Interchromosomal Colocalization with Parental Genes Is Linked to the Function and Evolution of Mammalian Retrocopies

Yubin Yan, Yuhan Tian, Zefeng Wu, Kunling Zhang, and Ruolin Yang*

College of Life Sciences, Northwest A&F University, Yangling, Shaanxi, China

*Corresponding author: E-mail: desert.ruolin@gmail.com.

Abstract

Retrocopies are gene duplicates arising from reverse transcription of mature mRNA transcripts and their insertion back into the genome. While long being regarded as processed pseudogenes, more and more functional retrocopies have been discovered. How the stripped-down retrocopies recover expression capability and become functional paralogs continually intrigues evolutionary biologists. Here, we investigated the function and evolution of retrocopies in the context of three-dimensional (3D) genome organization. By mapping retrocopy-parent pairs onto the chromatin contact maps of human and mouse cell lines, we found that retrocopies and their parental genes show a higher-than-expected interchromosomal colocalization frequency. The spatial interactions between retrocopies and parental genes occur frequently at loci in active subcompartments and near nuclear speckles. Accordingly, colocalized retrocopies are more actively transcribed and translated, and are more evolutionarily conserved than noncolocalized ones. The active transcription of colocalized retrocopies may result from their permissive epigenetic environment and shared regulatory elements with parental genes. Population genetic analysis on retroposed gene copy number variants (retroCNVs) in human populations revealed that retrocopy insertions are not entirely

random in regard to interchromosomal interactions and that colocalized retroCNVs are more likely to reach high frequency, suggesting that both insertion bias and natural selection contribute to the colocalization of retrocopy-parent pairs. Further dissection implies that reduced selection efficacy, rather than positive selection, contributes to the elevated allele frequency of colocalized retroCNVs. Overall, our results hint a role of interchromosomal colocalization in the "resurrection" of initially neutral retrocopies.

Key words: retrocopies, processed pseudogenes, 3D genome organization, interchromosomal interactions

Introduction

Gene duplication is a major process leading to the birth of new genes, which contributes significantly to the adaptive evolution and phenotypic novelty of species in diverse clades (Ohno 1970; Zhang 2003; Chen et al. 2013). Gene retrotransposition is an RNA-based gene duplication mechanism whereby mRNA derived from protein-coding genes are reverse transcribed to DNA and inserted back into the genome (Kaessmann et al. 2009). In mammals, gene retroduplication is mainly mediated by the non-LTR retrotransposon LINE-1s (L1s), and occasionally by LTR-retrotransposons (Esnault et al. 2000; Tan et al. 2016).

Although RNA-based gene duplication is less common and widespread than DNA-based duplication (i.e., segmental duplication and whole genome duplication) (Kaessmann 2010), thousands of retrotransposed gene copies (retrocopies) have been identified in humans and other mammalian species (Navarro and Galante 2015; Carelli et al. 2016; Rosikiewicz et al. 2017). However, to what extent this large amount of retrocopies contribute to new gene origination and genome evolution is still an open question. On the one hand, retrocopies have been stripped of introns and promoters during retrotransposition, rendering them to be "dead

50 on arrival", and were frequently dismissed as "processed pseudogenes" (Mighell et al. 2000;
51 Zhang et al. 2003). On the other hand, increasing number of protein-coding retrocopies
52 (retrogenes) with explicit functions have been reported in various organisms (Casola and
53 Betran 2017). For example, a recent study reported that a mammalian retrogene *HAPSTR2*
54 encodes a protein that retains the biochemical features of its parental gene *HAPSTR1* and acts
55 as a safeguard in stress signaling and resilience (Amici et al. 2023). Moreover, accumulating
56 evidence indicates that many retrocopies are robustly expressed, often in a tissue-specific
57 manner (Carelli et al. 2016; Qian et al. 2022). These retrocopy-derived RNAs may function as
58 natural antisense transcripts (NATs), long noncoding RNAs (lncRNAs), or sponging
59 microRNAs to regulate their progenitors or host genes (Kubiak and Makałowska 2017;
60 Ciomborowska-Basheer et al. 2021). Ribosome profiling data further suggest that a sizable
61 proportion of retrocopies may be actively translated (Kim et al. 2014; Ji et al. 2015; Qian et al.
62 2022), although their exact proteins and functions awaiting additional investigations.
63 How some retrocopies gain their expression capability and "awake" as functional paralogs
64 continually attracts evolutionary biologists (Vinckenbosch et al. 2006). Previous research
65 discovered that retrocopies may exploit preexisting promoters, evolve new promoters from
66 scratch, or recruit proto-promoters from their genomic vicinity (Carelli et al. 2016). For
67 example, the promoter of the aforementioned retrogene *HAPSTR2* originated from a nearby
68 proto-promoter (Amici et al. 2023). However, current studies mostly focused on the role of
69 local genomic context in the expression of retrocopies (Carelli et al. 2016; Machado and
70 Antunes 2020), while the influences of high-order chromatin organization on retrocopy
71 expression and functionality are largely unexplored.
72 In eukaryotes, high-order chromatin organization and transcription are tightly linked, with
73 reciprocal influence on each other (Bonev and Cavalli 2016; van Steensel and Furlong 2019).
74 Apart from the well-studied intrachromosomal (*cis*) structures, such as A/B compartments,

75 topologically associating domains (TADs), and chromatin loops, interchromosomal (*trans*)
76 interactions also play a critical role in gene regulation (Quinodoz et al. 2018; Maass et al. 2019)
77 and genome evolution (Dai et al. 2014; Xie et al. 2016). For example, human interchromosomal
78 paralogous gene pairs generated by whole genome duplication (i.e., ohnolog pairs) tend to be
79 colocalized in the nucleus, which may associate with stronger dosage balance in those ohnologs
80 (Xie et al. 2016). Likewise, neighboring genes in yeast exhibit interchromosomal spatial
81 proximity after their separation (Dai et al. 2014), which indicates a role for interchromosomal
82 interactions in the coregulation and evolution of functionally related genes. Inspired by these
83 previous studies and the fact that most retrocopies and their parental genes are located on
84 different chromosomes, we sought to ask whether retrocopy-parent pairs show
85 interchromosomal colocalization in the nucleus, and if so, how spatial proximity links to the
86 function and evolution of retrocopies.

87 Here, we used chromatin contact maps of human and mouse cell lines to detected significant
88 interchromosomal interactions between retrocopies and their parental genes. Through
89 extensive simulations, we demonstrated that retrocopy-parent pairs exhibit a higher
90 colocalization frequency than random chromatin fragment pairs under various null models.
91 Compared to noncolocalized retrocopies, colocalized retrocopies are enriched in active
92 subcompartments and speckles, and are more actively transcribed and translated, supporting a
93 connection between spatial proximity and functionality of retrocopies. Such differences might
94 be resulted from the higher number of shared regulatory elements with parental genes and the
95 more active epigenetic landscape of colocalized retrocopies. Furthermore, by investigating
96 retroposed gene copy number variants (retroCNVs) in human populations, we discovered that
97 insertion bias and nonadaptive evolution shaped the colocalization between retrocopies and
98 parental genes. Our results shed lights on the function and evolution of retrocopies in the
99 context of 3D genome organization.

