), while HSA18 is NOR+ in  
689 both chimpanzee and bonobo, but NOR- in human<sup>79</sup> (**Fig. 6b, Fig. Acro.S2**). Among great apes,  
690 we find the total number of NORs per haploid genome varies from as few as two in gorilla to 10  
691 in both orangutans, while the siamang maternal genome sequenced here harbors only a single  
692 NOR (**Fig. 6b, Fig. Acro.S2**). We also find NORs on both orangutan and siamang Y  
693 chromosomes<sup>10</sup>, and partial DJ fragments on the chimpanzee and bonobo chrY (**Fig. 6b, Fig.**  
694 **Acro.S2**), suggesting their ancestral chrY may have been NOR+. Except for rRNA genes, all ape  
695 NOR-bearing chromosome short arms appear to be satellite-rich and gene-poor (**Fig. 6c,d**), with  
696 the NORs restricted to the end of an autosomal short arm or the end of a Y chromosome long  
697 arm. We identify, however, multiple acrocentric chromosomes with heterochromatic sequence  
698 on their short arm, but without an NOR (e.g., gorilla HSA2A, HSA9, HSA13, HSA15, and  
699 HSA18). Unlike the NOR+ acrocentrics, these NOR- acrocentrics carry multiple predicted  
700 protein-coding genes on their short arms. Thus, short-arm heterochromatin is strongly associated  
701 with ape NORs though not always predictive of their presence.

702 Estimated rDNA copy number for ape arrays varies from 1 on chrY of Bornean orangutan to 287  
703 on HSA21 of chimpanzee; total diploid rDNA copy number similarly varies from 343 in siamang  
704 to 1,142 in chimpanzee (**Methods, Fig.6b, Fig. Acro.S2, Table Acro S3**), with total rDNA copy  
705 number varying widely between individual haplotypes of the same species, as expected<sup>80</sup>.  
706 Heterozygous NOR loss was observed in bonobo (HSA21), Sumatran orangutan (HSA13), and  
707 Bornean orangutan (HSA18), all of which were mediated by a truncation of the chromosome  
708 prior to the typical NOR location (**Fig. Acro.S2**). The structure and composition of both  
709 satellites and SDs varies considerably among the apes (**Fig.6a,c**). The orangutan acrocentrics are  
710 dominated by HSat3 and  $\alpha$ -satellite, compared to the more balanced satellite composition of the  
711 other apes. Gorilla is notable for the presence of double NORs on both haplotypes of HSA22,  
712 with the additional NORs inverted relative to the first and including a complete DJ but only a  
713 single, inactive rDNA unit (**Fig. Acro.S3**).

714 At the chromosome level, the high level of synteny on the long arms of the NOR+ chromosomes  
715 quickly degrades when transitioning to the short arm, with almost no sequence aligning uniquely  
716 between different ape species (**Fig. 6d**). Even the haplotypes of a single human genome aligned  
717 best to different reference chromosomes on their distal ends, supporting prior observations of  
718 extensive heterologous recombination<sup>78</sup>. Despite their widespread structural variation, the ape  
719 NOR+ chromosomes share common features such as homogeneous rDNA arrays containing  
720 highly conserved rRNA genes. We extracted representative rDNA units from each assembly to  
721 serve as a reference for each species and confirmed a similar sequence structure, including the  
722 presence of a central microsatellite region within the intergenic spacer sequence for all species  
723 (**Fig. Acro.S4**), but with relatively high nucleotide substitution rates outside of the >99%  
724 identical 18S and 5.8S coding regions<sup>81</sup>. Despite its conserved co-linear structure, nucleotide  
725 identity of the intergenic spacer varied from 95.19% for human versus bonobo to just 80.60% for  
726 human versus siamang (considering only SNVs, **Table Acro S2**). The DJ sequence was found to  
727 be conserved across all great apes and present as a single copy per NOR, including the  
728 palindromic structure typical of the human DJ, with the exception of siamang, which contains  
729 only one half of the palindrome on each haplotype but in opposite orientations (**Fig. Acro.S5**).  
730 The transcriptional direction of all rDNA arrays is consistent within each species, with the  
731 chimpanzee and bonobo arrays inverted relative to human (**Fig. 6a**). This inversion includes the  
732 entire DJ sequence, confirming a prior FISH analysis that found the chimpanzee DJ had been  
733 relocated to the centromeric side of the rDNA array<sup>77</sup>. Our comparative analysis supports the DJ  
734 as a functional component of ape NORs that is consistently positioned upstream of rRNA gene  
735 transcription, rather than distally on the chromosome arm.



736

737 **Figure 6. Organization and sequence composition of the ape acrocentric chromosomes.** a) Sequence  
 738 identity heatmaps and satellite annotations for the NOR+ short arms of both HSA22 haplotypes across all  
 739 the great apes, and siamang chr21 (the only NOR+ chromosome in siamang) drawn with ModDotPlot<sup>82</sup>.  
 740 The short arm telomere is oriented at the top of the plot, with the entirety of the short arm drawn to scale  
 741 up to but not including the centromeric  $\alpha$ -satellite. Heatmap colors indicate self-similarity within the  
 742 chromosome, and large blocks indicate tandem repeat arrays (rDNA and satellites) with their  
 743 corresponding annotations given in between. Human is represented by the diploid HG002 genome.  
 744 b) Estimated number of rDNA units per haplotype (hap) for each species. HSA numbers are given in the  
 745 first column, with the exception of “s-21” for siamang chr21, which is NOR+ but has no single human  
 746 homolog. c) Sum of satellite and rDNA sequence across all NOR+ short arms in each species. “unlabeled”  
 747 indicates sequences without a satellite annotation, which mostly comprise SDs. Total SD bases are given  
 748 for comparison, with some overlap between regions annotated as SDs and satellites. d) Top tracks: chr22  
 749 in the T2T-CHM13v2.0 reference genome displaying various gene annotation metrics and the satellite  
 750 annotation. Bottom tracks: For each primate haplotype, including the human HG002 genome, the



751 chromosome that best matches each 10 kbp window of T2T-CHM13 chr22 is color coded, as determined  
752 by MashMap<sup>83</sup>. On the right side of the centromere (towards the long arm), HSA22 is syntenic across all  
753 species; however, on the short arm synteny quickly degrades, with very few regions mapping uniquely to  
754 a single chromosome, reflective of extensive duplication and recombination on the short arms. Even the  
755 human HG002 genome does not consistently align to T2T-CHM13 chr22 in the most distal (left-most)  
756 regions.

757

758 **Centromere satellite evolution.** The assembly of five nonhuman great ape genomes allowed us  
759 to assess the sequence, structure, and evolution of centromeric regions at base-pair resolution for  
760 the first time. Using these assemblies, we identify 227 contiguous centromeres out of a possible  
761 230 centromeres across five NHPs, each of which were composed of tandemly repeating  $\alpha$ -  
762 satellite DNA organized into higher-order repeats (HORs) belonging to one or more  $\alpha$ -satellite  
763 suprachromosomal families (SFs) (**Fig. 7; Fig. CEN.S1; Supplemental Note XIX**). In specific  
764 primate lineages, different SFs have risen to high frequency, such as SF5 in the orangutan and  
765 SF3 in the gorilla. We carefully assessed the assembly of each of these centromeres, checking for  
766 collapses, false duplications and misjoins (**Methods**), and found that approximately 85% of  
767 bonobo, 69% of chimpanzee, 54% of gorilla, 63% of Bornean orangutan, but only 27% of  
768 Sumatran orangutan centromeres are complete and correctly assembled (**Fig. 7a; Fig. CEN.S1**).  
769 Most of the assembly errors are few (~2 per centromere haplotype, on average) and typically  
770 involve a few 100 kbp of centromere satellite sequence that still need further work to resolve.

771 Focusing on the completely assembled centromeres, we identify several unique characteristics  
772 specific to each primate species. First, we find that the bonobo centromeric  $\alpha$ -satellite HOR  
773 arrays are, on average, 0.65-fold the length of human centromeric  $\alpha$ -satellite HOR array and  
774 0.74-fold the length of its sister species, chimpanzee (**Fig. 7b**). A closer examination of bonobo  
775  $\alpha$ -satellite HOR array lengths reveals that they are bimodally distributed, with approximately half  
776 of the bonobo centromeres (27/48) having an  $\alpha$ -satellite HOR array with a mean length of 110  
777 kbp (range: 15–674 kbp) and the rest (21/48) having a mean length of 3.6 Mbp (range: 1.6–6.7  
778 Mbp; **Fig. 7c**). The bimodal distribution persists in both sets of bonobo haplotypes. This >450-  
779 fold variation in bonobo  $\alpha$ -satellite HOR array length has not yet been observed in any other  
780 primate species and implies a wide range of centromeric structures and sizes compatible with  
781 centromere function. Indeed, no “mini-centromere” arrays have been observed in the chimpanzee,  
782 despite its recent speciation from bonobo (~1.7 mya; **Fig. 7d**).

