

1 Human-specific expansion of 22q11.2 low copy repeats

2 Lisanne Vervoort¹, Nicolas Dierckxsens¹, Zjef Pereboom^{2,3}, Oronzo Capozzi⁴, Mariano Rocchi⁴, Tamim
3 H. Shaikh⁵, Joris R. Vermeesch^{1*}

4 ¹ Department of Human Genetics, KU Leuven, Leuven, Belgium

5 ² Centre for Research and Conservation, Royal Zoological Society of Antwerp, Antwerp, Belgium

6 ³ Department of Biology, Evolutionary Ecology Group, Antwerp University, Antwerp, Belgium

7 ⁴ Department of Biology, University of Bari, Bari, Italy

8 ⁵ Department of Pediatrics, Section of Clinical Genetics and Metabolism, University of Colorado Denver, Aurora,
9 Colorado, USA

10

11 *Corresponding author

12 E-mail: joris.vermeesch@uzleuven.be (JV)

13

14

15

16

17

18

19

20

21

22

23

24 **Abstract**

25 Segmental duplications or low copy repeats (LCRs) constitute complex regions interspersed in the
26 human genome. They have contributed significantly to human evolution by stimulating neo- or sub-
27 functionalization of duplicated transcripts. The 22q11.2 region carries eight LCRs (LCR22s). One of
28 these LCR22s was recently reported to be hypervariable in the human population. It remains unknown
29 whether this variability exists also in non-human primates. To assess the inter- and intra-species
30 variability, we *de novo* assembled the region in non-human primates by a combination of optical
31 mapping techniques. Orangutan carries three LCR22-mediated inversions of which one is the ancient
32 haplotype since it is also present in macaque. Using fiber-FISH, lineage-specific differences in LCR22
33 composition were mapped. The smallest and likely ancient haplotype is present in the chimpanzee,
34 bonobo and rhesus macaque. The absence of intra-species variation in chimpanzee indicates the
35 LCR22-A expansion to be unique to the human population. Further, we demonstrate that LCR22-
36 specific genes are expressed in both human and non-human primate neuronal cell lines and show
37 expression of several primate LCR22 transcripts for the first time. The human-specificity of the
38 expansions suggest an important role for the region in human evolution and adaptation.

39 **Author summary**

40 Low copy repeats or segmental duplications are DNA segments composed of various subunits which
41 are duplicated across the genome. Due to the high level of sequence identity between these segments,
42 homologous regions can misalign, resulting in reciprocal deletions and duplications, classified as
43 genomic disorders. These regions are subject to structural variation in the human population. We
44 recently detected extreme structural variation in one of the most complex segmental duplication
45 regions of the human genome, the low copy repeats on chromosome 22 (LCR22s). Rearrangements
46 between the LCR22s result in the 22q11.2 deletion/duplication syndrome, the most common human
47 genomic disorder. However, it remains unknown whether this variability is human-specific. In this
48 study, we investigated those LCR22s in several individuals of the different great apes and macaque.

49 We show only the smallest haplotype is present without any intra-species variation in the *Pan* genus,
50 our closest ancestors. Hence, LCR22 expansions are human-specific, suggesting a role of these LCR22s
51 in human evolution and adaptation and hypothesize the region contributes to the 22q11.2 deletion
52 syndrome inter-patient phenotypic variability.

53 **Introduction**

54 Segmental duplications or low copy repeats (LCRs) constitute over 5% of the genome [1] and are
55 complex patchworks of duplicated DNA fragments varying in length with over 90% sequence identity
56 [2,3]. This high sequence homology has so far impeded the accurate mapping and assembly of these
57 regions in the human reference genome [4,5]. Although it has become evident that assembly using
58 short read sequencing is unable to resolve these complex regions, some LCRs are often too long and
59 complex even for more recently developed long read technologies to resolve [4,6]. In addition, large
60 structural variation amongst haplotypes complicates the assembly of these LCR containing regions [7].
61 As a consequence, LCRs remain poorly mapped and characterized, despite their functional importance
62 in evolution and disease.

63 The impact of these LCRs on primate and human evolution is increasingly recognized [8,9]. It is
64 estimated that the origin of the LCRs coincide with the divergence of New and Old World Monkeys,
65 35-40 million years ago [10]. However, a genomic duplication burst was observed in the great ape
66 lineage, creating lineage-specific LCRs which are highly copy number variable [11]. These LCR-
67 containing regions in other great ape reference genomes are also enriched for gaps, since they are
68 subject to similar assembly difficulties as those encountered in the assembly of these regions in the
69 human reference genome [12,13]. In humans, the 22q11.2 region contains a relatively higher
70 proportion of LCRs compared with the rest of the genome. The origin of the human chromosome 22
71 LCRs (LCR22s) is concordant with the evolutionary timeline of LCRs in general. No duplicated
72 orthologous LCR22 sequences are present in the mouse [14,15], and FISH mapping and sequencing
73 experiments suggest lineage-specific LCR22 variation and mosaic structure in great apes [15–18].

74 However, since techniques to resolve the structure of the LCR22s were lacking, the great ape LCR22s
75 have not been assembled and their composition and structure remain unresolved.

76 Due to the high level of sequence identity, homologous segments within LCRs can misalign during
77 meiosis, via a mechanism known as non-allelic homologous recombination (NAHR), resulting in
78 genomic rearrangements including deletions, duplications, and inversions [19]. The four most proximal
79 LCR22 blocks are referred to as LCR22-A, -B, -C, and -D [20]. NAHR between these LCR22s underlies the
80 formation of the recurrent deletions associated with the 22q11.2 deletion syndrome (22q11.2DS)
81 (MIM: 188400/192430), the most common microdeletion disorder in humans [20] as well as the
82 reciprocal duplications of this region often associated with abnormal phenotypes (MIM: 608363) [20].

83 We demonstrated hypervariability in the organization and the copy number of duplicons within
84 LCR22s, especially LCR22-A [21]. By combining fiber-FISH and Bionano optical mapping we assembled
85 the LCR22s *de novo* and uncovered over 30 haplotypes of LCR22-A, with alleles ranging in size from
86 250 kb to 2000 kb within 169 normal diploid individuals [21]. Pastor et al. recently expanded the LCR22-
87 A catalogue by haplotyping the complete alleles of 30 22q11.2DS families [22]. To determine whether
88 this extreme haplotype variability is human-specific, we set out to chart the inter- and intra-species
89 variability of these LCR22s in non-human primates (S1 Table). The structures of the great apes,
90 including five chimpanzees (*Pan troglodytes*), one bonobo (*Pan paniscus*), two gorillas (*Gorilla gorilla*
91 and *Gorilla berengei graueri*), six orangutans (*Pongo pygmaeus* and *Ponglo abelii*), and one rhesus
92 macaque (Old World Monkey, *Macaca malutta*) were analyzed by using an LCR22-specific fiber-FISH.
93 To map the broader region, one representative of each species was analyzed by Bionano optical
94 mapping. We demonstrate the non-human primate haplotypes to be less complex compared to
95 humans. No intra-species variability similar to humans was observed suggesting that the
96 hypervariability of the human LCR22-A haplotype is of recent origin.