100

101 # Results

102 **Human Interchromosomal Retrocopy-Parent Pairs Exhibit Widespread Nuclear**
103 **Colocalization**

104 To determine the extent of spatial colocalization between retrocopies and their parental genes,
105 we retrieved 5,159 retrocopy-parent pairs from Carelli et al. (2016). We removed retrocopy-
106 parent pairs that were not located on autosomes or the X chromosome, and those with parental
107 genes inferred from other species (Materials and Methods). Among the remaining 4,638 pairs,
108 the vast majority (4,427; 95.5%) are interchromosomal, i.e., individual retrocopies and their
109 parental genes are located on different chromosomes, which are the focus of this study. Using
110 Hi-C contact maps at 250-kb resolution of seven human cell lines (GM12878, HMEC, HUVEC,
111 IMR90, K562, KBM7, and NEHK) (Rao et al. 2014), we found that between 738 (16.7%; in
112 IMR90) and 1,019 (23.0%; in GM12878) retrocopy-parent pairs exhibit significant
113 interchromosomal interactions, i.e., spatial colocalization (**fig. 1A**, **supplementary fig. S1** and
114 **table S1**). More than half (2,511; 56.7%) retrocopy-parent pairs are colocalized in at least one
115 cell lines, and 1,465 (33.1%) pairs exhibit spatial proximity in two or more cell lines (**fig. 1B**).

116 We reasoned that if the spatial proximity between retrocopies and parental genes had
117 biological significance, then a higher-than-expected colocalization frequency should be
118 observed. Therefore, we performed various simulations to determine the statistical significance
119 of retrocopy-parent colocalization (**supplementary fig. S2**). First, we simulated 1,000 datasets,
120 in each of which 4,427 random fragment pairs with matching chromosomes and sequence
121 lengths as the retrocopy-parent pairs were sampled. Notably, we found that the colocalization
122 frequencies in simulated data sets are significantly lower than true retrocopy-parent pairs ($p <$
123 0.001 for all cell lines; **fig. 1C** and **supplementary fig. S3**), suggesting nonrandom
124 colocalization between retrocopies and parental genes. Second, we kept the information of

125 retrocopies unchanged while sampling random coding and noncoding genes with matching
126 chromosomes as parental genes. Third, we kept the information of parental genes unchanged
127 while sampling random fragments with matching chromosomes and sequence lengths as
128 retrocopies. Forth, similar to the second simulation, but we restricted the randomly sampled
129 genes to protein-coding genes. For these simulations, we also observed a significant lower
130 colocalization frequencies in simulated data sets compared to true retrocopy-parent pairs ($p <$
131 0.001 in all cases; **fig. 1C** and **supplementary fig. S3**). However, compared to the first null
132 model, the colocalization frequencies in these simulated data are higher, indicating that some
133 properties (e.g., local genomic and epigenomic environments) of retrocopies or parental genes
134 may contribute to interchromosomal colocalization.

135 Lastly, to further investigate the influence of retrocopies and parental genes on spatial
136 proximity, we randomly shuffled the relationship between retrocopies and cognate genes for
137 1,000 times, and determined the distribution of colocalization frequency. Particularly, the
138 observed colocalization frequencies in GM12878, HMEC, HUVEC, K562, and NHEK are still
139 significantly higher than expected, while the significance is marginal in IMR90 and KBM7
140 (**fig. 1C** and **supplementary fig. S3**). In sum, we demonstrated that human interchromosomal
141 retrocopy-parent pairs exhibit significant nuclear colocalization.

142

143 **Colocalized Retrocopies Are Enriched in Active Subcompartments and Are Close to**
144 **Nuclear Speckles**

145 We next asked where did retrocopies and parental genes colocalize with respect to the 3D
146 genome organization in the nucleus. Interphase chromosomes of eukaryotes are organized
147 hierarchically into chromosome territories, A/B compartments, TADs, and chromatin loops
148 (Bonev and Cavalli 2016). Using ultra-high resolution Hi-C contact map, Rao et al. (2014)
149 further divides the active A and inactive B compartments of GM12878 into five primary

150 subcompartments (A1, A2, B1, B2, and B3), which are associated with distinct genomic and
151 epigenomic features in a more refined manner. We found that in GM12878, retrocopies are
152 slightly enriched in A2 subcompartment while depleted in B2 and B3 subcompartments (**fig.**
153 **2A**). Such enrichment of active subcompartments and depletion of inactive subcompartments
154 is more apparent for parental genes. In addition, compared to protein-coding genes as a whole,
155 retrocopies' parental genes show stronger enrichment of active subcompartments and depletion
156 of inactive subcompartments, which is consistent with previous findings that highly expressed
157 genes are more prone to generate retrocopies (Zhang et al. 2003; Podlaha and Zhang 2009; Sisu
158 et al. 2014).

159 Remarkably, colocalized retrocopies show strong enrichment of A1 and A2
160 subcompartments and strong depletion of B2 and B3 subcompartments, resembling the pattern
161 of parental genes (**fig. 2A**). In contrast, noncolocalized retrocopies are depleted in active A1
162 subcompartment and show little depletion of inactive subcompartments. Dividing colocalized
163 and noncolocalized retrocopies into intergenic and intragenic subclasses do not affect the
164 general pattern (**fig. 2A**), indicating that the subcompartment association is not affected by host
165 genes. Besides, parental genes and colocalized retrocopies are slightly enriched in TADs.
166 Contrary to retrocopies, unprocessed pseudogenes are depleted in A1 and A2 subcompartments
167 and TAD while slightly enriched in B1 and B2 subcompartments (**fig. 2A**), reflecting different
168 3D chromatin environments between processed and unprocessed "pseudogenes". Using
169 subcompartment annotations inferred from relatively low resolution Hi-C maps by deep
170 learning (Xiong and Ma 2019), we obtained similar results in HMEC, HUVEC, IMR90, and
171 K562 (**supplementary fig. S4**).

172 Of note, for colocalized retrocopy-parent pairs, both ends tend to be overlapped with A1 or
173 A2 subcompartments, whereas for noncolocalized pairs, a substantial proportion of retrocopies