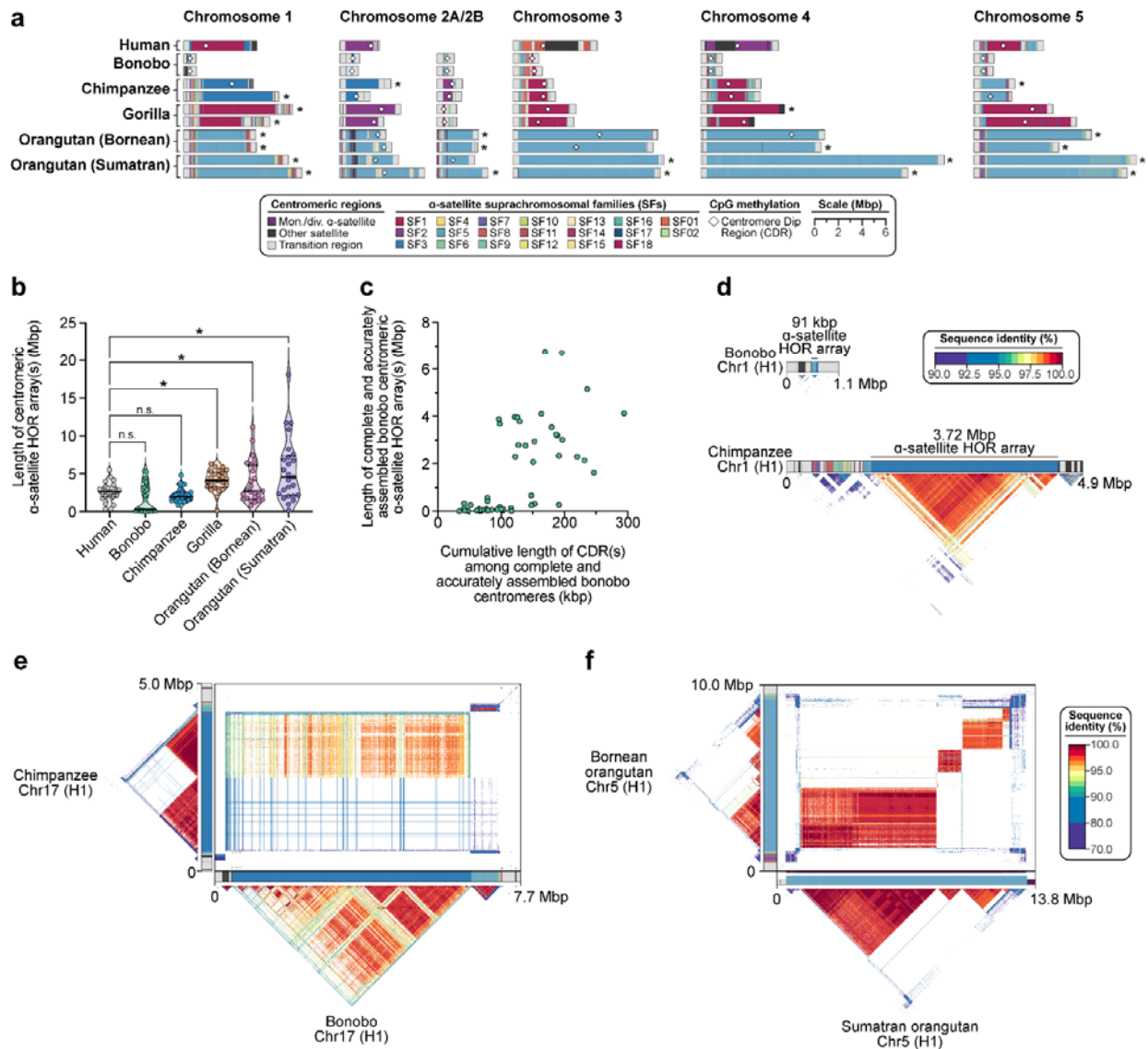
783 As previously noted<sup>84</sup>, chimpanzee  $\alpha$ -satellite HOR arrays are consistently smaller: 0.86-fold the  
784 length of their human counterparts (**Fig. 7a**). Additionally, the chimpanzee centromeres are  
785 typically composed of a single  $\alpha$ -satellite HOR array flanked by short stretches of divergent  $\alpha$ -  
786 satellite HORs and monomeric sequences, which are interspersed with TEs before extending into  
787 the p- and q-arms (**Fig. 7d**). In contrast, the gorilla  $\alpha$ -satellite HOR arrays are, on average, 1.58-  
788 fold larger than human (**Fig. 7a-b**), and unlike bonobo and chimpanzee, they are composed of  
789 punctuated regions of  $\alpha$ -satellite HORs, or regions of  $\alpha$ -satellite HORs that have high sequence  
790 identity within them but much lower sequence identity with neighboring regions, flanked by  
791 larger transition zones to monomeric  $\alpha$ -satellite sequence. The gorilla centromeres show a high

792 degree of haplotypic variation, with many paternal and maternal centromeres varying in size,  
793 sequence, and structure. We find that 30.4% (7 out of 23) gorilla  $\alpha$ -satellite HOR array pairs vary  
794 in size by >1.5-fold (especially HSA chromosomes 1, 2a, 4, 10, 15, 18 and 19), and 9 out of 23  
795 pairs (~39.1%) have  $\alpha$ -satellite HOR arrays with >5% sequence divergence between homologs  
796 (HSA chromosomes 1, 4–6, 10–12, 15, and 19). Finally, the Bornean and Sumatran orangutan  $\alpha$ -  
797 satellite HOR arrays are among the largest (1.52- and 2.11-fold larger, on average, than humans;  
798 **Fig. 7b**) and are characterized by multiple pockets of divergent  $\alpha$ -satellite HORs. A typical  
799 Bornean or Sumatran orangutan centromere has three or four distinct pockets of  $\alpha$ -satellite HORs,  
800 with up to nine distinct HOR arrays observed in a single centromere (Bornean chromosome 19).

801 Congeneric species of *Pan* and *Pongo* present an opportunity to assess the evolution of  
802 centromeric  $\alpha$ -satellites over evolutionary periods of time. Comparison of the centromeric  $\alpha$ -  
803 satellite HOR arrays from orthologous chromosomes across the bonobo and chimpanzee  
804 genomes reveals, for example, that 56% of them (14 out of 25 centromeres, including both X and  
805 Y) share a common identifiable ancestral sequence, such as that present in HSA chromosome 17  
806 (**Fig. 7e**). On this chromosome, the entire bonobo  $\alpha$ -satellite HOR array is ~92–99% identical to  
807 one domain of  $\alpha$ -satellite HORs present in the chimpanzee centromere. However, the  
808 chimpanzee centromere contains a second domain of  $\alpha$ -satellite HORs that spans approximately  
809 half of the  $\alpha$ -satellite HOR array. This domain is <70% identical to the bonobo  $\alpha$ -satellite HORs,  
810 indicating the formation of a new  $\alpha$ -satellite HOR array subregion acquired specifically in the  
811 chimpanzee lineage. Thus, over <2 mya, a new  $\alpha$ -satellite HOR arises and expands to become  
812 the predominant HOR distinguishing two closely related species (**Fig. 7e**). Given the shorter  
813 speciation time of orangutan (0.9 mya),  $\alpha$ -satellite HOR evolution is more tractable, with  $\alpha$ -  
814 satellite HORs sharing >97% sequence identity, including domains with 1:1 correspondence.  
815 However, in about a fifth of orangutan centromeres, we identify stretches of  $\alpha$ -satellite HORs  
816 present in Bornean but not Sumatran (or vice versa). The emergence of lineage-specific  $\alpha$ -  
817 satellite HOR sequences occurred on five chromosomes (HSA chromosomes 4, 5, 10, 11, 16; **Fig.**  
818 **7f**) and is marked by extremely high sequence identity (>99%) between  $\alpha$ -satellite HOR arrays,  
819 suggesting rapid turnover and homogenization of newly formed orangutan  $\alpha$ -satellite HORs.

820 We leveraged the new assemblies of these NHP centromeres to assess the location and  
821 distribution of the putative kinetochore—the large, proteinaceous structure that binds  
822 centromeric chromatin and mediates the segregation of chromosomes to daughter cells during  
823 mitosis and meiosis<sup>85,86</sup>. Previous studies in both humans<sup>8</sup> and NHPs<sup>84,87</sup> have shown that  
824 centromeres typically contain one kinetochore site, marked by one or more stretch of  
825 hypomethylated CpG dinucleotides termed the centromere dip region (CDR)<sup>88</sup>. We carefully  
826 assessed the CpG methylation status of all 237 primate centromeres and found that all contain at  
827 least one region of hypomethylation, consistent with a single kinetochore site. Focusing on the  
828 bonobo centromeres, where we find a bimodal distribution in  $\alpha$ -satellite HOR array length (**Fig.**  
829 **7b**), we show that CDR length and centromere length correlate ( $R^2=0.41$ ). In other words, the  
830 bonobo “minicentromeres” tend to associate with smaller CDRs when compared to larger  
831 centromeres (**Fig. 7c**). While much more in-depth functional studies need to be performed, this

832 finding suggests the reduced  $\alpha$ -satellite HOR arrays in bonobo are effectively limiting the  
 833 distribution of the functional component of the centromere.



834

835 **Figure 7. Assembly of 237 NHP centromeres reveals variation in  $\alpha$ -satellite HOR array size,**  
 836 **structure, and composition. a)** Sequence and structure of  $\alpha$ -satellite HOR arrays from the human (T2T-  
 837 CHM13), bonobo, chimpanzee, gorilla, Bornean orangutan, and Sumatran orangutan chromosome 1–5  
 838 centromeres, with the  $\alpha$ -satellite suprachromosomal family (SF) indicated for each centromere. The  
 839 sequence and structure of all completely assembled centromeres is shown in **Fig. CENSATS1**.

840 **b)** Variation in the length of the  $\alpha$ -satellite HOR arrays for NHP centromeres. Bonobo centromeres have a  
 841 bimodal length distribution, with 28 chromosomes showing “minicentromeres” (with  $\alpha$ -satellite HOR  
 842 arrays <700 kbp long). **c)** Correlation between the length of the bonobo active  $\alpha$ -satellite HOR array and  
 843 the length of the CDR for the same chromosome. **d)** Example showing that the bonobo and chimpanzee  
 844 chromosome 1 centromeres are divergent in size despite being from orthologous chromosomes.  
 845 **e)** Sequence identity heatmap between the chromosome 17 centromeres from bonobo and chimpanzee  
 846 show a common origin of sequence as well as the birth of new  $\alpha$ -satellite HORs in the chimpanzee

847 lineage. **f**) Sequence identity heatmap between the chromosome 5 centromeres from the Bornean and  
848 Sumatran orangutans show highly similar sequence and structure, except for one pocket of  $\alpha$ -satellite  
849 HORs that is only present in the Bornean orangutan. \*,  $p < 0.05$ ; n.s., not significant.