97 **Results**

98 **Conservation of the syntenic 22q11.2 locus**

99 To assess whether the overall structure of the 22q11.2 region was conserved, the syntenic regions
100 were investigated by Bionano optical mapping in non-human primate cell lines. Optical mapping allows
101 the detection of structural variation, by imaging long fluorescently labeled DNA molecules (>150kb)
102 followed by *de novo* assembly and local haplotyping [23]. Subsequently, the assembled alleles were
103 compared to the human reference genome hg38 (Fig 1A). The resulting 22q11.2 syntenic assemblies
104 were validated by fiber-FISH experiments using BAC (bacterial artificial chromosome) probes targeting
105 the regions flanking the proximal LCR22s (schematic representation in Fig 1B, S2 Table). Due to the low
106 mapping rate between the rhesus macaque sample and the human reference genome, the Bionano
107 analysis in this non-human primate could not be performed and the composition (Fig 1G) is only based
108 on fiber-FISH results.

109 The order and organization of LCR22-A through -H in chimpanzee (Fig 1C, S1A Fig), bonobo (Fig 1C, S1B
110 Fig), and gorilla (Fig 1D, S1C Fig) is identical to human. In contrast, three large rearrangements were
111 observed in the syntenic 22q11.2 locus of the orangutan (Fig 1E-F). First, the region between LCR22-F
112 and -H, including LCR22-G, is inverted. Second, an inversion is present between the LCR22-A and -F
113 blocks. Third, the orientation between LCR22-C and -D is not inverted compared with the human
114 reference. This could be interpreted as an extra inversion between LCR22-C and -D following the
115 rearrangement between LCR22-A and -F. However, investigating this locus in the macaque by fiber-
116 FISH uncovered the presence of this LCR22-C/D inversion, without the larger LCR22-A/F inversion (Fig
117 1G). Since we could not investigate the distal LCR22s, an inversion between these LCR22-F and -H
118 cannot be excluded. Hence, despite the unstable nature of the LCR22s themselves, the structural
119 organization between the LCR22 blocks is conserved between gorilla, chimpanzee, bonobo, and
120 human. Inversions, typically flanked by LCRs, are present in the orangutan and rhesus macaque
121 haplotype.

122 **Evolutionary analysis of LCR22-A**

123 The current reference genomes of great apes, except for the chimpanzee, are enriched for sequence
124 gaps within the loci orthologous to the LCR22s. As a consequence, it was not possible to fully rely on
125 the reference sequences and alleles had to be *de novo* assembled. For this, an LCR22-specific fiber-
126 FISH method was applied, which has proven its value to resolve these complex structures in humans
127 (Fig 2A) [21]. Exact probe identities were checked by changing the fluorophores of color-identical
128 probes (S2-5 Figs).

129 Based on the extensive variability observed in the overall size and duplicon content of human LCR22-
130 A (Fig 2B-C), we wanted to determine whether similar variation exists in the other great apes and
131 rhesus macaque. Towards this end, five chimpanzees, one bonobo, two gorillas, six orangutans, and
132 one rhesus macaque were analyzed (S1 Table). In contrast to the human variability, no structural
133 variation was observed in any of the ten alleles of LCR22-A observed in the chimpanzee samples (S6
134 Fig). In addition, both bonobo alleles also had the exact same composition as those in the chimpanzee.
135 This haplotype (Fig 2C) is the smallest haplotype observed in human. However, this haplotype is rare
136 in humans and only observed as a heterozygous allele in 5 of 169 human samples analyzed [21].

137 In the gorilla, the proximal and distal end are similar to the chimpanzee haplotype, except for a small
138 insertion (Fig 2C). This is considered as a gorilla-specific insertion, since it is not present in the other
139 non-human primate or human haplotypes. The same allele was observed in all four haplotypes of both
140 gorilla cell lines. In addition to the large-scale rearrangements in the orangutan, we also observed
141 major differences in the LCR22 compositions compared to the alleles of the other great apes (Fig 2C).
142 First, the SD22-5 (green) duplicon, the distal delineating LCR22-A end in other great apes, is located in
143 the middle of the allele, surrounded by SD22-6 duplicons. Second, tandem repeats, of probe
144 compositions (indicated between brackets in Fig 2C) characterize the proximal and distal end of the
145 allele. This characteristic is different from the interspersed mosaic nature of the LCR22s in humans. In
146 addition, structural variation is observed within these repeats in the six orangutan samples (S3 Table).
147 Thus, the haplotypes observed in the orangutan are very different from those observed in other great

148 apes (Fig 2C). In contrast, the rhesus macaque haplotype is mostly identical to the small chimpanzee
149 haplotype composition, except for an ~30kb insertion of unknown origin separating the SD22-5 and
150 SD22-6 duplicons.

151 In order to validate these results, we correlated the fiber-FISH data with the corresponding chimpanzee
152 reference genome. The human locus chr22:18,044,268-19,017,737 including the LCR22-A allele, can
153 be traced to the chimpanzee locus chr22:2,635,159-2,386,886 in the most recent reference genome
154 (Clint_PTRv2/panTro6/January 2018). The fiber-FISH probe order predicted from this sequence exactly
155 matches the obtained fiber-FISH pattern. Hence, this extra independent chimpanzee allele confirms
156 the presence of a single LCR22-A haplotype in chimpanzee.

157 In conclusion, due to the absence of LCR22-A structural variation in our closest ancestors, LCR22-A
158 hypervariability can be considered as human-specific.

159 **Evolutionary analysis of LCR22-B/C/D**

160 While LCR22-A is hypervariable in human genomes, LCR22-B and LCR22-C showed no variations, and
161 only six different alleles were observed for LCR22-D (Fig 3A-D) [21]. To evaluate the evolution of these
162 LCR22s and asses intra-species variation in non-human primates, we investigated the syntenic LCR22-
163 B, -C, and -D haplotypes in great apes and rhesus macaque by fiber-FISH as well. Since LCR22-B and -C
164 could be small and hard to distinguish above fiber-FISH noise, the probe set was supplemented with
165 BAC probes flanking these LCR22s (S2 Table).

166 For LCR22-B, the chimpanzee and bonobo were identical to the human haplotype, while the gorilla
167 haplotype was similar, with the deletion of one duplicon (SD22-2) (Fig 3E). In the orangutan, the distal
168 part is substituted by two probes (A3-D2). An extra insertion between these two probes creates the
169 haplotype of the rhesus macaque. LCR22-C carries lineage-specific insertions and deletions in the *Pan*
170 and *Gorilla* genus, while in the orangutan and rhesus macaque it is reduced to only one probe (D1) (Fig
171 3E). The human LCR22-D haplotype is subjected to structural variation, mainly in the SD22-3 duplicon

172 [21]. One variant, an internal inversion (indicated by the magenta arrow in Fig 3C-D), is present in 37%
173 of the human haplotypes. The same variant was observed in a heterozygous state in two LCR22-D
174 chimpanzee alleles (Fig 3E, S6 Fig), suggesting this variant precedes the split of the human lineage. The
175 proximal start and distal end were conserved in Gorilla, with extra insertions compared to the human
176 and *Pan* haplotype (Fig 3E). No structural variation was found at the distal end in these four
177 investigated alleles. However, we predict that this structural variant is likely to be present in the gorilla
178 population as well, since the composition is the same as in human, chimpanzee, and bonobo. The
179 LCR22-D haplotype in orangutan and rhesus macaque is composed of only two probes (Fig 3E). To
180 conclude, LCR22-B, -C, and -D haplotypes start to evolve towards their human structures in a common
181 ancestor of *Gorilla*, *Pan* and *Homo*, based on the very short haplotypes found in orangutan and
182 macaque.