174 are overlapped with B2 or B3 subcompartments, despite the majority of parental genes are still
175 overlapped with active subcompartments (**fig. 2B** and **supplementary fig. S5**).
176 Subcompartments are strongly associated with subnuclear structures, such as nuclear lamina,
177 nucleolus, and speckles, which are associated with genome regulation (Chen et al. 2018). Thus,
178 we compared the spatial positioning relative to subnuclear structures of colocalized and
179 noncolocalized retrocopies using TSA-seq data in K562 (Chen et al. 2018). TSA-seq is a
180 technique that combines tyramide signal amplification (TSA) and high-throughput sequencing
181 to map chromosome organization relative to subnuclear structures genome-wide. We found
182 that the $\log_2$(pull-down/input) signals of SON and pSC35 TSA-Seq of colocalized retrocopies
183 or parental genes are significantly higher than their noncolocalized counterparts, whereas an
184 opposite trend is observed for LaminAC and LaminB TSA-Seq signals (**fig. 2C** and
185 **supplementary figs. S6A-C**). SON and pSC35 are two marker proteins for nuclear speckles
186 while LaminAC and LaminB are markers for nuclear lamina, and high signal means shorter
187 cytological distance to the respective subnuclear structure. Thus, our results indicate that
188 colocalized retrocopies and parental genes are closer to speckles, while their noncolocalized
189 counterparts are closer to lamina.
190 SON TSA-Seq signals can be further converted to actual distance (in μm) between chromatin
191 and nuclear speckles and loci can be assigned to different TSA deciles based on their distance
192 to speckles (Chen et al. 2018). Leveraging these data, we found that colocalized retrocopies
193 have an average distance of 0.48μm to speckles, longer than that of colocalized parental genes
194 but significantly shorter than that of noncolocalized retrocopies (**supplementary fig. S6D**).
195 Consistently, colocalized parental genes have the highest proportion of fragments falling in the
196 9th and 10th SON TSA deciles (chromatin that is mostly close to speckle), followed by
197 colocalized retrocopies and noncolocalized parental genes, while only a small proportion of
198 noncolocalized retrocopies fall in the 9th and 10th deciles (**fig. 2D**). Furthermore, by integrating

199 TSA-Seq, DamID and Hi-C data in the K562 cell line, Wang et al. (2021) partitioned the
200 genome into 10 different spatial compartmentalization states relative to nuclear speckles,
201 lamina, and nucleolus. We found that colocalized retrocopies are enriched in interior active
202 regions and speckle-associated chromatin, while depleted in interior repressive regions and
203 lamina-associated chromatin, similar to the pattern of parental genes (**supplementary fig. S7**).
204

205 **RNA Biogenesis and Shared Regulatory Elements Are Linked to the Interchromosomal**
206 **Colocalization between Retrocopies and Parental Genes**

207 Although retrocopies are often thought to be processed pseudogenes, most of them show
208 evidence of transcription (Carelli et al. 2016). The distinct positioning of colocalized and
209 noncolocalized retrocopies in the nucleus suggests they may have different transcription profile,
210 we then asked whether the spatial colocalization with parental genes is associated with the
211 expression of retrocopies. Using RNA-seq data of GM12878, HUVEC, K562, and NHEK from
212 the ENCODE project (Djebali et al. 2012), we have also detected evidence of transcription for
213 most retrocopies (**supplementary table S1**), with a substantial proportion of them showing
214 robust expression. For example, 772 ~ 1,350 (17.4 ~ 30.5%) retrocopies expressed with an
215 FPKM ≥ 1 in the studied cell lines for polyadenylated (ployA+) RNAs in the whole cell
216 (**supplementary fig. S8A**). The proportion of robustly expressed retrocopies for polyA+ RNAs
217 is larger than 15% and 20% in the nucleus and cytosol, respectively (**supplementary figs. S8B**
218 **and C**). There are also substantial non-polyadenylated (polyA-) RNAs transcribed from
219 retrocopies in the nucleus and whole cell, but relatively few in the cytosol (**supplementary**
220 **figs. S8D-F**). More importantly, colocalized retrocopies and parental genes have significantly
221 higher expression level than their noncolocalized counterparts in all cell lines, for RNAs in
222 different compartments (whole cell, nucleus, and cytosol), and for both ployA+ and polyA-
223 RNAs (**fig. 3A** and **supplementary figs. S9 and S10**). In addition, the GRO-Seq and PolII

224 TSA-Seq data in K562 (Chen et al. 2018) shows that colocalized retrocopies and parental genes
225 generate more nascent RNAs and are closer to RNA polymerase II than their noncolocalized
226 counterparts, suggesting that RNA biogenesis is more active at colocalized loci (**fig. 3B**).
227 Enhancer sharing has been hypothesized to account for the spatial colocalization and
228 coregulation of paralogous genes located on the same chromosome (Ibn-Salem et al. 2017).
229 Also, transcription factors (TFs) play a part in interchromosomal interactions (Kim and
230 Shendure 2019). We thus interrogated the relationship between shared regulatory elements and
231 spatial colocalization. As expected, we found that the flanking regions (putative transcription
232 start site, i.e., TSS ± 3 kb) of colocalized retrocopies and parental genes harbor more
233 transcription factor binding sites (TFBSs) than noncolocalized ones (**supplementary fig. S11**).
234 In particular, the number of shared TFBSs between colocalized retrocopies and parental genes
235 is significantly greater than that of noncolocalized retrocopy-parent pairs (**fig. 3C**). It should
236 be noted that the choice of flaking regions (e.g., TSS ± 1 kb or upstream 3 kb) does not affect
237 our results (**supplementary fig. S12**). Interestingly, the corresponding TFs of shared motifs
238 are mostly zinc finger proteins, such as ZNF263, ZNF384, ZNF460, and ZNF135, consistent
239 with a recent report in which zinc finger factors are found to play a role in speckle-associated
240 interchromosomal interactions (Joo et al. 2023).
241 Colocalized and noncolocalized retrocopies are enriched in different subcompartments,
242 which are associated with distinct epigenomic features and may affect the expression of
243 retrocopies. We then explicitly examined the signals of 12 epigenomic features (including
244 DNase I sensitivity, the histone variant H2A.Z, and 10 histone modifications) for retrocopies.
245 Generally, colocalized retrocopies and their flanking regions have stronger signals of activating
246 histone modifications (e.g., H3K4me1, H3K27ac, H3K36me3, and H3K79me2) in comparison
247 with noncolocalized retrocopies (**fig. 4** and **supplementary figs. S13-S17**), indicating an
248 association between epigenetic landscape and interchromosomal colocalization. Notably,

249 retrocopies, even those noncolocalized ones, harbor stronger activating histone modifications
250 than unprocessed pseudogenes (**fig. 4** and **supplementary figs. S13-S17**), further supporting
251 the delimitation between processed and unprocessed "pseudogenes" in the context of 3D
252 genome environment.

253 To evaluate the relative importance of different features in the colocalization between
254 retrocopies and their parental genes, we trained a logistic regression model using 35 genetic
255 and epigenetic features of retrocopies and parental genes as predictors (**supplementary table**
256 **S2**) in each cell line where matched data were available. For GM12878, the machine learning
257 model is able to predict retrocopy-parent colocalization with an accuracy of 0.798 and an area
258 under the receiver operating characteristic curve (AUC) of 0.785 (**supplementary fig. S18A**).
259 Among the features included, the number of shared motifs between retrocopies and parental
260 genes is the most important, with various histone modifications of retrocopies also contribute
261 significantly (**supplementary fig. S18B**). The number of shared motifs and epigenetic features
262 of retrocopies are also top predictors in models of HUVEC, K562, and NHEK, but the
263 performance is worse for these cell lines (**supplementary figs. S19-21**).

264 Taken together, we revealed that colocalized retrocopies are more actively transcribed than
265 noncolocalized ones, and shared regulatory elements and permissive epigenetic environments
266 may contribute to the spatial colocalization between retrocopies and parental genes.