850

851 **Subterminal heterochromatin.** In addition to centromeres, we completely sequenced and  
852 assembled the subterminal heterochromatic caps of siamang, chimpanzee, bonobo, and gorilla  
853 (**Fig. 1c & Fig. 8a**). In total, these account for 1.05 Gbp of subterminal satellite sequences (642  
854 Mbp or 18.2% of the siamang genome). These massive structures (up to 26 Mbp in length) are  
855 thought to be composed almost entirely of tandem repetitive DNA: a 32 bp AT-rich satellite  
856 sequence, termed pCht7 in *Pan* and gorilla, or a 171 bp  $\alpha$ -satellite repeat present in a subset of  
857 gibbon species<sup>89-91</sup>. While their function is not known, these chromosomal regions have been  
858 implicated in nonhomologous chromosome exchange and unique features of telomeric RNA  
859 metabolism<sup>92,93</sup>. Our analysis indicates that we successfully sequenced 79 gapless subterminal  
860 caps in gorilla (average length=6.6 Mbp) and 57 and 46 caps in chimpanzee and bonobo,  
861 including both haplotypes (average lengths 4.8 and 5.2 Mbp, respectively) with less than 3.8% of  
862 pCht arrays flagged as potentially misassembled (**Fig. 8a**). Siamangs possess the largest (average  
863 length 6.7 Mbp) and most abundant subterminal satellite blocks (96 out of 100 chromosomal  
864 ends across the two haplotypes).

865 In gorilla and chimpanzees (*Pan*), the caps are organized into higher order structures where pCht  
866 subterminal satellites form tracts of average length of 335 to 536 kbp interrupted by spacer SD  
867 sequences of a modal length of 32 kbp (*Pan*) or 34 kbp (gorilla; **Fig. SubterminalS1**). The  
868 spacer sequences are each unique to the *Pan* and gorilla lineages but we confirm that each began  
869 originally as a euchromatic sequence that became duplicated interstitially in the common  
870 ancestor of human and apes. For example, the 34 kbp spacer in gorilla maps to a single copy  
871 sequence present in orangutans and human chromosome 10, which began to be duplicated  
872 interstitially on chromosome 7 in chimpanzee but only in gorilla became associated with pCht  
873 satellites expanding to over 477 haploid copies as part of the formation of the heterochromatic  
874 cap. Similarly, the ancestral sequence of pan lineage spacer maps syntenically to orangutan and  
875 human chr9. The ancestral sequence duplicated to multiple regions in gorilla (q-arms of chr4, 5,  
876 8, X, and p-arms of chr2A and 2B), before being captured and hyperexpanded (>345 copies) to  
877 form the structure of subterminal satellites of chimpanzee and bonobo. Analyzing CpG  
878 methylation, we find that each spacer demarcates a pocket of hypomethylation flanked by  
879 hypermethylated pCht arrays within the cap (**Fig. 8b-d**). Of note, this characteristic  
880 hypomethylation pattern is not observed at the ancestral origin or interstitially duplicated  
881 locations (**Fig. 8e**), suggesting an epigenetic feature not determined solely by sequence but by its  
882 association with the subterminal heterochromatic caps. Similar to the great apes, we find  
883 evidence of a hypomethylated spacer sequence also present in the siamang subterminal cap;  
884 however, its modal length is much larger (57.2 kbp in length) and its periodicity is less uniform  
885 occurring every 750 kbp (**Fig. SubterminalS2**). Nevertheless, the fact that these similar  
886 epigenetic features of the spacer evolved independently may suggest a functional role with  
887 respect to subterminal heterochromatic caps.

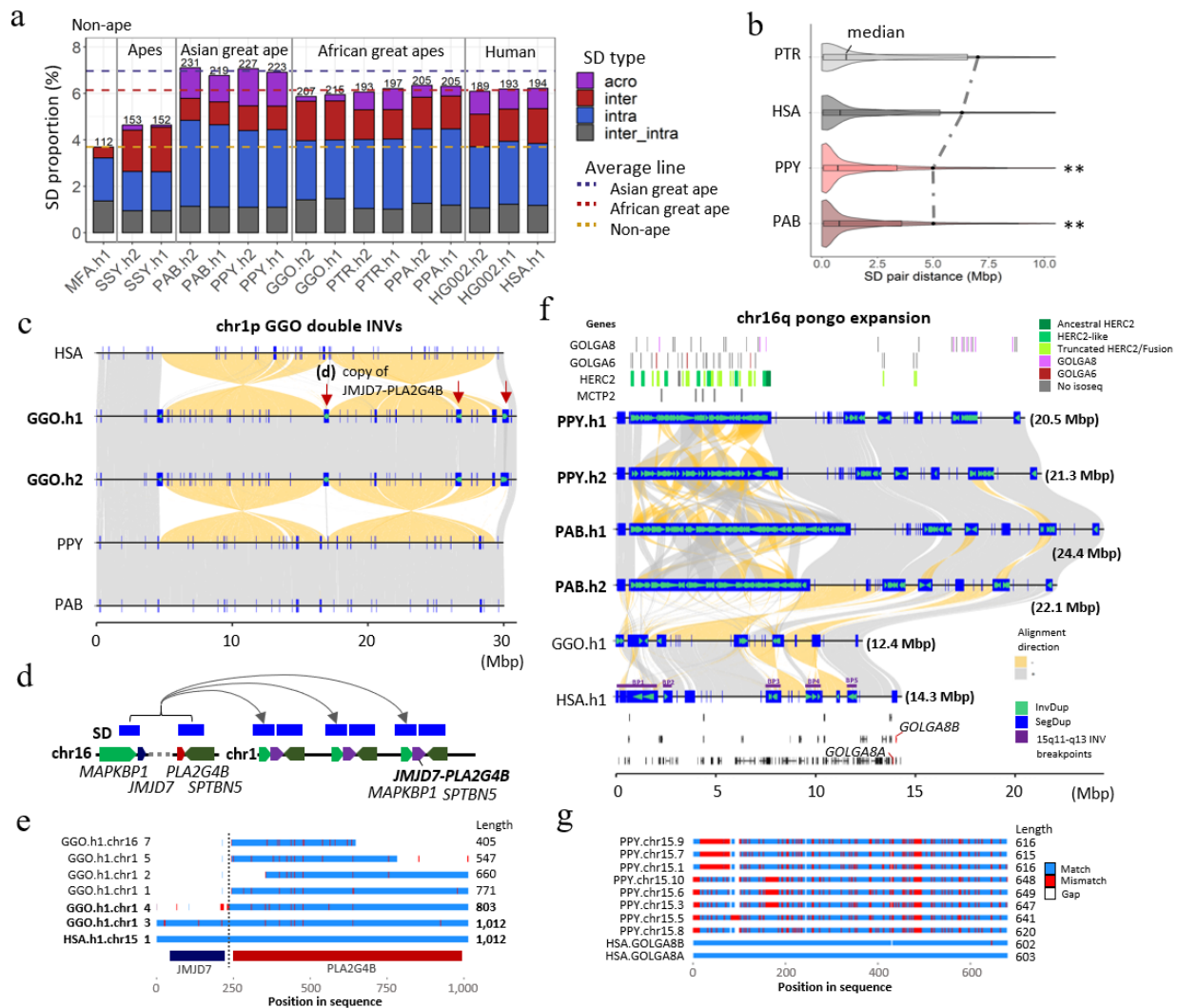


910 which also exhibit larger interspersions of intrachromosomal SDs (**Fig. 9b & SD.S3**). The  
911 increased SD content in orangutans is due to a greater number of acrocentric chromosomes (10  
912 vs. 5 on average for other apes) and a preponderance of clustered duplications. Consistent with  
913 the expansion of Asian great ape SDs, we find the largest number of lineage-specific SDs in the  
914 *Pongo* lineage (93.3 Mbp, followed by gorilla- and human-specific SDs (75.1 and 60.6 Mbp,  
915 respectively). Many SDs (79.3 to 95.6 Mbp per haplotype) in orangutan constitute massive,  
916 Mbp-scale SD clusters, including a mixture of tandem and inverted duplications up to 21.5 Mbp  
917 in size; in other species, the total number of such clustered duplications accounts for only 30 to  
918 40 Mbp per haplotype (except bonobo). In general, the number of SDs assigned to different  
919 lineages correlates with branch length ( $r^2=0.80$ ) (**Fig. SD.S4**) with the exception of siamang and  
920 some ancestral nodes reflecting the great ape expansion of SDs<sup>96</sup>.

921 Leveraging the increased sensitivity afforded by FLNC Iso-Seq, we annotated the transcriptional  
922 content of lineage-specific SDs identifying hundreds of potential genes, including gene family  
923 expansions often occurring in conjunction with chromosomal evolutionary rearrangements. We  
924 highlight two examples in more detail. First, at two of the breakpoints of a 30 Mbp double  
925 inversion of gorilla chromosome 1, we identify a gorilla-specific expansion of the genes  
926 *MAPKBP1* and *SPTBN5* as well as *PLA2G4B-JMJD7* (**Fig. 9c-e**) originating from an  
927 interchromosomal SD from ancestral loci mapping to HSA chromosome 15 (duplicated in other  
928 chromosomes in chimpanzee and bonobo; **Table Genes.S5**). We estimate these duplications  
929 occurred early after gorilla speciation, 6.1 mya (**Fig. SD.S5a**) followed by subsequent expansion  
930 resulting in the addition of eight copies (one truncated) mapping to two of the breakpoints of the  
931 double inversion. Investigating the Iso-Seq transcript model of this gene, revealed that five of the  
932 new gorilla copies are supported by multi-exon transcripts. Two of these additional copies  
933 possess valid start and stop codons spanning at least 70% of the homologous single-copy  
934 ortholog gene in humans (**Fig. 9e & SD.S5a**). Notably, the ancestral copy of this gene in gorilla  
935 (HSA chromosome 15q) is found to be highly truncated (40% of original protein) suggesting that  
936 the new chromosome 1 copies may have assumed and refined the function.