183 **LCR22-specific transcript expression in human and non-human primates**

184 According to the human reference transcriptome and the GTEx expression profiles [24], the LCR22s
185 contain several expressed genes, pseudogenes, and long non-coding RNAs (Figs 2B, 3B, 3D). However,
186 an expression study analyzing the LCR22 genes and their paralogs has not yet been accomplished in
187 human nor in non-human primates. In addition, very few LCR22-specific genes have been annotated
188 in the non-human primates. Based on the fiber-FISH composition of the non-human primate LCR22
189 alleles, we could predict the presence or absence of certain transcripts, since probes used in the fiber-
190 FISH assays typically cover those genes. Hence, we set out to explore the conservation of the LCR22
191 specific genes and the expression pattern similarities with humans, by mining published primate
192 transcriptome datasets. We explored gene expression in two publicly available brain transcriptome
193 studies on human and non-human primates [25–27] (Table 1). The brain model was chosen since it is
194 known that part of the human LCR22 transcripts are expressed in this tissue type, and genes within
195 LCRs in general play a role in synaptogenesis, neuronal migration, and neocortical expansion in the
196 human lineage [8,24].

197 **Table 1: Transcriptome analysis of LCR22 genes for two publicly available transcriptome studies**

		PRJNA393104			PRJNA415990		
		Homo sapiens	Pan troglodytes	Gorilla gorilla	Homo sapiens	Pan troglodytes	Pongo abelii
LCR22-A	USP18	Green	Green	Green	Green	Green	Green
	TMEM191B	Green	Yellow	Dark Green	Yellow	Green	Yellow
	PI4KAP1	Green	Red	Red	Green	Red	Red
	RIMBP3	Red	Red	Red	Green	Yellow	Yellow
	DGCR6	Green	Green	Green	Green	Yellow	Yellow
	PRODH	Green	Green	Green	Green	Green	Green
	DGCR5	Green	Dark Green	Dark Green	Dark Green	Dark Green	Red
	DGCR2	Green	Green	Green	Green	Green	Green
LCR22-B	DGCR6L	Green	Green	Yellow	Green	Green	Dark Green
LCR22-C	TMEM191A	Green	Dark Green	Yellow	Green	Dark Green	Yellow
	PI4KA	Green	Green	Green	Green	Green	Green
LCR22-D	HIC2	Green	Green	Green	Green	Green	Green
	TMEM191C	Green	Green	Yellow	Yellow	Yellow	Yellow
	PI4KAP2	Green	Dark Green	Dark Green	Green	Dark Green	Red
	UBE2L3	Green	Green	Green	Green	Green	Green

198 *Confirmed transcripts are indicated as green, absent transcripts as red, and inconclusive as yellow. Dark green boxes present*
 199 *novel reported transcripts.*

200 We relied on two independent methods for the detection of transcripts: alignment and *de novo*
 201 assembly. Transcript assembly or alignment were seen as inconclusive when the coverage was below
 202 four reads or when paralogs could not be distinguished from each other. The latter was more
 203 frequently the case for non-human primates, as we lack a reference sequence of LCR22s and few
 204 orthologous transcripts have yet been annotated. Consequently, the *de novo* assembly has led to the
 205 discovery of several new transcripts for each of the non-human primates and a number of new splice
 206 forms of the *TMEM191* transcripts in the human genome. Moreover, *DGCR5* and *TMEM191A* are
 207 detected for the first time in non-human species (Table 1).

208 We attempted to distinguish between paralogs by adapting an assembly method originally developed
 209 for heteroplasmy detection in mitochondrial genomes [28]. Since this method needs sufficient
 210 coverage, we selected *PI4KA* and the two pseudogenes *PI4KAP1* and *PI4KAP2*. *PI4KA* was present in
 211 high coverage for all samples, while *PI4KAP1* was only found in human and *PI4KAP2* was only absent
 212 for orangutan (S1 Appendix). This pattern correlates with the fiber-FISH duplicons. *PI4KAP1* is located
 213 in SD22-3 in LCR22-A, which is unique to human (Fig 2). *PI4KAP2* is located in SD22-3 of LCR22-D, which
 214 is present in all great apes (Fig 3E). However, the SD22-3 duplicon in orangutan probably expresses the
 215 *PI4KA* gene, since the partial SD22-3 is absent in LCR22-C and the region is inverted. Therefore, absence

216 of the *PI4KAP2* pseudogene in this species correlates with the absence of an extra SD22-3 duplicon in
217 the fiber-FISH pattern. Although we were also able to identify some of the *TMEM191* paralogs with
218 this method, low coverage, and the presence of multiple splice variants made it impossible to verify all
219 paralogs in each study.

220 We looked at the expression of 39 LCR22 genes for two publicly available transcriptome studies
221 (PRJNA393104 and PRJNA415990). Genes without distinct evidence of expression in any of the samples
222 were excluded from Table 1 (a full list can be found in the S4 Table). For both transcriptome studies,
223 there is clear evidence of LCR22 specific transcripts with a conserved expression pattern across both
224 human and non-human primates (Table 1).

225 **Discussion**

226 FISH mapping studies of metaphase chromosomes from great apes using 22q11.2 BAC probes and
227 analysis of sequencing data had demonstrated the LCR22 expansion to precede the divergence of old
228 and New world monkeys, and suggested species specific LCR22 variation had occurred during primate
229 speciation [15–18]. However, the FISH studies were mainly focusing on interrogation of the copy
230 number of a limited number of genic segments and sequencing analysis was inevitably interpreted
231 against human reference genome 37 (hg19), carrying important inconsistencies compared to the most
232 recent reference genome hg38. By *de novo* assembling the LCR22s using LCR22-specific probes in the
233 fiber-FISH assay we resolved the haplotype composition in five chimpanzees, one bonobo, two gorillas,
234 six orangutans and a macaque. This evolutionary analysis of the complex segmental duplications on
235 chromosome 22 in different members of each species reveals that hypervariability of the LCR22-A
236 allele is human-specific.

237 Human-specific expansions of LCR22s had introduced additional substrates for LCR22-mediated
238 rearrangements which can result in genomic disorders associated with the 22q11.2 locus. As
239 demonstrated by Demaerel et al. [21], the region of overlap between LCR22-A and LCR22-D is within a
240 long stretch of homology encompassing SD22-4 flanked by SD22-6 on both sides, where recombination

241 was shown to have taken place in case of an LCR22-A/D deletion. This locus is not present in any of the
242 LCR22 blocks of the *Pan* genus. Pastor et al. [22] narrowed this region to SD22-6, the duplicon
243 encompassing the *FAM230* gene member. Guo et al. [29] predicted the rearrangement breakpoint was
244 located in the *BCR* (Breakpoint Cluster Region) locus, present in the distal part of SD22-4 (end of arrow).
245 This locus was present twice in the *Pan* haplotype, once in LCR22-C and once in LCR22-D, but in
246 opposite orientation preventing recombination leading to deletions and duplications. In the human
247 lineage, the prevalence of both SD22-4 and SD22-6 increases in LCR22-A and LCR22-D. Hence, human-
248 specific expansion of the region likely increases the susceptibility of chromosome 22q11.2 to
249 rearrangements, similar to observations made in other diseases resulting from LCR-mediated
250 rearrangement [30].