267

268 **Colocalized Retrocopies Are More Likely to Be Functional and Evolutionarily Conserved**

269 Retrocopies may preserved the coding potentials of their parental genes before degenerative
270 mutations disrupting the open reading frames (ORFs), therefore, gaining expression is the first
271 and most important step of being functional for a retrocopy. The resulting RNAs may serve as
272 functional units *per se*, such as NATs, lncRNAs, or microRNAs (Kubiak and Makałowska
273 2017). There are also sporadic cases that retrocopies act as important protein-coding genes in

274 development and disease (Casola and Betran 2017). To further assess the functional relevance
275 of retrocopies at the translation level, we leveraged the ribosome profiling (i.e., ribo-seq) data
276 across 101 studies curated on RPFdb (Wang et al. 2019), and found that colocalized retrocopies
277 have significantly higher RPKM (Reads Per Kilobase per Million) values than noncolocalized
278 retrocopies in five of the seven cell lines (**fig. S22**). The difference in ribo-seq signals is
279 statistically marginal between retrocopies that are colocalized in zero and one cell line, while
280 retrocopies with recurrent colocalized in two or more cell lines are covered by significantly
281 more reads than the other two categories of retrocopies (**fig. 5A**). Because ribo-seq reads could
282 be derived from genomic fragments randomly co-purified with ribosomes, we further screened
283 retrocopy transcripts with actively translated ORFs. These ORFs were detected using RibORF
284 (Ji et al. 2015), which makes use of the 3-nt periodicity and uniformity of read distribution
285 across codons of translated regions, thus reduce spurious signals of translation. When only
286 transcripts with at least one detected ORFs were considered, colocalized retrocopies still
287 present higher ribo-seq signals than noncolocalized ones (**supplementary fig. S23**). These
288 results indicate colocalized retrocopies, especially those exhibiting recurrent colocalization,
289 are more prone to be actively translated.

290 If the higher levels in transcription and translation of colocalized retrocopies did confer any
291 functional significance, they would be more evolutionarily conserved than noncolocalized ones.
292 As expected, we observed that the average phastCons conservation scores of colocalized
293 retrocopies are significantly higher than noncolocalized ones (**fig 5B** and **supplementary fig.**
294 **S24**). This pattern is retained when we use LINSIGHT scores (Huang et al. 2017) to measure
295 the conservation of retrocopies (**supplementary fig. S25**). In addition, colocalized retrocopies
296 and their putative promoters (1 kb upstream of TSSs) are more likely to be overlapped with
297 NHGRI-EBI GWAS single nucleotide polymorphisms (SNPs) (Buniello et al. 2019) than
298 noncolocalized ones (Fisher exact test, $p < 0.01$ in six out of the seven cell lines; **fig. 5C**),

299 further supporting their higher potential of functionality. We found that the phastCons scores
300 of retrocopies are positively correlated with their levels of RNA-seq (R = 0.45, $p < 2.2e{-}16$)
301 and ribo-seq (R = 0.28, $p < 2.2e{-}16$; **fig. 5D**). However, after accounting the variation of RNA
302 expression level, the partial correlation coefficient between phastCons score and ribo-seq
303 signal is only 0.088 ($p = 5.5 \times 10^{-8}$), while the partial correlation coefficient between
304 phastCons score and RNA expression level is 0.374 ($p = 1.8 \times 10^{-12}$) when controlling for ribo-
305 seq signal, suggesting that retrocopies may be of more functional relevance at the RNA level.
306 Of particular note is that for two sister retrocopies derived from the same parental gene, the
307 colocalized one generally has higher levels of expression, motif sharedness, and conservation
308 than the noncolocalized one in GM12878 (**supplementary fig. S26**). But the trends are not
309 apparent in other cell lines (data not shown), possibly due to the small sample size of desired
310 retrocopies and the nosier nature of colocalization in cell lines other than GM12878.
311 In addition, to ask whether the colocalization pattern of retrocopy-parent pairs and its
312 functional significance is conserved across mammalian species, we performed similar analyses
313 with 4,532 mouse interchromosomal retrocopy-parent pairs (**supplementary table S3**) (Carelli
314 et al. 2016). Using Hi-C map of mouse lymphoblastoid cells (CH12-LX) at 250-kb resolution
315 (Rao et al. 2014), we found that 744 (16.4%) retrocopy-parent pairs are in spatial proximity,
316 significantly higher than the colocalization frequencies under various null models
317 (**supplementary fig. S27**). In CH12-LX, colocalized retrocopies and parental genes also
318 expressed at a higher level than noncolocalized ones (**supplementary fig. S28**). Comparison
319 of ribo-seq signals and phastCons scores indicate colocalized retrocopies are more actively
320 translated and tend to be more conserved than noncolocalized ones in mouse (**supplementary**
321 **fig. S29**). Besides, the conservation of retrocopies is positively correlated with their levels of
322 transcription and translation, with a greater correlation coefficient between conservation and
323 RNA expression (**supplementary fig. S30**). These results show that the spatial colocalization

324 between retrocopies and parental genes is an evolutionarily conserved feature of mammalian
325 genome organization.
326

327 **Biased Insertion and Natural Selection Shaped the Interchromosomal Colocalization**
328 **Between Retrocopies and Parental Genes**

329 Lastly, we explored what molecular and evolutionary mechanisms might have shaped the
330 spatial colocalization between retrocopies and parental genes. To this end, we identified 625
331 polymorphic retrocopies (retroCNVs) in five human populations (CEU, CHS, LWK, PEL, and
332 YRI; 491 individuals) using sideRETRO (Miller et al. 2021) (**supplementary table S4**;
333 Materials and Methods). As did in Zhang et al. (2021), we did not distinguish homozygotes
334 and heterozygotes, only accounted for the presence or absence of each allele in different
335 individuals. Our list captures most retroCNVs identified in other studies (Abyzov et al. 2013;
336 Ewing et al. 2013; Schrider et al. 2013; Zhang et al. 2017; Feng and Li 2021; Batcher et al.
337 2022) (**supplementary table S5**) and the neighbor-joining tree based on retroCNV presence
338 frequency is topologically the same as the tree built from genome-wide SNPs (**supplementary**
339 **fig. S31**), indicating the overall accuracy of retroCNV identification. The presence frequency
340 spectrum of retroCNVs is more skewed toward low-frequency in comparison with SNPs
341 (**supplementary fig. S32**), congruent with the highly deleterious effect of retroCNV insertions.
342 Ninety-eight retroCNVs are present in only one individual, which may be enriched in *de novo*
343 retrocopy insertions. We then performed the same simulations for these presumably *de novo*
344 retrocopies and their parental genes as did for putative "fixed" retrocopy-parent pairs
345 (**supplementary fig. S2**), and found that the colocalization frequency between *de novo*
346 retrocopy insertion sites and their parental genes is only significantly higher than expected in
347 some simulations (**fig. 6A** and **supplementary fig. S33**). This result suggests that the insertion

348 of retrocopies is not random, but nonrandom insertion cannot fully explain the current
349 colocalization pattern between retrocopies and parental genes.