937 Second, in orangutan, we find a restructured 20-Mbp region corresponding to the Prader-Willi  
938 syndrome (PWS)<sup>65</sup> and the 15q13 microdeletion syndrome<sup>97</sup> region in humans. This includes a  
939 massive 6.8–10.8 Mbp expansion of clustered tandem and inverted duplications mapping distally  
940 to breakpoint 1 of PWS as well as smaller 200–550 kbp expansions of *GOLGA6/8* repeats distal  
941 to PWS BP3/4 (**Fig. 9f**). We estimate that the larger region, alone, is composed of 87–111 copies  
942 of fragments of *GOLGA6/8*, *HERC2* and *MCTP2*. We find Iso-Seq transcript support for 37–39  
943 distinct orangutan copies. Using *GOLGA8* as a marker, we show that it has expanded to 10–12  
944 copies (>70% of original length) in orangutan but exists as a single copy in gorilla and bonobo  
945 and in two copies (*GOLGA8A* and *B*) in human out of multiple *GOLGA8* genes, retaining at least  
946 70% of sequence compared to orangutan sequence (**Fig. 9f-g, Fig. SD.S5b**). We estimate that the  
947 *Pongo* expansion of *GOLGA8* occurred 7.3 mya (**Fig. SD.S5b**), long before the species diverged.  
948 Alignment of the translated peptide sequence, we observe 17.1–23.7% divergence from the  
949 human copy (*GOLGA8A*; **Fig. 9g**). Based on studies of the African great ape genomes and  
950 humans, *GOLGA8* was among more than a dozen loci defined as “core duplicons” promoting the

951 interspersion of SDs and genomic instability via palindromic repeat structures<sup>98,99</sup>. Our findings  
 952 extend this recurrent genomic feature for the *GOLGA8* duplicons to the Asian ape genomes.



953  
 954 **Figure 9. Ape SD content and new genes.** **a**) Comparative analysis of primate SDs comparing the  
 955 proportion of acrocentric (purple), interchromosomal (red), intrachromosomal (blue), and shared  
 956 inter/intrachromosomal SDs (gray). The total SD Mbp per genome is indicated above each histogram with  
 957 the colored dashed lines showing the average Asian, African great ape, and non-ape SD (*MFA*=*Macaca*  
 958 *fascicularis*<sup>98</sup>; see **Fig.SD.S2** for additional non-ape species comparison). **b**) A violin plot distribution of  
 959 pairwise SD distance to the closest paralog where the median (black line) and mean (dashed line) are  
 960 compared for different apes (see **Fig. SD.S3** for all species and haplotype comparisons). An excess of  
 961 interspersed duplications ( $p < 0.001$  one-sided Wilcoxon rank sum test) is observed for chimpanzee and  
 962 human when compared to orangutan. **c**) Alignment view of chr1 double inversion. Alignment direction is  
 963 indicated by + as gray and - as yellow. SDs as well as those with inverted orientations are indicated by  
 964 blue rectangles and green arrowheads. The locations in which the *JMJD7-PLA2G4B* gene copy was found  
 965 are indicated by the red arrows. **d**) duplication unit containing three genes including *JMJD7-PLA2G4B*.  
 966 **e**) Multiple sequence alignment of the translated *JMJD7-PLA2G4B*. Match, mismatch and gaps are  
 967 indicated by blue, red and white. Regions corresponding to each of *JMJD7* or *PLA2G4B* are indicated by  
 968 the track below. **f**) Alignment view of chr16q. The expansion of *GOLGA6/8*, *HERC2*, and *MCTP2* genes

969 are presented in the top track. 16q recurrent inversion breakpoints are indicated in the human genome.  
970 The track at the bottom indicates the gene track with *GOLGA8* human ortholog in red. g) Multiple  
971 sequence alignment of the translated *GOLGA8*.

972

## 973 **DISCUSSION**

974 The completion of the ape genomes significantly refines previous analyses providing a more  
975 definitive resource for all future evolutionary comparisons. These include an improved and more  
976 nuanced understanding of species divergence, human ancestral alleles, incomplete lineage  
977 sorting, gene annotation, repeat content, divergent regulatory DNA, and complex genic regions  
978 as well as species-specific epigenetic differences involving methylation. These preliminary  
979 analyses reveal hundreds of new candidate genes and regions to account for phenotypic  
980 differences among the apes. For example, we observed an excess of HAQERS corresponding to  
981 bivalent promoters thought to contain gene regulatory elements that exhibit precise  
982 spatiotemporal activity patterns in the context of development and environmental response<sup>100</sup>.  
983 Bivalent chromatin state enrichments have not yet been observed in fast-evolving regions from  
984 other great apes, which may reflect limited cross-species transferability of epigenomic  
985 annotations from human. The finding of a HAQER enriched gene, *ADCYAP1*, that is  
986 differentially regulated in speech circuits and methylated in the layer 5 projection neurons that  
987 make the more specialized direct projections to brainstem motor neurons in humans, shows the  
988 promise of T2T genomes to identify hard to sequence regions important for complex traits.  
989 Perhaps most importantly, we provide an evolutionary framework for understanding the ~10%–  
990 15% of highly divergent, previously inaccessible regions of ape genomes. In this regard, we  
991 highlight a few noteworthy findings.

992 *Orangutans show the greatest amount of recent segmental duplication.* Comparative analyses  
993 suggest expansion of SDs in the common ancestor of the great ape lineage as opposed to the  
994 African great ape lineage as we originally proposed based on sequence read-depth analyses back  
995 to the human reference genome<sup>95,96</sup>. This discrepancy highlights the importance of *ab initio*  
996 sequence genome assembly of related lineages that are comparable in quality and contiguity. The  
997 assembly of the acrocentric chromosomes (of which orangutans have the maximum at 9/10) and  
998 the resolution of massive (10–20 Mbp) tandem SDs in the orangutan species account for the  
999 increase in SD content among the Asian great apes. The African great ape lineage still stands out  
1000 for having the largest fraction of interspersed SDs—a genomic architectural feature that  
1001 promotes recurrent rearrangements facilitating syndromic disease associated with autism and  
1002 developmental delay in the human species<sup>101</sup>. Complete sequence resolution of NHP interspersed  
1003 SDs provides a framework for understanding disease-causing copy number variants in these  
1004 other NHP lineages<sup>102</sup>.

1005 *Large-scale differences in acrocentric chromosomes.* The short arms of NOR+ ape acrocentric  
1006 chromosomes appear specialized to encode rRNA genes. On the autosomes, ape NORs exist  
1007 exclusively on the acrocentric chromosomes, embedded within a gene-poor and satellite-rich  
1008 short arm. On the Y chromosome, NORs occur occasionally toward the end of the chromosome



1009 and adjacent to satellites shared with other acrocentric chromosomes. Prior analysis of the human  
1010 pangenome suggested heterologous recombination between chromosomes with NORs as a  
1011 mechanism for concerted evolution of the rRNA genes<sup>78,103</sup>. Comparative analysis of ape  
1012 genomes provides further support for this hypothesis. For example, the uniform direction of all  
1013 rDNA arrays within a species would permit crossover recombination between heterologous  
1014 chromosomes without substantial karyotypic consequence. However, rare translocations,  
1015 mediated by the large SDs commonly surrounding the NORs, have occurred during ape  
1016 evolution, resulting in a different complement of NOR+ acrocentric chromosomes and possibly  
1017 creating reproductive barriers associated with speciation<sup>104</sup>.

1018 *Lineage-specific gene family expansions/explosions and rearrangements.* The number of lineage-  
1019 specific duplications that encode transcripts and potential genes is now estimated at hundreds per  
1020 ape lineage often occurring at sites of evolutionary chromosomal rearrangements that have been  
1021 historically difficult to sequence resolve (**Table SD.S1**). Our analysis has uncovered hundreds of  
1022 fixed inversions frequently associated with the formation of these lineage-specific duplications.  
1023 These findings challenge the predominant paradigm that subtle changes in regulatory DNA<sup>105</sup> are  
1024 the major mechanism underlying ape species differentiation. Rather, the expansion, contraction,  
1025 and restructuring of SDs lead to not only dosage differences but concurrent gene innovation and  
1026 chromosomal structural changes<sup>106</sup>. Indeed, in the case of human, four such gene family  
1027 expansions, namely *NOTCH2NL*<sup>107</sup>, *SRGAP2C*<sup>108,109</sup>, *ARHGAP11*<sup>99</sup> and *TBC1D3*<sup>110,111</sup>, have  
1028 been functionally implicated over the last decade in the frontal cortical expansion of the human  
1029 brain<sup>107-110</sup> as well as human-specific chromosomal changes<sup>99</sup>. Detailed characterization of the  
1030 various lineage-specific expansions in NHPs will no doubt be more challenging yet it is clear  
1031 that such SDRs are an underappreciated genic source of interspecific difference and potential  
1032 gene neofunctionalization.

1033 *Bonobo minicentromeres.* We identified several idiosyncratic features of centromere  
1034 organization and structure that characterize the different ape lineages, significantly extending  
1035 earlier observations based on the characterization of five select centromeres<sup>84</sup>. Perhaps the most  
1036 remarkable is the bimodal distribution of centromere HOR length in the bonobo lineage—19 of  
1037 the 48 bonobo centromeres are, in fact, less than 100 kbp in size. Given the estimated divergence  
1038 of the *Pan* lineage, such 300-fold reductions in size must have occurred very recently—in the  
1039 last million years. These bonobo “minicentromeres” appear fully functional with a well-defined  
1040 CDR (encompassing all of the  $\alpha$ -satellite DNA). Thus, their discovery may provide a roadmap  
1041 for the design of smaller, more streamlined artificial chromosomes for the delivery and stable  
1042 transmission of new genetic information in human cells<sup>112</sup>.