251 The *Pan-Rhesus* LCR22-A haplotype is the smallest amongst the apes and was present in a homozygous
252 way. Hence, this is likely the ancestral haplotype, with lineage-specific insertions and deletions. This
253 ancestral haplotype is composed of three core duplicons (SD22-1, SD22-6, and SD22-5). Compared
254 with most human haplotypes, three other core duplicons are missing (SD22-2, SD22-3, and SD22-4).
255 These elements are present in respectively LCR22-B/D, LCR22-D, and LCR22-C of the *Pan* genus.
256 Babcock et al. [31] presented a model of insertion of duplicons into LCR22-A combining homologous
257 recombination in the absence of a crossover with non-homologous repair. The model was proposed
258 for an interchromosomal recombination, but can be applied for intrachromosomal events as well.
259 Following insertion in the LCR22-A block, allelic homologous recombination is a possible mechanism
260 for the creation and expansion of new haplotypes. Since *Alu* elements are frequently delineating LCR
261 blocks in general and on chromosome 22 specifically, they form a perfect substrate for this type of
262 rearrangements [18,31,32].

263 This study provides the hitherto highest resolution map of the LCR22s across our closest evolutionary
264 relatives, showing lineage-specific inversions, insertions, and deletions. Bionano optical mapping
265 identifies three LCR22-mediated inversions in the orangutan lineage, and one in the rhesus macaque.

266 A previous study focusing on the identification of inversion variants between human and primate
267 genomes, observed the inversion between LCR22-C/D in the rhesus macaque, but was not able to
268 identify any in the orangutan [33]. The extreme LCR22 amplification in gorilla, as described by Babcock
269 et al. [17], was not identified in this study. It seems likely that some of the LCR22 duplicons are
270 amplified at other regions in the gorilla genome. Since metaphase and interphase FISH studies have a
271 lower level of resolution, the exact location of these amplifications is not known but some
272 amplifications appear to be located at telomeric bands. Hence, they will not be identified by our LCR22
273 targeted fiber-FISH analysis.

274 It remains to be uncovered how this LCR22 variability influences the human phenotype and which
275 elements in the regions are under selective pressure. Human-specific expansions were also observed
276 in LCRs present on other chromosomes that are known to cause genomic disorders [34,35] and have
277 been associated with human adaptation and evolution [8]. Gene duplications are a source for
278 transcript innovation and expansion of the transcript diversity due to exon shuffling, novel splice
279 variants, and fusion transcripts by the juxtaposition of duplicated subunits [36–38]. The human-specific
280 *SRGAP2C* gene on chromosome 1 is an example of neofunctionalization [39]. The LCR-located gene,
281 created by incomplete duplication, exerts an antagonistic effect on the ancestral *SRGAP2A* transcripts,
282 resulting in human-specific neocortical changes [39,40]. Another example is the partial
283 intrachromosomal duplication of *ARHGAP11A* (chromosome 15) leading to *ARHGAP11B*, which is
284 associated with brain adaptations during evolution [41]. Hence, human-specific (incomplete)
285 duplications of genic segments can render those genes into functional paralogs with possible
286 innovating functions. These genes present evidence of positive selection and show a general increase
287 in copy number in the human lineage [11].

288 The LCRs on chromosome 22 might be considered as an extreme source for expansion of the transcript
289 catalogue. Several genes are present in the copy number variable duplicons of the LCR22s: *PI4KA*
290 (SD22-3) and paralogs *PI4KAP1* and *PI4KAP2*, *RIMBP3* (SD22-3) and paralogs *RIMBP3B* and *RIMBP3C*,

291 *FAM230* non-coding RNAs (SD22-6). First, most of these paralogs are not well characterized and
292 classified (possibly incorrectly) as non-coding. We have clearly demonstrated expression of *PI4KA*
293 (LCR22-C) and its non-processed pseudogenes *PI4KAP1* (LCR22-A) and *PI4KAP2* (LCR22-D). The
294 expression is correlated with the presence or absence of the SD22-3 duplicons in the different species.
295 Second, due to the high variability of these haplotypes in the human population, not every individual
296 will have the same LCR22 genes or genic copy number. For example, due to this LCR22-A haplotype
297 variability, the *PI4KAP1* pseudogene is not present in every human. Hence, the presence of specific
298 paralogs and their possible functional importance might be underestimated. Transcriptome studies
299 may help to unravel the role of these human-specific expansions. Short-read RNA-Seq datasets can be
300 used to detect transcript expression (Table 1, S4 Table). Due to the duplicated nature of the LCR22s,
301 paralogs share a high level of sequence identity. Therefore, short-read data are not always able to
302 resolve the differences between transcripts arising from different paralogs. To unravel the predicted
303 transcriptome complexity and the contribution of individual paralogs, long read full-length
304 transcriptome analysis will be required. In addition, tools to obtain the full-length sequences of the
305 LCR22s and map the paralog variability will be essential to fully comprehend the extent of sequence
306 variation present. Our analysis focused on brain RNA-Seq datasets because of the importance of LCRs
307 in the human brain development. However, absence of a transcript in the dataset does not
308 automatically mean that the gene is absent. For example, *FAM230* and *RIMBP3* paralogs are mainly
309 expressed in testis [42–44]. LCR22-specific tissue transcriptome mapping or mining of the human cell
310 atlas will be required to determine the full impact of the genes in those regions.

311 In summary, optical mapping of the LCRs on chromosome 22 unraveled lineage-specific differences
312 between non-human primates and demonstrated the LCR22-A expansions and variability unique to
313 the human population. It seems likely this expansion renders the region unstable and triggers NAHR
314 resulting in the 22q11 deletions or duplications. To counter the paradox that LCR22 expansions reduce
315 overall fitness, we hypothesize an important role for the region in human evolution and adaptation,
316 previously described as the ‘core duplicon hypothesis’ [45–47]. Further research will be needed to

317 unravel the functional importance of LCR22 expansion, including the role of paralog-specific
318 transcripts.

319 **Materials and Methods**

320 **Sample collection and cell culture**

321 Four chimpanzee samples (*Pan troglodytes* 7, 8, 15, and 17), one gorilla cell line (*Gorilla gorilla* 1), and
322 five orangutans (*Pongo pygmaeus* 6, 7, 8, 9, and 10) were kindly provided by Professor Mariano Rocchi
323 (University of Bari, Italy). All these samples were Epstein-Barr virus (EBV) transfected cell lines and
324 cultured according to standard protocols. One chimpanzee fibroblast cell line was purchased from the
325 Coriell Cell Repository (AG 06939A). One gorilla fibroblast cell line (Gorilla Kaisi) was originally obtained
326 from the Antwerp Zoo (Antwerp, Belgium). The orangutan fibroblast cell line and the rhesus macaque
327 kidney cell line were obtained from the European Collection of Authenticated Cell Cultures (ECACC)
328 Repository. One EBV cell line was established from bonobo Banya from the Planckendael Zoo
329 (Mechelen, Belgium). More information on the samples is provided in S1 Table.