350 Then we investigated the potential role of post-insertion selection in determining retrocopy-
351 parent colocalization. In contrast to *de novo* retrocopy insertions, high-frequency (occurs in ≥
352 28 individuals; n = 96, a sample size that is comparable to singletons) retroCNVs and their
353 parental genes exhibit a higher-than-expected colocalization frequency in all simulations (**fig.
354 6A** and **supplementary fig. S34**), indicating the action of post-insertion selection on retrocopy
355 survival in the context of interchromosomal interactions. Furthermore, we found that the
356 presence frequency spectrum of retroCNVs exhibiting colocalization in two or more cell lines
357 is less skewed toward low frequency compared to retroCNVs showing colocalization in zero
358 or one cell line (**fig. 6B**), suggesting relaxed negative selection on colocalized retroCNVs. The
359 difference in frequency spectrum holds true for retroCNV subclasses base on colocalization
360 status in individuals cell lines (**supplementary fig. S35**). With respect to the 3D genome
361 organization, recurrently colocalized retroCNVs are enriched in the active A1 subcompartment
362 (**fig. 2A** and **supplementary fig. S4**) and interior active regions and speckles (**supplementary
363 fig. S6**). These active chromatin regions may have higher density of selected sites and/or more
364 intense selection, which are negatively correlated with local effective population size
365 (Gossmann et al. 2011), leading to reduced selection efficacy on recurrently colocalized
366 retroCNVs. Indeed, the density of conserved sites (phastCons score > 0.5) flanking colocalized
367 retroCNVs (insertion site ± 50 kb) is significantly higher than noncolocalized retroCNVs
368 (**supplementary fig. S36**). Also, the unfolded site frequency spectrum of SNPs in the flanking
369 regions (insertion site ± 50 kb) of recurrently colocalized retroCNVs is more skewed toward
370 low frequency (**supplementary fig. S37**).

371 Because there is also a higher proportion of intermediate-to-high-frequency alleles for
372 recurrently colocalized retroCNVs, we sought to detect whether the elevated frequency of

colocalized retroCNVs is driven by positive selection. If a retroCNV insertion was under strong positive selection, selective sweep would: 1) reduce the nucleotide diversity (π) among sequences near the insertion for retroCNV carriers compared to noncarriers; 2) reduce the Tajima's D among sequences near the insertion for retroCNV carriers compared to genomic background; 3) increase the linkage disequilibrium (LD) between SNPs near the insertion for retroCNV carriers compared to genomic background (Zhang and Tautz 2022). Based on these principles, we scanned signals of positive selection for each retroCNV with allele frequency ≥ 0.5 in individual populations using SNPs from the 1000 Genomes Project (Altshuler et al. 2015) (Materials and Methods). We found that neither colocalized nor noncolocalized high-frequency retroCNVs are under positive selection using the above criteria (**supplementary table S6**). Therefore, our results indicate that the intermediate-to-high-frequency of both colocalized and noncolocalized retroCNVs is probably the result of nonadaptive process or weak positive selection that left untraceable signal.

Discussion

In the present study, we revealed that retrocopy-parent pairs exhibit higher frequency of interchromosomal colocalization than expected under various null models and that colocalized retrocopies are more active in terms of transcription and translation. We further explored the potential mechanisms, functional consequences, and evolutionary implications of the nonrandom colocalization between retrocopies and their parental genes.

We focused exclusively on interactions between retrocopies and parental genes that are located on different chromosomes, i.e., interchromosomal contacts. Compared to intrachromosomal structures (e.g., TADs and chromatin loops), the mechanisms and functional consequences of interchromosomal interactions are less well understood. Although often being thought as noise or technical artifacts, emerging evidence suggests that interchromosomal

interactions are relevant to fundamental cellular activity, such as transcription and splicing (Maass et al. 2019). Using CRISPR/Cas9 live-cell imaging, Maass et al. (2018) showed that interchromosomal contacts are as stable and frequent as intrachromosomal interactions. In another study, Quinodoz et al. (2018) developed split-pool recognition of interactions by tag extension (SPRITE) and detected two hubs of interchromosomal interactions that shape 3D genome organization in the nucleus. Along with accumulating evidences of *trans* gene regulation in health and in disease (Patel et al. 2014; Bashkirova and Lomvardas 2019; Maass et al. 2019; Lei et al. 2023; Moon et al. 2023), recent progress indicates that interchromosomal interactions play structural and functional roles in cells. Our finding that retrocopy function is linked to *trans* interactions with parental genes also hints nonrandom organization of interchromosomal contacts.

Interchromosomal contacts often occur at farther spatial distances than *cis* interactions (Maass et al. 2018), rendering their detection difficult for regular ligation-based methods (e.g., Hi-C). In this study, we leveraged the ultra-high-depth Hi-C data of GM12878, which has ~5 billion mapped read pairs, with 740 million interchromosomal read pairs (Rao et al. 2014). Even at such sequencing depth, we primarily depend our analysis of interchromosomal interaction at 250-kb resolution. For the other six human cell lines and the mouse CH12-LX cell line with shallower sequencing depth, we chose to select the top-5M contacts with lowest *p*-values, as a compromise between the number and reliability of significant interactions. We recognized that the resulting interactions in relatively low-depth cell lines may be noisier than in GM12878, but significant interactions should also be enriched in those top-5M contacts. To further evaluate the influence of Hi-C resolution and the number of significant interchromosomal interactions, we detected spatial colocalization between retrocopies and parental genes at 100-kb and 500kb resolutions in GM12878. Using the same criterion (q-value < 0.05 in FitHiC2), the number of colocalized pairs of retrocopy-parent pairs decreases with

423 the increase of resolution. However, the colocalization pairs based on high-resolution data is
424 almost a subset of pairs detected using low-resolution contact map (**supplementary fig. S38**).
425 Furthermore, although the observed colocalization frequencies changes under different
426 resolutions, they are always significantly greater than that under various null models
427 (**supplementary fig. S39**).

428 Subsequently, we established that colocalized and noncolocalized retrocopies exhibit distinct
429 properties in terms of subcompartment annotations and cytological distance to subnuclear
430 structures (**fig. 2 and supplementary figs. S4-7**). For subcompartments, both A1 and A2
431 harbor highly expressed genes and activating chromatin marks, and are depleted in nucleolus-
432 associated domains (NADs), but they show difference in replication timing and GC content.
433 B1 represents facultative heterochromatin that exhibits intermediate level of transcriptional
434 activity and high level of H3K27me3. B2 and B3 are associated with nuclear lamina, but B2 is
435 enriched in NADs whereas B3 is strongly depleted in NADs (Rao et al. 2014; Xiong and Ma
436 2019). Using TSA-seq data, Chen et al. (2018) further revealed that chromatin in A1
437 subcompartment is close to nuclear speckles while A2 subcompartment corresponds to region
438 with non-speckle associated transcription. Therefore, our observations (**fig. 2 and**
439 **supplementary figs. S4-7**) imply that many interchromosomal interactions between
440 retrocopies and their parental genes may be mediated by nuclear speckles. Nuclear speckles
441 are domains enriched in pre-mRNA splicing factors (Spector and Lamond 2011) and are also
442 hubs of interchromosomal interactions where active transcription takes place nearby
443 (Galganski et al. 2017; Quinodoz et al. 2018; Chen and Belmont 2019; Kim et al. 2020).
444 Remarkably, zinc finger proteins are found to facilitate the establishment of speckle-associated
445 interchromosomal contacts (Joo et al. 2023), in accordance with our findings that the number
446 of shared regulatory motifs is the most importance feature in predicting retrocopy-parent
447 colocalization and that most shared motifs are bind by TFs in the zinc finger family. Thus, our

448 findings further support a role of TFs in interchromosomal interactions (Kim et al. 2019; Kim
449 and Shendure 2019).