1043 *Epigenetic architecture of subterminal heterochromatin.* Our analysis suggests that the  
1044 subterminal chromosomal caps of chimpanzee, gorilla, and siamang have evolved independently  
1045 to create multi-Mbp of heterochromatin in each species. In chimpanzee and gorilla, we define a  
1046 common organization of a subterminal spacer (~30 kbp in size) that is hypomethylated and  
1047 flanked by hypermethylated heterochromatic satellite with a periodicity of one spacer every 335–  
1048 536 kbp of satellite sequence (pCht satellite in hominoids and  $\alpha$ -satellite in hylobatids). In each  
1049 case, the spacer sequence differs in its origin but has arisen as an ancestral SD<sup>89</sup> that has become

1050 integrated and expanded within the subterminal heterochromatin. In contrast to the ancestral  
1051 sequences located in euchromatin, the spacer sequences embedded within the subterminal caps  
1052 acquire more pronounced hypomethylation signatures suggesting an epigenetic feature. This  
1053 subterminal hypomethylation pocket is reminiscent of the CDRs identified in centromeres that  
1054 define the sites of kinetochore attachment<sup>113</sup> as well as methylation dip region observed among  
1055 some acrocentric chromosomes<sup>26</sup>. It is tempting to speculate that the subterminal  
1056 hypomethylation pocket may represent a site of protein binding or a “punctuation” mark perhaps  
1057 facilitating ectopic exchange and concerted evolution driving persistent subtelomeric  
1058 associations and meiotic exchanges between nonhomologous chromosomes<sup>114</sup>.

1059 While the ape genomes sampled here are nearly complete, some limitations remain. Sequence  
1060 gaps still exist in the acrocentric centromeres and a few other remaining complex regions where  
1061 the largest and most identical tandem repeats reside. This is especially the case for the Sumatran  
1062 orangutan centromeres where only 27% are completely assembled. The length (nearly double the  
1063 size) and the complex compound organization of orangutan  $\alpha$ -satellite HOR sequence will  
1064 require specialized efforts to completely order and orientate<sup>84</sup>. Nevertheless, with the exception  
1065 of these and other large tandem repeat arrays, we estimate that ~99.5% of the content of each  
1066 genome has been characterized and is correctly placed. Second, although we completed the  
1067 genomes of a representative individual, we sequenced and assembled only two haplotypes from  
1068 each species and more than 15 species/subspecies of apes remain<sup>115</sup>. Sampling more closely  
1069 related species that diverged within the last million years will provide a unique opportunity to  
1070 understand the evolutionary processes shaping the most dynamic regions of our genome. High-  
1071 quality assemblies of all chimpanzee species<sup>116</sup>, as well as the numerous gibbon species<sup>117</sup>, will  
1072 provide critical insight into selection, effective population size, and the rapid structural  
1073 diversification of ape chromosomes at different time points. Finally, while high-quality genomes  
1074 help eliminate reference bias, they do not eliminate annotation biases that favor the human. This  
1075 will be especially critical for both genes and regulatory DNA that have rapidly diverged between  
1076 the species.

1077

## 1078 DATA AVAILABILITY

1079 The raw genome sequencing data generated by this study are available under NCBI BioProjects,  
1080 PRJNA602326, PRJNA976699–PRJNA976702, and PRJNA986878–PRJNA986879 and  
1081 transcriptome data are deposited under BioProjects, PRJNA902025 (UW Iso-Seq) and  
1082 PRJNA1016395 (UW and PSU Iso-Seq and short-read RNA-seq). The genome assemblies are  
1083 available from GenBank under accessions: GCA\_028858775.2, GCA\_028878055.2,  
1084 GCA\_028885625.2, GCA\_028885655.2, GCA\_029281585.2 and GCA\_029289425.2. Genome  
1085 assemblies can be downloaded via NCBI  
1086 ([https://www.ncbi.nlm.nih.gov/datasets/genome/?accession=GCF\\_028858775.2,GCF\\_029281585.2,GCF\\_028885625.2,GCF\\_028878055.2,GCF\\_028885655.2,GCF\\_029289425.2](https://www.ncbi.nlm.nih.gov/datasets/genome/?accession=GCF_028858775.2,GCF_029281585.2,GCF_028885625.2,GCF_028878055.2,GCF_028885655.2,GCF_029289425.2)).  
1087  
1088 Convenience links to the assemblies and raw data are available on GitHub  
1089 (<https://github.com/marbl/Primates>) along with a UCSC Browser hub

1090 (<https://github.com/marbl/T2T-Browser>). The UCSC Browser hub includes genome-wide  
1091 alignments, CAT annotations, methylation, and various other annotation and analysis tracks used  
1092 in this study. The T2T-CHM13v2.0 and HG002v1.0 assemblies used here are also available via  
1093 the same browser hub, and from GenBank via accessions GCA\_009914755.4 (T2T-CHM13),  
1094 GCA\_018852605.1 (HG002 paternal), and GCA\_018852615.1 (HG002 maternal). The  
1095 alignments are publicly available to download or browse in HAL118 MAF and UCSC Chains  
1096 formats (<https://cglgenomics.ucsc.edu/february-2024-t2t-apes>).

1097

## 1098 **CODE AVAILABILITY**

1099 All code used for the reported analyses is available from our project's GitHub repository:  
1100 <https://github.com/marbl/Primates>

1101

## 1102 **ACRONYMS & ABBREVIATIONS**

1103 AQER: ancestor quickly evolved region  
1104 cDNA: complementary deoxyribonucleic acid  
1105 CDR: centromere dip region  
1106 CRE: cis-regulatory element  
1107 DJ: distal junction [region]  
1108 ENC: evolutionary neocentromere  
1109 ERV: endogenous retrovirus  
1110 FLNC: full-length non-chimeric  
1111 GGO: gorilla  
1112 HAQER: human ancestor quickly evolved region [human branch]  
1113 HAS: human  
1114 HiFi: high-fidelity  
1115 HOR: higher-order repeat  
1116 ILS: incomplete lineage sorting  
1117 LINE: long interspersed nuclear element  
1118 LTR: long terminal repeats  
1119 MEI: mobile element insertion  
1120 MHC: major histocompatibility complex  
1121 mya: million years ago  
1122 ncRNA: noncoding RNA  
1123 Ne: effective population sizes  
1124 NHP: nonhuman primate  
1125 NOR: nucleolar organizer region  
1126 NUMT: nuclear sequence of mitochondrial DNA origin  
1127 ONT: Oxford Nanopore Technologies  
1128 ORF: open reading frame  
1129 PAB: Sumatran orangutan  
1130 PacBio: Pacific Biosciences Inc.

1131 PGGB: pangenome graph builder  
1132 PLE: Penelope-Like Retroelements  
1133 PPA: bonobo  
1134 PPY: Bornean orangutan  
1135 PTR: chimpanzee  
1136 PWS: Prader-Willi syndrome  
1137 RC: rolling circle repeats  
1138 rDNA/rRNA: ribosomal deoxyribonucleic/ribonucleic acid  
1139 SD: segmental duplication  
1140 SDR: structurally divergent region  
1141 SF: suprachromosomal family  
1142 SINE: short interspersed nuclear element  
1143 SNV: single-nucleotide variant  
1144 SVA: SINE-VNTR-Alu element  
1145 T2T: telomere-to-telomere  
1146 TE: transposable element  
1147 TOGA: Tool to infer Orthologs from Genome Alignments  
1148 UL: ultra-long  
1149 VNTR: variable number tandem repeat

1150

## 1151 **COMPETING INTERESTS**

1152 E.E.E. is a scientific advisory board (SAB) member of Variant Bio, Inc. C.T.W. is a co-  
1153 founder/CSO of Clareo Biosciences, Inc. W.L. is a co-founder/CIO of Clareo Biosciences, Inc.  
1154 The other authors declare no competing interests.

1155

## 1156 **ACKNOWLEDGMENTS**

1157 We thank Richard Buggs for useful suggestions and Tonia Brown, Agostinho Antunes and  
1158 Mohamed Emam for editing the manuscript and supplementary note. This research was  
1159 supported, in part, by the Intramural Research Program of the National Human Genome  
1160 Research Institute, National Institutes of Health (NIH), and extramural NIH grants  
1161 R35GM151945 (to K.D.M.), R01 HG002385, R01 HG010169, U24 HG007497 (to E.E.E.), R35  
1162 GM146926 (to Z.A.S.), R35 GM146886 (to C.D.H.), R35GM142916 (to P.H.S.), R01HG012416  
1163 (to P.H.S. and E.G.), R35HG011332 (to C.B.L.), R01GM123312 (to R.J.O.), U24 HG010263 (to  
1164 M.C.S), R35GM133747 (to R.C.M.), U41HG007234 (to P.H., M.D., B.P.), R01HG010329 (to  
1165 P.H., M.D.), R01MH120295 (to M.D.), 1P20GM139769 (to M.K.K. and M.L.), R35GM133600  
1166 (to C.R.B and P.B.), and 1U19 AG056169-01A1, UH3 AG064706, and U19 AG023122 (to  
1167 N.J.S.) as well as a Vallee Scholars Award to P.H.S. We also acknowledge financial support by  
1168 the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 437857095,  
1169 444810852 (to T.L.L.), by Verne M. Willaman Endowment Professorship (to K.D.M.), and by  
1170 the John and Donna Krenicki Endowment Professorship (to R.J.O.). Sequencing was partially  
1171 supported by the NIH Intramural Sequencing Center. This work used the computational

1172 resources of the NIH HPC Biowulf cluster (<https://hpc.nih.gov>), the HPC in the Computational  
1173 Biology Core in the Institute for Systems Genomics at UConn, Clemson University's HPC  
1174 Palmetto Cluster, and the HPC in the Genomics Institute at UCSC. RNA-seq sequencing was  
1175 performed at PSU Genomics Core facility, and Hi-C sequencing was performed at the Genome  
1176 Sciences core facility at the Penn State College of Medicine. This work was supported by the  
1177 National Library of Medicine Training Program in Biomedical Informatics and Data Science  
1178 (T15LM007093 to B.K.) and in part by the National Institute of Allergy and Infectious Diseases  
1179 (P01-AI152999 to B.K.). We acknowledge financial support under the National Recovery and  
1180 Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.1, Call for tender No. 104  
1181 published on 2.2.2022 by the Italian Ministry of University and Research (MUR), funded by the  
1182 European Union – NextGenerationEU– Project Title Telomere-to-telomere sequencing: the new  
1183 era of Centromere and neocentromere eVolution (CenVolution) – CUP H53D23003260006 -  
1184 Grant Assignment Decree No. 1015 adopted on 07/07/2023 by the Italian MUR. – Project Title  
1185 SUDWAY: Substance Use Disorders through Whole genome, psychological and neuro-  
1186 endophenotypes AnalYsis CUP H53D2300331- Grant Assignment Decree No. 1015 adopted on  
1187 7 July 2023 by the Italian MUR. F.M. was supported by Fondazione con il Sud (2018-PDR-  
1188 01136). We are grateful to the Frozen Zoo® at San Diego Zoological Society for providing the  
1189 fibroblast cell lines. This work was supported by the Italian Ministry of University and Research  
1190 (MUR) grant PRIN 2020 (project code 2020J84FAM, CUP H93C20000040001) to F.A. We  
1191 thank the Genome in a Bottle Consortium for sharing preliminary RNA-seq data for HG002.  
1192 E.E.E. and E.D.J. are investigators of the Howard Hughes Medical Institute. The work of DH,  
1193 PM and FTN was supported by the National Center for Biotechnology Information of the  
1194 National Library of Medicine (NLM), National Institutes of Health.