330 **Fiber-FISH**

331 Long DNA fibers were extracted from the cultured cell lines using the FiberPrep® DNA extraction kit
332 (Genomic Vision). The slides were hybridized with the LCR22-specific customized probe set[21],
333 supplemented with BAC probes targeting the unique regions between the LCR22s (S2 Table). Probes
334 were labeled with digoxigenin-dUTP (Jena Bioscience), fluorescein-dUTP (Jena Bioscience), biotin-
335 dUTP (Jena Bioscience), or combinations of these, using the BioPrime DNA Labeling System (Thermo
336 Fisher Scientific). Indirect labeling with Alexa Fluor 647 IgG Fraction Monoclonal Mouse Anti-
337 Digoxigenin (pseudocolored blue, Jackson Immunoresearch), Cy3 IgG Fraction Monoclonal Mouse
338 Anti-Fluorescein (pseudocolored green, Jackson Immunoresearch), and BV480 Streptavidin
339 (pseudocolored red, BD Biosciences) detected the primary labeled probes. The slides were scanned by
340 an automated fluorescence microscope (Genomic Vision) at three excitation levels, corresponding to

341 the three fluorophores. Images were automatically compiled by the system. The slides were visualized
342 in FiberStudio (Genomic Vision) and manually inspected for regions of interest. Based on matching
343 colors and distances between the probes, alleles were *de novo* assembled.

344 **Bionano optical mapping**

345 High-molecular weight DNA from one chimpanzee (*Pan troglodytes* 15), one bonobo (Bonobo Banya),
346 one gorilla (*Gorilla gorilla* 1), one orangutan (*Pongo pygmaeus* 8), and the rhesus macaque was
347 extracted using the SP Blood & Cell Culture DNA Isolation kit (Bionano Genomics) and labeled using
348 the DLS DNA labeling kit (DLE-1 labeling enzyme, Bionano Genomics). Samples were loaded onto
349 Saphyr Chips G2.3 (Bionano Genomics), linearized, and visualized using the Saphyr Instrument
350 (Bionano Genomics), according to the Saphyr System User Guide.

351 All analyses were performed in Bionano Access (Bionano Genomics). After general quality assessment
352 via the Molecule Quality Report, a *de novo* assembly was performed against the most recent human
353 reference genome hg38. Structural variants could be detected at the genome-wide level in the
354 generated circos plot. The 22q11.2 region was visually inspected for structural rearrangements by
355 zooming in to this region and comparing the compiled haplotypes with the hg38 reference.

356 **Transcriptome analysis of LCR22 genes**

357 We selected two transcriptome studies, one across eight brain regions (PRJNA393104) [25], and one
358 of neuronal differentiated induced pluripotent stem cells (PRJNA415990) [26,27]. The publicly
359 available transcriptome datasets were downloaded from the European Nucleotide Archive: 32
360 datasets were from PRJNA393104 and 22 from PRJNA415990. A full list of accession numbers can be
361 found in S4 Table. For each study, samples of the same individual were pooled together to generate a
362 higher overall coverage of each transcript. The pooled FASTQ files were aligned to the human reference
363 genome with BWA-MEM [48] and converted to BAM files with SAMtools [49]. The *de novo* assemblies
364 were executed with NOVOloci, a targeted assembler under development that was modified for

365 transcriptome data. NOVOloca needs a seed to initiate the assembly, therefore we prepared a short
366 seed (100-250bp) for each of the 39 genes of the LCR22s. The resulting assemblies were aligned to the
367 NCBI database with BLAST. Paralogs were identified and assembled separately with the heteroplasmy
368 module [28] of NOVOPlasty [50]. We manually inspected the transcriptome alignments to the LCR22s
369 and observed a large fraction of reads within introns, which also manifests in the *de novo* assemblies
370 as additional intronic sequences at the end of some transcripts. As we did not observe genomic
371 contamination, the presence of intronic sequences most likely originates from nascent RNA [51]. These
372 nascent RNA sequences were removed from the *de novo* assemblies based on coverage difference and
373 visual alignment.

374 **References**

- 375 1. Numanagić I, Gökkaya AS, Zhang L, Berger B, Alkan C, Hach F. Fast characterization of
376 segmental duplications in genome assemblies. *Bioinformatics*. 2018;34(17):i706–14.
- 377 2. Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE. Segmental Duplications: Organization and
378 Impact Within the Current Human Genome Project Assembly. *Genome Res*. 2001;11(6):1005–
379 17.
- 380 3. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte R V., Schwartz S, et al. Recent Segmental
381 Duplications in the Human Genome. *Science* (80-). 2002;297(5583):1003–7.
- 382 4. Chaisson MJP, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, et al. Resolving
383 the complexity of the human genome using single-molecule sequencing. *Nature*.
384 2015;517(7536):608–11.
- 385 5. Vollger MR, Dishuck PC, Sorensen M, Welch AME, Dang V, Dougherty ML, et al. Long-read
386 sequence and assembly of segmental duplications. *Nat Methods*. 2019;16(1):88–94.
- 387 6. Genovese G, Handsaker RE, Li H, Altemose N, Lindgren AM, Chambert K, et al. Using
388 population admixture to help complete maps of the human genome. *Nat Genet*.

- 389 2013;45(4):406–14.
- 390 7. Levy-Sakin M, Pastor S, Mostovoy Y, Li L, Leung AKY, McCaffrey J, et al. Genome maps across
391 26 human populations reveal population-specific patterns of structural variation. *Nat*
392 *Commun.* 2019;10(1):1–14.
- 393 8. Dennis MY, Eichler EE. Human adaptation and evolution by segmental duplication. *Curr Opin*
394 *Genet Dev.* 2016;41:44–52.
- 395 9. Dennis MY, Harshman L, Nelson BJ, Penn O, Cantsilieris S, Huddleston J, et al. The evolution
396 and population diversity of human-specific segmental duplications. *Nat Ecol Evol.* 2017;1:1–
397 23.
- 398 10. Bailey JA, Eichler EE. Primate segmental duplications: Crucibles of evolution, diversity and
399 disease. *Nat Rev Genet.* 2006;7(7):552–64.
- 400 11. Marques-Bonet T, Kidd JM, Ventura M, Graves TA, Cheng Z, Hillier LW, et al. A burst of
401 segmental duplications in the genome of the African great ape ancestor. *Nature.*
402 2009;457(7231):877–81.
- 403 12. Mikkelsen TS, Hillier LW, Eichler EE, Zody MC, Jaffe DB, Yang SP, et al. Initial sequence of the
404 chimpanzee genome and comparison with the human genome. *Nature.* 2005;437(7055):69–
405 87.
- 406 13. Gordon D, Huddleston J, Chaisson MJP, Hill CM, Kronenberg ZN, Munson KM, et al. Long-read
407 sequence assembly of the gorilla genome. *Science (80-).* 2016;352(6281).
- 408 14. Puech A, Saint-Joke B, Funke B, Gilbert DJ, Sirotkin H, Copeland NG, et al. Comparative
409 mapping of the human 22q11 chromosomal region and the orthologous region in mice reveals
410 complex changes in gene organization. *Proc Natl Acad Sci U S A.* 1997;94(26):14608–13.
- 411 15. Shaikh TH, Kurahashi H, Saitta SC, Mizrahy O’Hare A, Hu P, Roe BA, et al. Chromosome 22-