450 Speckle-associated interchromosomal interactions, along with more active chromatin marks
451 and higher number of shared motifs associated with colocalized retrocopies, signify colocalized
452 retrocopies should be more actively transcribed. The implications of higher expression of
453 colocalized retrocopies could be twofold. First, retrocopy-derived RNAs could function as
454 NATs, lncRNAs, or microRNAs in diverse cellular activities, including regulation of their
455 parental genes (Kubiak and Makałowska 2017). Besides, RNAs transcribed from retrocopies
456 may be processed and translated, and function at the protein level (Qian et al. 2022). Second,
457 the action of transcription and the resulting RNAs can orchestrate both intra- and inter-
458 chromosomal chromatin organization (van Steensel and Furlong 2019; Bertero 2021;
459 Quinodoz and Guttman 2021). In the nucleus, 3D genome organization and transcription are
460 intermingled, with reciprocal influence on each other. Recent single-cell Hi-C and RNA-seq
461 data discovered widespread chromatin interaction rewiring before transcription activation,
462 strongly supporting a regulatory role of chromatin organization on gene expression (Liu et al.
463 2023). Meanwhile, TFs can bring DNA fragments on the same or different chromosomes to
464 spatial proximity through direct oligomerization or cofactor oligomerization, leading to
465 transcription-mediated interactions in *cis* or *trans* (Kim et al. 2019; Kim and Shendure 2019).
466 In addition, non-coding RNAs in the nucleus may serve as chromosome scaffolds or seeds to
467 guide chromatin organization (Mele and Rinn 2016; Quinodoz et al. 2021; Bouwman et al.
468 2022). However, the actual impact of retrocopy transcription and retrocopy-derived RNAs on
469 interchromosomal interactions needs additional investigations.

470 We further showed that both biased insertion and natural selection contribute to the
471 nonrandom colocalization between retrocopies and parental genes (**fig. 6**). Mechanistically,
472 new L1 insertions are linked to DNA replication, with a preference in early-replicating regions

473 of the genome (Flasch et al. 2019; Sultana et al. 2019). Because early-replicating regions are
474 enriched with active subcompartments and speckles and are more like to be involved in
475 interchromosomal interactions (Vouzas and Gilbert 2021), if retrocopy insertions have the
476 same property as L1s, this insertion bias would partly explain the higher-than-expected
477 colocalization frequency between *de novo* retroCNVs and parental genes.

478 Our population genetic analysis indicates that colocalized retroCNVs are more likely to
479 reach high frequency, but this elevated frequency is probably driven by nonadaptive forces.
480 Then, what is the evolutionary significance of retrocopy-parent colocalization? Given the fact
481 that colocalized "fixed" retrocopies are more actively transcribed and have higher potential of
482 functionality than noncolocalized ones, we argue that the spatial proximity with parental genes
483 may facilitate the "resurrection" of initially neutral retrocopies. This is similar to the situation
484 where new genes frequently originated in the promiscuous testis (Kaessmann 2010).
485 Colocalized retrocopies are also residing in chromatin with activating epigenetic marks, are
486 cytologically close to speckles, and extensively share regulatory elements with their
487 progenitors, all of which may facilitate their exploitation of cellular transcription machinery
488 and the exaptation of these seemingly neutral retroposed gene copies.

489 However, we are aware of some potential caveats should be considered in interpreting the
490 population genetic analysis of retroCNVs. First, our search for retroCNVs is not exhaustive.
491 Only five representative human populations were used to identify retroCNVs, and only 98
492 singletons were identified to represent potential *de novo* retroCNVs, which may reduce the
493 statistical power of our analysis. Second, our approach ignored retroCNVs that are present in
494 the reference genomes, but absent in some other individuals. However, such retroCNVs may
495 represent a relatively small proportion of total retroduplication variants (Zhang et al. 2017).
496 Third, we were unable to derive the ancestral state of retroCNVs. Some low-frequency
497 retroCNVs may arise from losses of ancestral retrocopy insertions. Nevertheless, our

498 investigation exemplifies a first yet valuable attempts to understand the evolutionary dynamics
499 of retrocopies in the context of 3D genome organization. In the future, more thorough survey
500 of retroCNVs at large scale, coupled with long-read sequencing (as did in Feng and Li (2021)
501 and Troskie et al. (2021)), should give us a deeper understanding of the link between 3D
502 genome organization and the function and evolution of retrocopies.

503 In conclusion, we demonstrated that retrocopies and their parental genes displayed
504 significant colocalization in the nucleus, and characterized the spatial positioning and
505 functional profiling of colocalized and noncolocalized retrocopies in great detail. Our results
506 highlight the importance of 3D genome organization in the function and evolution of stripped-
507 down retrocopies.

508

509 ## Materials and Methods

510 ### Retrocopy-Parent Pair Datasets

511 A total of 5,159 human (genome assembly: GRCh37) and 5,569 mouse (genome assembly:
512 NCBIM37) retrocopies and their cognate parental genes were retrieved from the
513 Supplementary Table S1 of Carelli et al. (2016). Retrocopy-parent pairs fit any of the following
514 criteria were removed from further analyses: 1) a retrocopy or its parental gene was not located
515 on autosomes or the X chromosome; 2) a retrocopy and its parental gene were located on the
516 same chromosome; 3) the information of parental gene of a retrocopy was inferred from other
517 species (Carelli et al. 2016). Consequently, 4,427 human and 4,532 mouse interchromosomal
518 retrocopy-parent pairs were retained (**supplementary tables S1 and S3**). The number of
519 associated unique parental genes were 1,430 and 1,201 in human and mouse, respectively.

520

521 **Identification of Colocalized Retrocopy-Parent Pairs**

522 The raw interchromosomal contact matrices at 250-kb resolution for seven human cell lines
523 and one mouse cell line were downloaded from NCBI (GEO accession: GSE63525) (Rao et al.
524 2014). These contact matrices were constructed using Hi-C read pairs that were uniquely
525 mapped to the respective reference genomes (mapping quality ≥ 30). The studied cell lines are
526 GM12878 (human B-lymphoblastoids), HMEC (human mammary epithelial), HUVEC
527 (human umbilical vein endothelial), IMR90 (human lung fibroblasts), K562 (human
528 erythroleukemia), KBM7 (near haploid human myelogenous leukemia), NHEK (normal
529 human epidermal keratinocytes), and CH12-LX (murine B-lymphoblasts).