1195 This article is subject to HHMI's Open Access to Publications policy. HHMI lab heads have  
1196 previously granted a nonexclusive CC BY 4.0 license to the public and a sublicensable license to  
1197 HHMI in their research articles.

1198

## 1199 **AUTHOR CONTRIBUTIONS**

1200 Individual analysis leads are indicated with an asterisk. Lu.C., La.C., O.A.R., Cy.S., Ma.H., B.M.,  
1201 and K.D.M. managed sampling. B.M. and A.P.L. performed transcriptome data generation. K.H.,  
1202 G.G.B., S.Y.B., J.C. generated ONT long-read data. J.C., R.E.G. and Sa.S. provided Illumina  
1203 sequencing data. G.H.G., K.M.M., P.H.S. and J.L.R. generated HiFi sequencing data. R.E.G. and  
1204 Sa.S. made Hi-C libraries that were later sequenced by B.M. \*B.D.P. and A.R. managed data,  
1205 processed data submissions, and coordinated administrative tasks. W.T.H., J.W., A.R., and  
1206 B.D.P. performed assembly QC. \*D.A. and S.K. assembled the genomes. \*A.R. performed  
1207 polishing and created the genome browsers. Assembly generation was supervised by S.K. S.M.  
1208 performed chromosome recognition and M.V. led definition of chromosome nomenclature. L.S.,  
1209 K.K. and K.D.M. analyzed non-B DNA. B.K., W.W., A.G., E.M., E.G., G.F. and P.H.S. created  
1210 pangenome graph alignment. \*G.H. generated the Cactus alignments. P.H.S., R.S.H., S.K.M.,  
1211 B.K., W.W., A.G., E.M., E.G., G.F. and K.D.M. performed divergence analysis. Co.S. and B.P.

1212 performed ancestral allele analysis. B.P., P.H., M.D., D.H., J.F.M., P.M., F.R.R., F.T., S.C. K.P.,  
1213 and K.D.M. analyzed and annotated genes. \*P.H. annotated lineage-specific genes, integrated  
1214 gene annotations and managed sharing of the annotation data. B.P. supervised gene annotation  
1215 analysis. \*F.M. and I.R. performed ILS analysis and predicted speciation times. \*J.M.S., P.B.,  
1216 C.R.B., C.F., P.Z., G.A.H. and R.J.O. analyzed repeat content. E.T. and K.D.M. investigated  
1217 NUMTs. M.L. and M.K.K. investigated specifically for species-specific MEIs. P.B. and C.R.B.  
1218 performed ORF analysis on species-specific FL-L1s. R.J.O. supervised and integrated the results.  
1219 \*A.N.S., Ar.B., Q.L., M.C.S., M.G.T., Z.A.S., C.D.H., R.C.M., and K.D.M. analyzed population  
1220 data and investigated selective sweeps. \*Y.S., An.B., E.E., I.G., W.L., M.P., P.A.P., Sw.S., Z.Z.,  
1221 Yi.Z. and C.T.W. analyzed immunoglobulin loci. Y.S. and C.T.W. led the analysis and  
1222 integrated the results. Jo.M., B.S.M., and T.L.L. performed annotation of MHC genes. P.H.  
1223 supported and validated the MHC annotation. Mi.T., performed phylogenetic tree analysis across  
1224 MHC loci. Y.E.L., D.R.S. and S.V.Y. analyzed epigenetic data focusing on methylation and gene  
1225 expression. \*J.M., M.L.Y., Y.Z. and T.S. analyzed replication timing. D.G. and T.S. generated  
1226 Repli-seq data. \*F.A., M.V. L.G., and D.Y. analyzed inversions and large-scale chromosome  
1227 rearrangements. F.A., D.Y. and D.P. visualized the data. \*J.L., J.H., S.Z. and Y.M. performed  
1228 SDR analysis. A.P.C., Mi.H. and N.J.S. performed TOGA analysis. A.P.C. integrated the results.  
1229 \*Ya.L., \*R.J.M., M.K., S.A.Z., C.B.L. analyzed divergent regions of the genome by predicting  
1230 AQERs. C.B.L. supervised the analysis and summarized the results. C.L., Yo.L. and E.D.J.  
1231 investigated further into candidate genes using AQER regions. \*S.J.S., A.P.S., J.L.G., T.P.,  
1232 G.M.A. and M.B. analyzed acrocentric regions. T.P. and G.M.A. performed NOR chromosome  
1233 imaging and quantification. S.J.S. and M.B. analyzed rDNA. A.P.S. generated dot plots. J.L.G.,  
1234 M.V. and A.M.P. supervised the analysis. \*G.A.L., K.H.M. and H.L. investigated centromeres.  
1235 G.A.L. integrated the section. \*D.Y. and E.E.E. investigated subterminal heterochromatin. D.Y.  
1236 performed the analyses and E.E.E. supervised the analysis. \*D.Y. and E.E.E. analyzed SD. E.E.E.  
1237 supervised the SD analysis. H.J. optimized the pipeline and D.P. and D.Y. visualized the data.  
1238 P.H. analyzed novel genes and curated gene annotation across the region. E.E.E., D.Y., and  
1239 A.M.P. wrote and edited the manuscript with input from all authors. E.E.E., A.M.P., and K.D.M.  
1240 initiated and supervised the project, acquired the funding along with other senior authors. E.E.E.  
1241 and A.M.P. coordinated the study.

1242 **REFERENCES**

- 1243 1 US DOE Joint Genome Institute, Initial sequencing and analysis of the human genome. *Nature*  
1244 **409**, 860-921 (2001).
- 1245 2 Venter, J. C. *et al.* The sequence of the human genome. *science* **291**, 1304-1351 (2001).
- 1246 3 Blanchette, M., Green, E. D., Miller, W. & Haussler, D. Reconstructing large regions of an  
1247 ancestral mammalian genome in silico. *Genome research* **14**, 2412-2423 (2004).
- 1248 4 Chimpanzee Sequencing and Analysis Consortium, A. Initial sequence of the chimpanzee  
1249 genome and comparison with the human genome. *Nature* **437** (2005).
- 1250 5 Gordon, D. *et al.* Long-read sequence assembly of the gorilla genome. *Science* **352**, aae0344  
1251 (2016).
- 1252 6 Prüfer, K. *et al.* The bonobo genome compared with the chimpanzee and human genomes.  
1253 *Nature* **486**, 527-531 (2012).
- 1254 7 Mao, Y. *et al.* A high-quality bonobo genome refines the analysis of hominid evolution. *Nature*  
1255 **594**, 77-81 (2021).
- 1256 8 Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44-53 (2022).
- 1257 9 Rhie, A. *et al.* The complete sequence of a human Y chromosome. *Nature* **621**, 344-354 (2023).
- 1258 10 Makova, K. D. *et al.* The complete sequence and comparative analysis of ape sex chromosomes.  
1259 *Nature*, 1-11 (2024).
- 1260 11 Rautiainen, M. *et al.* Telomere-to-telomere assembly of diploid chromosomes with Verkko.  
1261 *Nature Biotechnology* **41**, 1474-1482 (2023).
- 1262 12 Cheng, H., Asri, M., Lucas, J., Koren, S. & Li, H. Scalable telomere-to-telomere assembly for  
1263 diploid and polyploid genomes with double graph. *Nature Methods*, 1-4 (2024).
- 1264 13 Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality,  
1265 completeness, and phasing assessment for genome assemblies. *Genome biology* **21**, 1-27 (2020).
- 1266 14 Weissensteiner, M. H. *et al.* Accurate sequencing of DNA motifs able to form alternative (non-B)  
1267 structures. *Genome research* **33**, 907-922 (2023).
- 1268 15 Armstrong, J. *et al.* Progressive Cactus is a multiple-genome aligner for the thousand-genome  
1269 era. *Nature* **587**, 246-251 (2020).
- 1270 16 Herrero, J. *et al.* Ensembl comparative genomics resources. *Database* **2016**, bav096 (2016).
- 1271 17 Hickey, G. *et al.* Pangenome graph construction from genome alignments with Minigraph-Cactus.  
1272 *Nature biotechnology* **42**, 663-673 (2024).
- 1273 18 Liao, W.-W. *et al.* A draft human pangenome reference. *Nature* **617**, 312-324 (2023).
- 1274 19 Garrison, E. *et al.* Building pangenome graphs. *bioRxiv*, 2023.2004.2005.535718 (2023).
- 1275 20 Locke, D. P. *et al.* Comparative and demographic analysis of orang-utan genomes. *Nature* **469**,  
1276 529-533 (2011).
- 1277 21 Mattle-Greminger, M. P. *et al.* Genomes reveal marked differences in the adaptive evolution  
1278 between orangutan species. *Genome biology* **19**, 1-13 (2018).
- 1279 22 Shao, Y. *et al.* Phylogenomic analyses provide insights into primate evolution. *Science* **380**, 913-  
1280 924 (2023).
- 1281 23 Rivas-González, I., Schierup, M. H., Wakeley, J. & Hobolth, A. TRAILS: Tree reconstruction of  
1282 ancestry using incomplete lineage sorting. *Plos Genetics* **20**, e1010836 (2024).
- 1283 24 Rivas-González, I. *et al.* Pervasive incomplete lineage sorting illuminates speciation and selection  
1284 in primates. *Science* **380**, eabn4409 (2023).
- 1285 25 Frankish, A. *et al.* GENCODE: reference annotation for the human and mouse genomes in 2023.  
1286 *Nucleic acids research* **51**, D942-D949 (2023).