- 412 specific low copy repeats and the 22q11.2 deletion syndrome: genomic organization and
413 deletion endpoint analysis. *Hum Mol Genet.* 2000;9(4):489–501.
- 414 16. Bailey JA, Yavor AM, Viggiano L, Misceo D, Horvath JE, Archidiacono N, et al. Human-Specific
415 Duplication and Mosaic Transcripts: The Recent Paralogous Structure of Chromosome 22. *Am*
416 *J Hum Genet.* 2002;70(1):83–100.
- 417 17. Babcock M, Yatsenko S, Hopkins J, Brenton M, Cao Q, De Jong P, et al. Hominoid lineage
418 specific amplification of low-copy repeats on 22q11.2 (LCR22s) associated with velo-cardio-
419 facial/digeorge syndrome. *Hum Mol Genet.* 2007;16(21):2560–71.
- 420 18. Guo X, Freyer L, Morrow B, Zheng D. Characterization of the past and current duplication
421 activities in the human 22q11.2 region. *BMC Genomics.* 2011;12(1):71.
- 422 19. Gu W, Zhang F, Lupski JR. Mechanisms for human genomic rearrangements. *Pathogenetics.*
423 2008;1(1):4.
- 424 20. McDonald-McGinn D, Sullivan K, Marino B, Philip N, Swillen A, Vorstman J, et al. 22q11.2
425 Deletion Syndrome. *Nat Rev Dis Prim.* 2015;1(15071).
- 426 21. Demaerel W, Mostovoy Y, Yilmaz F, Vervoort L, Pastor S, Hestand MS, et al. The 22q11 low
427 copy repeats are characterized by unprecedented size and structural variability. *Genome Res.*
428 2019;29:1389–401.
- 429 22. Pastor S, Tran O, Jin A, Carrado D, Silva BA, Uppuluri L, et al. Optical mapping of the
430 22q11.2DS region reveals complex repeat structures and preferred locations for non-allelic
431 homologous recombination (NAHR). *Sci Rep.* 2020;10(1):1–13.
- 432 23. Chan S, Lam E, Saghbini M, Bocklandt S, Hastie A, Cao H, et al. Structural variation detection
433 and analysis using bionano optical mapping. *Methods Mol Biol.* 2018;1833:193–203.
- 434 24. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue

- 435 Expression (GTEx) project. *Nat Genet.* 2013;45(6):580–5.
- 436 25. Xu C, Li Q, Efimova O, He L, Tatsumoto S, Stepanova V, et al. Human-specific features of
437 spatial gene expression and regulation in eight brain regions. *Genome Res.* 2018;28(8):1097–
438 110.
- 439 26. Field AR, Jacobs FMJ, Fiddes IT, Phillips APR, Reyes-Ortiz AM, LaMontagne E, et al. Structurally
440 Conserved Primate LncRNAs Are Transiently Expressed during Human Cortical Differentiation
441 and Influence Cell-Type-Specific Genes. *Stem Cell Reports.* 2019;12(2):245–57.
- 442 27. Fiddes IT, Lodewijk GA, Mooring M, Bosworth CM, Ewing AD, Mantalas GL, et al. Human-
443 Specific NOTCH2NL Genes Affect Notch Signaling and Cortical Neurogenesis. *Cell.*
444 2018;173(6):1356-1369.e22.
- 445 28. Dierckxsens N, Mardulyn P, Smits G. Unraveling heteroplasmy patterns with NOVOPlasty. *NAR*
446 *Genomics Bioinforma.* 2020;2(1):1–10.
- 447 29. Guo X, Delio M, Haque N, Castellanos R, Hestand MS, Vermeesch JR, et al. Variant discovery
448 and breakpoint region prediction for studying the human 22q11.2 deletion using BAC clone
449 and whole genome sequencing analysis. *Hum Mol Genet.* 2015;25(17):3754–67.
- 450 30. Sudmant PH, Huddleston J, Catacchio CR, Malig M, Hillier LW, Baker C, et al. Evolution and
451 diversity of copy number variation in the great ape lineage. *Genome Res.* 2013;23:1373–82.
- 452 31. Babcock M, Pavlicek A, Spiteri E, Kashork CD, Ioshikhes I, Shaffer LG, et al. Shuffling of Genes
453 Within Low-Copy Repeats on 22q11 (LCR22) by Alu-Mediated Recombination Events During
454 Evolution. *Genome Res.* 2003;13:2519–32.
- 455 32. Bailey JA, Liu G, Eichler EE. An Alu Transposition Model for the Origin and Expansion of Human
456 Segmental Duplications. *Am J Hum Genet.* 2003;73(4):823–34.
- 457 33. Catacchio CR, Angela F, Maggiolini M, Addabbo PD, Bitonto M, Capozzi O, et al. Inversion

- 458 variants in human and primate genomes. *Genome Res.* 2018;28(6):1–11.
- 459 34. Boettger LM, Handsaker RE, Zody MC, Mccarroll SA. Structural haplotypes and recent
460 evolution of the human 17q21.31 region. *Nat Genet.* 2012;44(8):881–5.
- 461 35. Antonacci F, Dennis MY, Huddleston J, Sudmant PH, Steinberg KM, Rosenfeld JA, et al.
462 Palindromic GOLGA8 core duplicons promote chromosome 15q13.3 microdeletion and
463 evolutionary instability. *Nat Genet.* 2014;46(12):1293–302.
- 464 36. Nuttle X, Giannuzzi G, Duyzend MH, Schraiber JG, Sudmant PH, Penn O, et al. Emergence of a
465 *Homo sapiens*-specific gene family and chromosome 16p11.2 CNV susceptibility. *Nature.*
466 2016;536(7615):205–9.
- 467 37. Dougherty ML, Nuttle X, Penn O, Nelson BJ, Huddleston J, Baker C, et al. The birth of a human-
468 specific neural gene by incomplete duplication and gene fusion. *Genome Biol.* 2017;18(1):1–
469 16.
- 470 38. McCartney AM, Hyland EM, Cormican P, Moran RJ, Webb AE, Lee KD, et al. Gene Fusions
471 Derived by Transcriptional Readthrough are Driven by Segmental Duplication in Human.
472 *Genome Biol Evol.* 2019;11(9):2676–90.
- 473 39. Dennis MY, Nuttle X, Sudmant PH, Antonacci F, Graves TA, Nefedov M, et al. Evolution of
474 human-specific neural SRGAP2 genes by incomplete segmental duplication. *Cell.*
475 2012;149(4):912–22.
- 476 40. Charrier C, Joshi K, Coutinho-Budd J, Kim JE, Lambert N, De Marchena J, et al. Inhibition of
477 SRGAP2 function by its human-specific paralogs induces neoteny during spine maturation.
478 *Cell.* 2012;149(4):923–35.
- 479 41. Florio M, Albert M, Taverna E, Namba T, Brandl H, Lewitus E, et al. Human-specific gene
480 ARHGAP11B promotes basal progenitor amplification and neocortex expansion. *Science* (80-).
481 2015;347(6229):1465–70.

- 482 42. Mittelstaedt T, Schoch S. Structure and evolution of RIM-BP genes: Identification of a novel
483 family member. *Gene*. 2007;403(1–2):70–9.
- 484 43. Delihans N. A family of long intergenic non-coding RNA genes in human chromosomal region
485 22q11.2 carry a DNA translocation breakpoint/AT-rich sequence. *PLoS One*. 2018;13(4):1–19.
- 486 44. Delihans N. Formation of human long intergenic noncoding RNA genes, pseudogenes, and
487 protein genes: Ancestral sequences are key players. *PLoS One*. 2020;15(3):1–19.
- 488 45. Jiang Z, Tang H, Ventura M, Cardone MF, Marques-Bonet T, She X, et al. Ancestral
489 reconstruction of segmental duplications reveals punctuated cores of human genome
490 evolution. *Nat Genet*. 2007;39(11):1361–8.
- 491 46. Johnson ME, Cheng Z, Morrison VA, Scherer S, Ventura M, Gibbs RA, et al. Recurrent
492 duplication-driven transposition of DNA during hominoid evolution. *Proc Natl Acad Sci U S A*.
493 2006;103(47):17626–31.
- 494 47. Marques-Bonet T, Girirajan S, Eichler EE. The origins and impact of primate segmental
495 duplications. *Trends Genet*. 2009;25(10):443–54.
- 496 48. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*.
497 2013;1303.3997v:1–3.
- 498 49. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence
499 Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078–9.
- 500 50. Dierckxsens N, Mardulyn P, Smits G. NOVOPlasty: De novo assembly of organelle genomes
501 from whole genome data. *Nucleic Acids Res*. 2017;45(4):e18.
- 502 51. Kapranov P, St Laurent G, Raz T, Ozsolak F, Reynolds CP, Sorensen PHB, et al. The majority of
503 total nuclear-encoded non-ribosomal RNA in a human cell is “dark matter” un-annotated RNA.
504 *BMC Biol*. 2010;8(1):149.