530 In the raw contact matrix, an entry $M_{i,j}$ is the number of filtered read pairs that support the
531 interaction between fragments i and j with fixed non-overlapping size (e.g., 250 kb). Only
532 contacts that were supported by at least one high-mapping-quality read pairs were listed in the
533 raw matrices. Significant interchromosomal interactions at 250-kb were identified using
534 FitHiC2 (Ay et al. 2014; Kaul et al. 2020). The Knight-Ruiz matrix balancing algorithm
535 (Knight and Ruiz 2013) was applied to normalized Hi-C data during identification. Unlike
536 intrachromosomal interactions where the observed contacts can be compared with the null
537 distribution of contacts based on the linear distances between fragments to determine statistical
538 significance, p-values of interchromosomal interactions in FitHiC2 were calculated using a
539 uniform probability model (Duan et al. 2010). At the level of q-value (Benjamini-Hochberg
540 adjusted p-value) < 0.05, 5,152,588 significant interchromosomal interactions at 250-kb were
541 identified. Due to the much shallow depth of Hi-C sequencing in the other cell lines, far fewer
542 significant interactions at q-value < 0.05 were detected. In order to perform comparable
543 analyses, we selected the top-5 million fragment pairs (~7% of all possible interchromosomal
544 pairs) with lowest p-values as significant interactions for all cell lines, as suggested by the
545 authors of FitHiC2 (Kaul et al. 2020).

546 The colocalization status between retrocopies and their parental genes were determined from
547 the interaction of their residing genomic fragments using the "findOverlaps" function of the R
548 package "GenomicInteractions" (Harmston et al. 2015). If a retrocopy and its parental gene
549 were overlapped respectively with two ends of an interchromosomal fragment pair that was
550 identified as significant interaction, then they were denoted as "colocalized", otherwise they
551 were "noncolocalized". Visualization of significant interaction between retrocopies and
552 parental genes was plotted with the R package "circlize" (Gu et al. 2014).
553

554 **Statistical Tests on Chromatin Spatial Colocalization of Retrocopy-Parent Pairs**

555 To assess the statistical significance of interchromosomal interactions between retrocopies
556 and parental genes, we performed five different simulations. 1) Random fragment pairs with
557 matching chromosomes and sequence lengths as the retrocopy-parent pairs were generated to
558 represent simulated chromatin pairs. 2) The information of retrocopies was kept unchanged
559 while sampling random genes (including coding and noncoding ones) with matching
560 chromosomes as parental genes to form simulated chromatin pairs. 3) The information of
561 parental genes was kept unchanged while sampling random fragments with matching
562 chromosomes and sequence lengths as retrocopies to be simulated chromatin pairs. 4) Similar
563 to the second simulation, but the randomly sampled genes were restricted to protein-coding
564 genes only. 5) The pairing relationship between retrocopies and parental genes was randomly
565 shuffled. For each sampling, the same number of chromatin pairs as true retrocopy-parent pairs
566 was generated, and 1,000 samplings were carried out for each simulation. Colocalization of
567 simulated chromatin pairs was determined using Hi-C data of respective cell lines, and the *p*-
568 value was calculated as the fraction of simulated data sets that exhibited a colocalization
569 frequency higher than that of true retrocopy-parent pairs. See also **supplementary fig. S2** for
570 graphic illustrations.

571

572 **Genome-wide Enrichment with Subcompartments, TADs, and SPIN States**

573 Genome-wide annotations of subcompartments in GM12878 and TADs in the seven human
574 cell lines were downloaded from NCBI (GEO accession: GSE63525) (Rao et al. 2014).
575 Subcompartment annotations for HMEC, HUVEC, IMR90, and K562 were retrieved from
576 https://cmu.box.com/s/n4jh3utmitzl88264s8bzsfcjhqnhaa0. In cell lines other than GM12878,
577 subcompartment annotation was achieved through the SNIPER computational approach, in
578 which interchromosomal interactions were imputed and subcompartments were inferred by
579 denoising autoencoder and multilayer perceptron classifier using typical Hi-C datasets with
580 moderate depth (Xiong and Ma 2019).

581 The annotation of 10 spatial compartmentalization states for K562 was downloaded from
582 https://github.com/ma-compbio/SPIN.

583 Enrichment or depletion of retrocopies, parental genes, unprocessed pseudogenes, and
584 retroCNV insertion sites in subcompartments, TAD, and SPIN states were computed using the
585 Genomic Association Tester (GAT) (Heger et al. 2013). Only mappable regions of the hg19
586 genome were included for the sampling of random regions to determine the enrichment or
587 depletion of test genomic elements. Retrocopies and parental genes were divided into
588 "colocalized" and "noncolocalized" ones based on Hi-C data of individual cell lines. To assess
589 the impact of host genes, retrocopies were further divided into "intragenic" and "intergenic"
590 subcategories, based on whether they overlap with protein-coding genes. Protein-coding genes
591 and unprocessed pseudogenes (GENCODE release 39; lifted to GRCh37) were included for
592 comparison. Note that "transcribed unprocessed pseudogenes" annotated by GENCODE were
593 not included. For retroCNVs, we used the estimated coordinates of insertion sites for
594 enrichment analysis.

595

596 **TSA-seq and Related Data**

597 Genomic tracks of SON, pSC35, lamin B, lamin A/C, and pol II TSA-seq signals, GRO-seq
598 signal, and speckle distance in bigwig format for the K562 cell line were downloaded from
599 NCBI (GEO accession: GSE81553) (Chen et al. 2018). The signal strength in 20 kb non-
600 overlapping bins were calculated using deepTools (Ramirez et al. 2016) and intersected with
601 retrocopies and parental genes using bedtools (Quinlan and Hall 2010).

602

603 **RNA-seq, TFBS, and Histone Modification Analyses**

604 The genomic alignment files in bam format of RNA-seq for GM12878, HUVEC, K562, and
605 NHEK were downloaded from the ENCODE RNA Dashboard website (https://public-
606 docs.crg.eu/rguigo/Data/jlagarde/encode_RNA_dashboard//hg19/). RNA-seq data of mouse
607 CH12-LX cell line was downloaded from ENCODE (ENCSR000AJV). Only uniquely mapped
608 reads were kept for downstream analysis. The FPKM values of retrocopies and other genes
609 were estimated using StringTie (Pertea et al. 2016) and are averaged across replicates.

610 The position frequency matrices (PFMs) of 949 human TFs were downloaded from the
611 JASPAR database (https://jaspar.genereg.net) (Castro-Mondragon et al. 2022), and FIMO
612 (Grant et al. 2011) was used to scan the occurrences of TFBSs in the flanking sequences (± 3
613 kb, ± 1 kb, or upstream 3 kb of putative TSSs) of retrocopies and parental genes. Only TFBSs
614 with q-value < 0.01 were retained for further analysis.