- 1287 26 Hoyt, S. J. *et al.* From telomere to telomere: The transcriptional and epigenetic state of human  
1288 repeat elements. *Science* **376**, eabk3112 (2022).
- 1289 27 Walker, J. A. *et al.* Orangutan Alu quiescence reveals possible source element: support for  
1290 ancient backseat drivers. *Mobile DNA* **3**, 1-12 (2012).
- 1291 28 Prado-Martinez, J. *et al.* Great ape genetic diversity and population history. *Nature* **499**, 471-475  
1292 (2013).
- 1293 29 DeGiorgio, M., Huber, C. D., Hubisz, M. J., Hellmann, I. & Nielsen, R. SweepFinder2: increased  
1294 sensitivity, robustness and flexibility. *Bioinformatics* **32**, 1895-1897 (2016).
- 1295 30 DeGiorgio, M. & Szpiech, Z. A. A spatially aware likelihood test to detect sweeps from haplotype  
1296 distributions. *PLoS genetics* **18**, e1010134 (2022).
- 1297 31 Cagan, A. *et al.* Natural selection in the great apes. *Molecular biology and evolution* **33**, 3268-  
1298 3283 (2016).
- 1299 32 McManus, K. F. *et al.* Inference of gorilla demographic and selective history from whole-genome  
1300 sequence data. *Molecular biology and evolution* **32**, 600-612 (2015).
- 1301 33 Rodriguez, O. L., Sharp, A. J. & Watson, C. T. Limitations of lymphoblastoid cell lines for  
1302 establishing genetic reference datasets in the immunoglobulin loci. *Plos one* **16**, e0261374  
1303 (2021).
- 1304 34 Sirupurapu, V., Safonova, Y. & Pevzner, P. A. Gene prediction in the immunoglobulin loci.  
1305 *Genome research* **32**, 1152-1169 (2022).
- 1306 35 Rodriguez, O. L. *et al.* Genetic variation in the immunoglobulin heavy chain locus shapes the  
1307 human antibody repertoire. *Nature communications* **14**, 4419 (2023).
- 1308 36 Radwan, J., Babik, W., Kaufman, J., Lenz, T. L. & Winternitz, J. Advances in the evolutionary  
1309 understanding of MHC polymorphism. *Trends in Genetics* **36**, 298-311 (2020).
- 1310 37 Heijmans, C. M., de Groot, N. G. & Bontrop, R. E. Comparative genetics of the major  
1311 histocompatibility complex in humans and nonhuman primates. *International Journal of*  
1312 *Immunogenetics* **47**, 243-260 (2020).
- 1313 38 Lenz, T. L., Spirin, V., Jordan, D. M. & Sunyaev, S. R. Excess of deleterious mutations around HLA  
1314 genes reveals evolutionary cost of balancing selection. *Molecular Biology and Evolution* **33**,  
1315 2555-2564 (2016).
- 1316 39 Lenz, T. L. HLA Genes: A Hallmark of Functional Genetic Variation and Complex Evolution. *HLA*  
1317 *Typing: Methods and Protocols*, 1-18 (2024).
- 1318 40 Fortier, A. L. & Pritchard, J. K. Ancient Trans-Species Polymorphism at the Major  
1319 Histocompatibility Complex in Primates. *bioRxiv*, 2022.2006. 2028.497781 (2022).
- 1320 41 Elango, N. & Yi, S. V. DNA methylation and structural and functional bimodality of vertebrate  
1321 promoters. *Molecular biology and evolution* **25**, 1602-1608 (2008).
- 1322 42 Jeong, H. *et al.* Evolution of DNA methylation in the human brain. *Nature communications* **12**,  
1323 2021 (2021).
- 1324 43 Yang, Y. *et al.* Continuous-trait probabilistic model for comparing multi-species functional  
1325 genomic data. *Cell systems* **7**, 208-218. e211 (2018).
- 1326 44 Yunis, J. J. & Prakash, O. The origin of man: a chromosomal pictorial legacy. *Science* **215**, 1525-  
1327 1530 (1982).
- 1328 45 Ventura, M. *et al.* Recurrent sites for new centromere seeding. *Genome Research* **14**, 1696-1703  
1329 (2004).
- 1330 46 Marshall, O. J., Chueh, A. C., Wong, L. H. & Choo, K. A. Neocentromeres: new insights into  
1331 centromere structure, disease development, and karyotype evolution. *The American Journal of*  
1332 *Human Genetics* **82**, 261-282 (2008).



- 1333 47 Oneda, B. *et al.* Low-level chromosomal mosaicism in neurodevelopmental disorders. *Molecular*  
1334 *syndromology* **8**, 266-271 (2017).
- 1335 48 Cardone, M. F. *et al.* Evolutionary history of chromosome 11 featuring four distinct centromere  
1336 repositioning events in Catarrhini. *Genomics* **90**, 35-43 (2007).
- 1337 49 Porubsky, D. *et al.* Recurrent inversion toggling and great ape genome evolution. *Nature*  
1338 *genetics* **52**, 849-858 (2020).
- 1339 50 Müller, S., Finelli, P., Neusser, M. & Wienberg, J. The evolutionary history of human  
1340 chromosome 7. *Genomics* **84**, 458-467 (2004).
- 1341 51 Kehrer-Sawatzki, H., Szamalek, J. M., Tänzer, S., Platzer, M. & Hameister, H. Molecular  
1342 characterization of the pericentric inversion of chimpanzee chromosome 11 homologous to  
1343 human chromosome 9. *Genomics* **85**, 542-550 (2005).
- 1344 52 Carbone, L., Ventura, M., Tempesta, S., Rocchi, M. & Archidiacono, N. Evolutionary history of  
1345 chromosome 10 in primates. *Chromosoma* **111**, 267-272 (2002).
- 1346 53 Kehrer-Sawatzki, H., Sandig, C., Goidts, V. & Hameister, H. Breakpoint analysis of the pericentric  
1347 inversion between chimpanzee chromosome 10 and the homologous chromosome 12 in  
1348 humans. *Cytogenetic and Genome Research* **108**, 91-97 (2004).
- 1349 54 Kehrer-Sawatzki, H. *et al.* Molecular characterization of the pericentric inversion that causes  
1350 differences between chimpanzee chromosome 19 and human chromosome 17. *The American*  
1351 *Journal of Human Genetics* **71**, 375-388 (2002).
- 1352 55 Cardone, M. F. *et al.* Hominoid chromosomal rearrangements on 17q map to complex regions of  
1353 segmental duplication. *Genome biology* **9**, 1-11 (2008).
- 1354 56 Goidts, V., Szamalek, J. M., Hameister, H. & Kehrer-Sawatzki, H. Segmental duplication  
1355 associated with the human-specific inversion of chromosome 18: a further example of the  
1356 impact of segmental duplications on karyotype and genome evolution in primates. *Human*  
1357 *genetics* **115**, 116-122 (2004).
- 1358 57 Misceo, D. *et al.* Evolutionary history of chromosome 20. *Molecular Biology and Evolution* **22**,  
1359 360-366 (2005).
- 1360 58 Ventura, M. *et al.* Gorilla genome structural variation reveals evolutionary parallelisms with  
1361 chimpanzee. *Genome research* **21**, 1640-1649 (2011).
- 1362 59 Capozzi, O. *et al.* A comprehensive molecular cytogenetic analysis of chromosome  
1363 rearrangements in gibbons. *Genome Research* **22**, 2520-2528 (2012).
- 1364 60 Catacchio, C. R. *et al.* Inversion variants in human and primate genomes. *Genome Research* **28**,  
1365 910-920 (2018).
- 1366 61 Kronenberg, Z. N. *et al.* High-resolution comparative analysis of great ape genomes. *Science* **360**,  
1367 eaar6343 (2018).
- 1368 62 Maggiolini, F. A. M. *et al.* Single-cell strand sequencing of a macaque genome reveals multiple  
1369 nested inversions and breakpoint reuse during primate evolution. *Genome research* **30**, 1680-  
1370 1693 (2020).
- 1371 63 Mercuri, L. *et al.* A high-resolution map of small-scale inversions in the gibbon genome. *Genome*  
1372 *Research* **32**, 1941-1951 (2022).
- 1373 64 Nuttle, X. *et al.* Emergence of a *Homo sapiens*-specific gene family and chromosome 16p11. 2  
1374 CNV susceptibility. *Nature* **536**, 205-209 (2016).
- 1375 65 Paparella, A. *et al.* Structural Variation Evolution at the 15q11-q13 Disease-Associated Locus.  
1376 *International Journal of Molecular Sciences* **24**, 15818 (2023).
- 1377 66 Zody, M. C. *et al.* Evolutionary toggling of the MAPT 17q21. 31 inversion region. *Nature genetics*  
1378 **40**, 1076-1083 (2008).