505 **Figure captions**

506 **Fig 1. Composition of the 22q11.2 locus in human and non-human primates.** Schematic
507 representations of the 22q11.2 region, including LCR22-A through -H, based on Bionano optical
508 mapping and fiber-FISH. As represented in (A) the human reference genome hg38, (B) human, (C)
509 chimpanzee and bonobo, (D) gorilla, and (E) orangutan. (F) Bionano optical mapping results of
510 orangutan compared to the human reference genome. The top bar represents the human hg38
511 reference genome with blocks indicating the LCR22s (corresponding to Figure 1A). The bottom bar
512 represents the assembled haplotype for this orangutan. Grey lines between the maps indicate
513 orthologous loci. Blue labels in the maps are aligned labels, and yellow labels unaligned. Arrows below
514 depict rearrangements between the human and the orangutan genomes. (G) Schematic 22q11.2
515 representation of the macaque, only based on fiber-FISH results. Colored lines indicate the BAC probes
516 used in the fiber-FISH experiments (S2 Table). Different sizes and colors of the LCR22 blocks indicate
517 LCR22 differences in size and composition, respectively, based on fiber-FISH results. Cartoons are not
518 to scale.

519 **Fig 2. Human duplication structure and evolutionary analysis of LCR22-A.** (A) *De novo* assembly of a
520 LCR22-A haplotype based on matching colors and distances between the probes. SD22 duplicons are
521 assigned to specific probe combinations. (B) UCSC Genome Browser hg38 reference screenshot, with
522 tracks for fiber-FISH probe BLAT positions, segmental duplications, gaps, and RefSeq genes. Assigned
523 duplicons in (A) are decomposed to their corresponding fiber-FISH probes in this reference screenshot.
524 (C) Evolutionary tree representation of the observed LCR22-A haplotypes. Only a subset of assembled
525 haplotypes are depicted for human, to emphasize the human hypervariability. Filled, colored arrows
526 represent copies of duplicons, and hatched arrows represent partial copies of duplicons of the same
527 color.

528 **Fig 3. Human duplicon structure and evolutionary analysis of LCR22-B, -C, and -D.** (A) *De novo*
529 assembly of a LCR22-B (left) and LCR22-C (right) haplotype. SD22 duplicons are assigned to specific

530 probe combinations, based on the probe composition in LCR22-A (Figure 2A). (B) UCSC Genome
531 Browser hg38 reference screenshot of LCR22-B (left) and LCR22-C (right), with tracks for fiber-FISH
532 probe BLAT positions, segmental duplications, and RefSeq genes. (C) *De novo* assembly of an LCR22-D
533 haplotype based on matching colors and distances between the probes. SD22 duplicons are assigned
534 to specific probe combinations. (D) UCSC Genome Browser hg38 reference screenshot, with tracks for
535 fiber-FISH probe BLAT positions, segmental duplications, and RefSeq genes. The extended SD22-3
536 duplicon is decomposed to the corresponding fiber-FISH probes in the reference genome. (E)
537 Evolutionary tree representation of the observed LCR22-B, -C, and -D haplotypes. Filled, colored
538 arrows represent copies of duplicons, and hatched arrows represent partial copies of duplicons of the
539 same color.

540 **Supporting information captions**

541 **S1 Fig. Bionano optical mapping of the 22q11.2 region in chimpanzee, bonobo, and gorilla.** Regional
542 organization of the 22q11.2 locus in (A) chimpanzee, (B) bonobo, and (C) gorilla. *De novo* assembled
543 non-human primate maps are compared to the human reference genome (hg38). The top bar
544 represents the human hg38 reference genome with blocks indicating the LCR22s. The bottom bar
545 represents the assembled non-human primate haplotype. Grey lines between the maps indicate
546 orthologous signals between them. Blue labels in the maps are aligned labels, and yellow labels
547 unaligned.

548 **S2 Fig. Exact probe composition of the LCR22 chimpanzee haplotypes.** To derive the exact probe
549 composition of the chimpanzee haplotype, color-identical probes were differently labeled and
550 hybridized to the slides. Changes of the pattern indicate the presence of the differently labeled probe.
551 Red, cyan, and yellow probes were checked.

552 **S3 Fig. Exact probe composition of the LCR22 gorilla haplotypes.** To derive the exact probe
553 composition of the gorilla haplotype, color-identical probes were differently labeled and hybridized to

554 the slides. Changes of the pattern indicates that the presence of the differently labeled probe. Red,
555 cyan, blue, and yellow probes were checked.

556 **S4 Fig. Exact probe composition of the LCR22-A and –B orangutan haplotypes.** To derive the exact
557 probe composition of the orangutan haplotype, color-identical probes were differently labeled and
558 hybridized to the slides. Changes of the pattern indicate the presence of the differently labeled probe.
559 Red, cyan, blue, magenta, green, and yellow probes were checked. LCR22-C and –D were not included
560 in the analysis, since they only consist of one and two probes, respectively. The probes are linked to
561 unique BAC probes, predicting their composition.

562 **S5 Fig. Exact probe composition of the LCR22-A and –B rhesus macaque haplotypes.** To derive the
563 exact probe composition of the rhesus macaque haplotype, color-identical probes were differently
564 labeled and hybridized to the slides. Changes of the pattern indicates the presence of the differently
565 labeled probe. Red, cyan, and yellow probes were checked. LCR22-C and –D are not included in the
566 analysis, since they only consist of one and two probes, respectively. The probes are linked to unique
567 BAC probes, predicting their composition.

568 **S6 Fig. Chimpanzee LCR22-A and –D haplotypes in investigated samples.** De novo assembled
569 haplotypes for LCR22-A and LCR22-D in the six investigated samples. Two chimpanzees (*Pan*
570 *troglydytes* 7 and 15) showed structural variation distal in the LCR22-D haplotype. A white line
571 distinguishes the two haplotypes.

572 **S1 Appendix. Paralog analysis by heteroplasmy mode of NOVOPlasty.**

573 **S1 Table. Overview of non-human primate samples.**

574 **S2 Table. BAC probes targeting unique regions surrounding the LCR22s.**

575 **S3 Table. LCR22-A structural variation in the orangutan samples.**

576 **S4 Table. LCR22 specific transcript expression in human and non-human primates.**

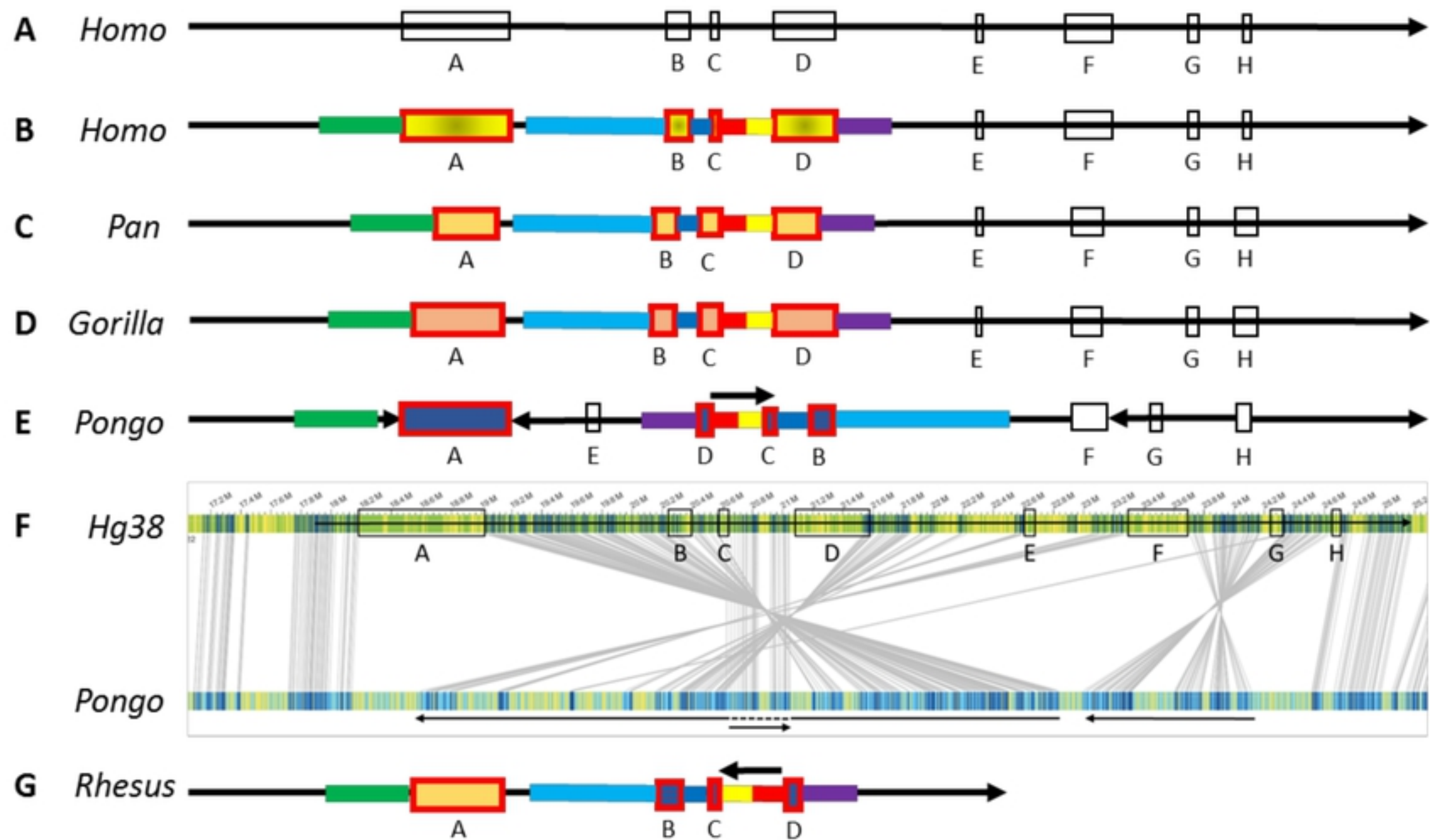


Figure 1

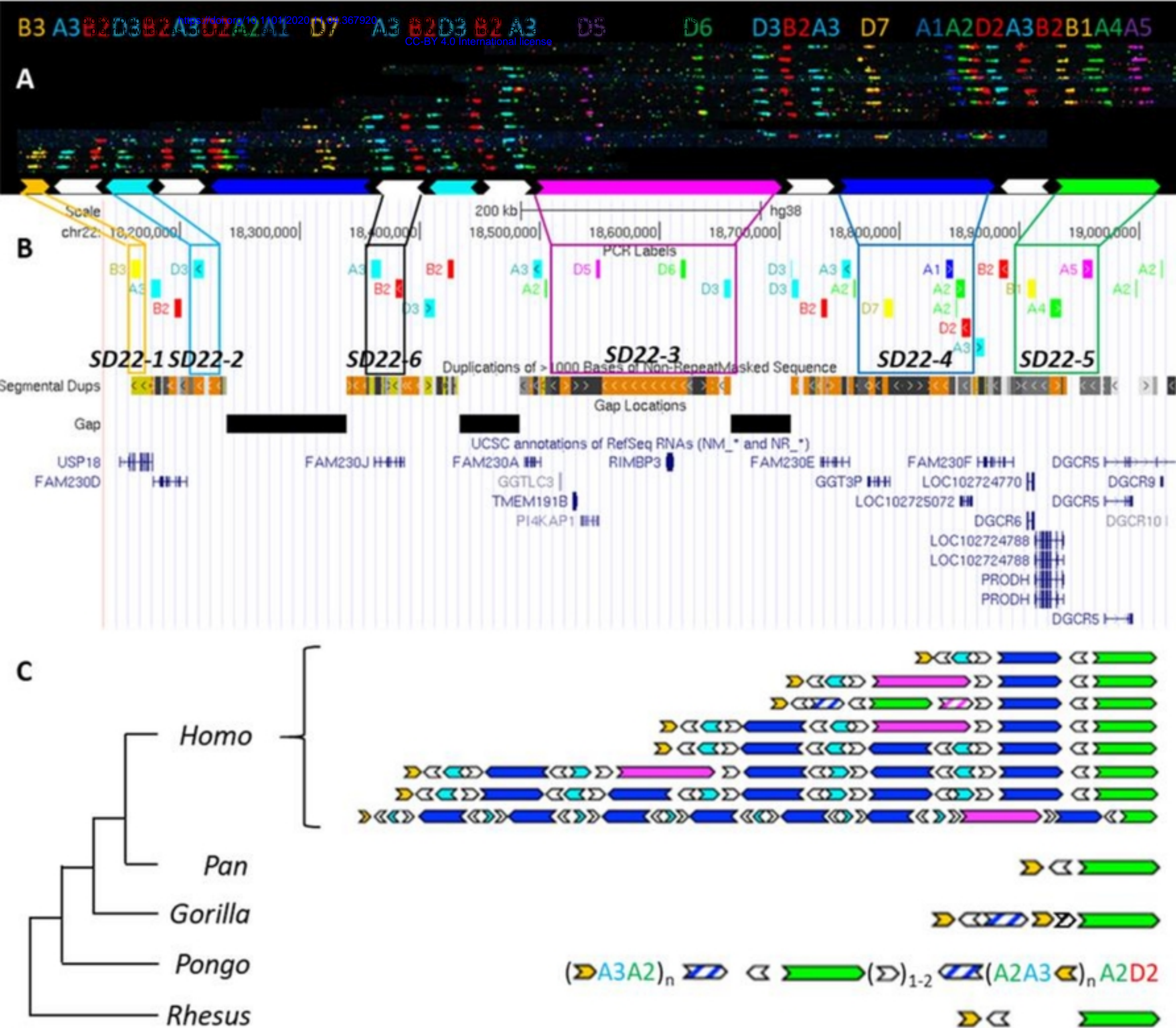


Figure 2