- 1379 67 Maggiolini, F. A. *et al.* Genomic inversions and GOLGA core duplicons underlie disease instability  
1380 at the 15q25 locus. *PLoS Genetics* **15**, e1008075 (2019).
- 1381 68 Antonacci, F. *et al.* Characterization of six human disease-associated inversion polymorphisms.  
1382 *Human molecular genetics* **18**, 2555-2566 (2009).
- 1383 69 Mao, Y. *et al.* Structurally divergent and recurrently mutated regions of primate genomes. *Cell*  
1384 **187**, 1547-1562. e1513 (2024).
- 1385 70 Mangan, R. J. *et al.* Adaptive sequence divergence forged new neurodevelopmental enhancers  
1386 in humans. *Cell* **185**, 4587-4603. e4523 (2022).
- 1387 71 Gedman, G. L. *et al.* Convergent gene expression highlights shared vocal motor microcircuitry in  
1388 songbirds and humans. *bioRxiv*, 2022.2007. 2001.498177 (2022).
- 1389 72 Lovell, P. V. *et al.* ZEBRA: Zebra finch Expression Brain Atlas—A resource for comparative  
1390 molecular neuroanatomy and brain evolution studies. *Journal of Comparative Neurology* **528**,  
1391 2099-2131 (2020).
- 1392 73 Kirilenko, B. M. *et al.* Integrating gene annotation with orthology inference at scale. *Science* **380**,  
1393 eabn3107 (2023).
- 1394 74 Willcox, B. J. *et al.* FOXO3A genotype is strongly associated with human longevity. *Proceedings*  
1395 *of the National Academy of Sciences* **105**, 13987-13992 (2008).
- 1396 75 Nassar, L. R. *et al.* The UCSC genome browser database: 2023 update. *Nucleic acids research* **51**,  
1397 D1188-D1195 (2023).
- 1398 76 Zemke, N. R. *et al.* Conserved and divergent gene regulatory programs of the mammalian  
1399 neocortex. *Nature* **624**, 390-402 (2023).
- 1400 77 van Sluis, M. *et al.* Human NORs, comprising rDNA arrays and functionally conserved distal  
1401 elements, are located within dynamic chromosomal regions. *Genes & Development* **33**, 1688-  
1402 1701 (2019).
- 1403 78 Guarracino, A. *et al.* Recombination between heterologous human acrocentric chromosomes.  
1404 *Nature* **617**, 335-343 (2023).
- 1405 79 Chiatante, G., Giannuzzi, G., Calabrese, F. M., Eichler, E. E. & Ventura, M. Centromere destiny in  
1406 dicentric chromosomes: new insights from the evolution of human chromosome 2 ancestral  
1407 centromeric region. *Molecular biology and evolution* **34**, 1669-1681 (2017).
- 1408 80 Stults, D. M., Killen, M. W., Pierce, H. H. & Pierce, A. J. Genomic architecture and inheritance of  
1409 human ribosomal RNA gene clusters. *Genome research* **18**, 13-18 (2008).
- 1410 81 Agrawal, S. & Ganley, A. R. The conservation landscape of the human ribosomal RNA gene  
1411 repeats. *PloS one* **13**, e0207531 (2018).
- 1412 82 Sweeten, A. P., Schatz, M. C. & Phillippy, A. M. ModDotPlot—Rapid and interactive visualization  
1413 of complex repeats. *bioRxiv* (2024).
- 1414 83 Kille, B., Garrison, E., Treangen, T. J. & Phillippy, A. M. Minmers are a generalization of  
1415 minimizers that enable unbiased local Jaccard estimation. *Bioinformatics* **39**, btad512 (2023).
- 1416 84 Logsdon, G. A. *et al.* The variation and evolution of complete human centromeres. *Nature* **629**,  
1417 136-145 (2024).
- 1418 85 Cheeseman, I. M. The kinetochore. *Cold Spring Harbor perspectives in biology* **6**, a015826 (2014).
- 1419 86 Musacchio, A. & Desai, A. A molecular view of kinetochore assembly and function. *Biology* **6**, 5  
1420 (2017).
- 1421 87 Logsdon, G. A. *et al.* The structure, function and evolution of a complete human chromosome 8.  
1422 *Nature* **593**, 101-107 (2021).
- 1423 88 Gershman, A. *et al.* Epigenetic patterns in a complete human genome. *Science* **376**, eabj5089  
1424 (2022).

- 1425 89 Ventura, M. *et al.* The evolution of African great ape subtelomeric heterochromatin and the  
1426 fusion of human chromosome 2. *Genome research* **22**, 1036-1049 (2012).
- 1427 90 Lisitsyn, N. *et al.* Isolation of rapidly evolving genomic sequences: construction of a differential  
1428 library and identification of a human DNA fragment that does not hybridize to chimpanzee DNA.  
1429 *Biomedical Science* **1**, 513-516 (1990).
- 1430 91 Koga, A., Hirai, Y., Hara, T. & Hirai, H. Repetitive sequences originating from the centromere  
1431 constitute large-scale heterochromatin in the telomere region in the siamang, a small ape.  
1432 *Heredity* **109**, 180-187 (2012).
- 1433 92 Hirai, H. *et al.* Chimpanzee chromosomes: retrotransposable compound repeat DNA  
1434 organization (RCRO) and its influence on meiotic prophase and crossing-over. *Cytogenetic and*  
1435 *genome research* **108**, 248-254 (2004).
- 1436 93 Wallace, B. & Hulten, M. Meiotic chromosome pairing in the normal human female. *Annals of*  
1437 *human genetics* **49**, 215-226 (1985).
- 1438 94 Marques-Bonet, T. & Eichler, E. in *Cold Spring Harbor symposia on quantitative biology*. 355-362  
1439 (Cold Spring Harbor Laboratory Press).
- 1440 95 Cheng, Z. *et al.* A genome-wide comparison of recent chimpanzee and human segmental  
1441 duplications. *Nature* **437**, 88-93 (2005).
- 1442 96 Marques-Bonet, T. *et al.* A burst of segmental duplications in the genome of the African great  
1443 ape ancestor. *Nature* **457**, 877-881 (2009).
- 1444 97 Sharp, A. J. *et al.* A recurrent 15q13. 3 microdeletion syndrome associated with mental  
1445 retardation and seizures. *Nature genetics* **40**, 322-328 (2008).
- 1446 98 Jiang, Z. *et al.* Ancestral reconstruction of segmental duplications reveals punctuated cores of  
1447 human genome evolution. *Nature genetics* **39**, 1361-1368 (2007).
- 1448 99 Antonacci, F. *et al.* Palindromic GOLGA8 core duplicons promote chromosome 15q13. 3  
1449 microdeletion and evolutionary instability. *Nature genetics* **46**, 1293-1302 (2014).
- 1450 100 Bernstein, B. E. *et al.* A bivalent chromatin structure marks key developmental genes in  
1451 embryonic stem cells. *Cell* **125**, 315-326 (2006).
- 1452 101 Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003-1007  
1453 (2002).
- 1454 102 Sudmant, P. H. *et al.* Evolution and diversity of copy number variation in the great ape lineage.  
1455 *Genome research* **23**, 1373-1382 (2013).
- 1456 103 McStay, B. The p-arms of human acrocentric chromosomes play by a different set of rules.  
1457 *Annual Review of Genomics and Human Genetics* **24**, 63-83 (2023).
- 1458 104 Lloyd Jr, F. & Goldrosen, M. The production of a bispecific anti-CEA, anti-hapten (4-amino-  
1459 phthalate) hybrid-hybridoma. *Journal of the National Medical Association* **83**, 901 (1991).
- 1460 105 King, C. A model for transposon-based eucaryote regulatory evolution. *Journal of theoretical*  
1461 *biology* **114**, 447-462 (1985).
- 1462 106 Navarro, A. & Barton, N. H. Chromosomal speciation and molecular divergence--accelerated  
1463 evolution in rearranged chromosomes. *Science* **300**, 321-324 (2003).
- 1464 107 Fiddes, I. T. *et al.* Human-specific NOTCH2NL genes affect notch signaling and cortical  
1465 neurogenesis. *Cell* **173**, 1356-1369. e1322 (2018).
- 1466 108 Dennis, M. Y. *et al.* Evolution of human-specific neural SRGAP2 genes by incomplete segmental  
1467 duplication. *Cell* **149**, 912-922 (2012).
- 1468 109 Schmidt, E. R., Kupferman, J. V., Stackmann, M. & Polleux, F. The human-specific paralogs  
1469 SRGAP2B and SRGAP2C differentially modulate SRGAP2A-dependent synaptic development.  
1470 *Scientific reports* **9**, 18692 (2019).

1471 110 Guitart, X. *et al.* Independent expansion, selection and hypervariability of the TBC1D3 gene  
1472 family in humans. *bioRxiv* (2024).  
1473 111 Vollger, M. R. *et al.* Segmental duplications and their variation in a complete human genome.  
1474 *Science* **376**, eabj6965 (2022).  
1475 112 Moralli, D. & Monaco, Z. L. Gene expressing human artificial chromosome vectors: Advantages  
1476 and challenges for gene therapy. *Experimental Cell Research* **390**, 111931 (2020).  
1477 113 Logsdon, G. A. & Eichler, E. E. The dynamic structure and rapid evolution of human centromeric  
1478 satellite DNA. *Genes* **14**, 92 (2022).  
1479 114 Hirai, H. *et al.* Structural variations of subterminal satellite blocks and their source mechanisms  
1480 as inferred from the meiotic configurations of chimpanzee chromosome termini. *Chromosome*  
1481 *Research* **27**, 321-332 (2019).  
1482 115 Schoch, C. L. *et al.* NCBI Taxonomy: a comprehensive update on curation, resources and tools.  
1483 *Database* **2020**, baaa062 (2020).  
1484 116 Hey, J. The divergence of chimpanzee species and subspecies as revealed in multipopulation  
1485 isolation-with-migration analyses. *Molecular biology and evolution* **27**, 921-933 (2010).  
1486 117 Roos, C. Phylogeny and classification of gibbons (Hylobatidae). *Evolution of gibbons and siamang:*  
1487 *Phylogeny, morphology, and cognition*, 151-165 (2016).  
1488