615 Genome-wide fold change signals of DNase I sensitivity, histone variant H2A.Z and 10
616 histone modifications (H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K9me3, H3K27ac,
617 H3K27me3, H3K36me3, H3K9me2, and H4K20me1) in bigwig format for GM12878, HMEC,
618 HUVEC, IMR90, K562, and NHEK were obtained from the Roadmap Epigenomics Project
619 (https://egg2.wustl.edu/roadmap/web_portal/). The signals of retrocopies and unprocessed

620 pseudogenes and their flanking regions (± 3 kb) were computed and visualized with DeepTools
621 (Ramirez et al. 2016).
622
623 **Logistic Regression and Evaluation of Feature Importance**
624 For GM12878, HUVEC, K562, and NHEK, we built a logistic regression model to predict
625 the colocalization status of retrocopy-parent pairs using 35 genetic and epigenetic features
626 related to retrocopies and parental genes as predictors (**supplementary table S2**). The model
627 was trained using 70% randomly selected retrocopy-parent pairs, and the remaining 30% of
628 data was used to evaluate the performance of trained model. The R package Boruta (Kursa and
629 Rudnicki 2010) was applied to assess the importance of individual features.
630
631 **Coding Potential and Evolutionary Importance of Retrocopies**
632 The RPKM values and ORF information of human and mouse genes inferred from ribosome
633 profiling data were downloaded from RPFdb (Wang et al. 2019). For human and mouse, 101
634 and 52 ribo-seq studies were included in the RPFdb database. The maximum RPKM of a given
635 gene among all ribo-seq samples was used to compare the coding potential between colocalized
636 and noncolocalized retrocopies, as did in Qian et al. (2022). The ORFs were detected with
637 RibORF (Ji et al. 2015), which combines the 3-nt periodicity and uniformity of read
638 distribution across codons of translated regions to identify translated ORFs from ribo-seq data.
639 The human 46-way phastCons scores and the mouse 60-way phastCons scores were
640 downloaded from UCSC Genome Browser. For both species, the phastCons score of the
641 placental mammal subset was used. For human, we also downloaded the precomputed
642 LINSIGHT score from http://compgen.cshl.edu/LINSIGHT/LINSIGHT.bw. LINSIGHT
643 leverages both functional and population genomic data to detect deleterious noncoding variants.

644 The average phastCons/LINSIGHT scores of retrocopies were computed with the
645 bigWigAverageOverBed tool of UCSC.
646 A total of 516,945 GWAS SNPs associating with various diseases were downloaded from
647 the National Human Genome Research Institute's (NHGRI) GWAS catalog (Buniello et al.
648 2019). Duplicated recodes were removed, leaving 254,286 unique SNPs for further analysis.
649
650 **Identification and Population Genetic Analysis of retroCNVs**
651 To investigate the insertion bias and evolutionary forces of newly inserted retrocopies in
652 relation to interchromosomal interactions, we used sideRETRO (Miller et al. 2021) to identify
653 retroCNVs using whole genome sequencing (WGS) data from the 1000 Genomes Project.
654 sideRETRO utilize abnormal alignments (i.e., discordantly aligned paired-end reads and split
655 reads) in WGS or whole exome sequencing (WES) to identify unfixed retrocopies absent in the
656 reference genome, but present in the sequenced individual. Considering the computation
657 complexity, only the WGS mapping files of 491 individuals from five diverse populations
658 (CEU, CHS, LWK, PEL, and YRI) were used. These individuals were sequenced to a high
659 depth (30×), and were good resources for detecting retroCNVs using sideRETRO.
660 The initial list of retroCNVs was further filtered with the following criteria: 1) retroCNVs
661 that were not located on autosomes and the X chromosome were removed. 2) If the insertion
662 site of a retroCNV and its parental gene were on the same chromosome, they were excluded
663 from further analysis. 3) retroCNVs with a missing rate greater than 20% were discarded.
664 To detect possible positive selection on retroCNVs, we calculated three population genetic
665 statistics for each retroCNV that had a frequency ≥ 50% in any of the five populations: 1) the
666 average nucleotide diversity (π) of 100 kb flanking regions centered at the insertion sites of
667 retroCNV carriers over that of non-carriers (π_{ratio}) (Schrider et al. 2013). 2) average Tajima's
668 D of 100 kb flanking regions centered at the insertion sites of retroCNV carriers (Llopart et al.

669 2002). 3) average LD (r^2) of 10 kb flanking regions centered at the insertion sites of retroCNV
670 carriers (Cardoso-Moreira et al. 2016). These statistics were computed for individual
671 populations using matching polymorphism data from the 1000 Genomes Project. In order to
672 evaluate the statistical significance of the estimates, we randomly generated "pseudo" insertion
673 site in the genome, while keeping the same setup of carriers and noncarriers as the focal
674 retroCNV. A total of 200 samplings were carried out and the Wilcoxon rank-sum test was used
675 to determine how likely to get a smaller π_{ratio}, a smaller Tajima's D, or a greater LD from
676 "pseudo" insertion sites as that from a true retroCNV insertion (Zhang and Tautz 2022).
677
678 **Conversion of Genomic Coordinates between Different Reference Genomes**
679 When necessary, the genomic coordinates of human hg19 (GRCh37) were converted to hg38
680 (GRCh38), and the genomic coordinates of mouse mm9 (NCBI37) were converted to mm10
681 (GRCm38) using the UCSC liftover tool, and vice versa.
682

683 ## Data Availability

684 No new data were generated or analysed in support of this research.
685

686 ## Acknowledgements

687 We thank members of the Yang laboratory for their discussion. This work was supported
688 by the Fund of Northwest A&F University (Z111021404) and the "100-Talent Program" of
689 Shaanxi Province of China (A289021612) to R.Y.
690

691 ## Author Contributions

692 R.Y. conceived and supervised the project. Y.Y. and R.Y. designed the study. Y.Y., Y.T.,
693 Z.W., and K. Z. performed data analysis. Y.Y. wrote the manuscript with input from R.Y. and
694 Y.T. All authors reviewed and approved the manuscript.

695

696 ## Supplementary Material

697 Supplementary tables S1-S6 and supplementary figures S1-S39 are available as separate
698 files.

699

700 ## References

701 Abyzov A, Iskow R, Gokcumen O, Radke DW, Balasubramanian S, Pei B, Habegger L,
702 Genomes Project C, Lee C, Gerstein M. 2013. Analysis of variable retroduplications in
703 human populations suggests coupling of retrotransposition to cell division. *Genome Res*.
704 23:2042-2052.

705 Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Donnelly P,
706 Eichler EE, Flicek P, Gabriel SB, et al. 2015. A global reference for human genetic
707 variation. *Nature*. 526:68-74.

708 Amici DR, Cingoz H, Alasady MJ, Alhayek S, Phoumyvong CM, Sahni N, Yi SS, Mendillo
709 ML. 2023. The HAPSTR2 retrogene buffers stress signaling and resilience in mammals.
710 *Nat Commun*. 14:152.

711 Ay F, Bailey TL, Noble WS. 2014. Statistical confidence estimation for Hi-C data reveals
712 regulatory chromatin contacts. *Genome Res*. 24:999-1011.

713 Bashkirova E, Lomvardas S. 2019. Olfactory receptor genes make the case for inter-
714 chromosomal interactions. *Current Opin Genet Dev*. 55:106-113.

715 Batcher K, Varney S, York D, Blacksmith M, Kidd JM, Rebhun R, Dickinson P, Bannasch D.
716 2022. Recent, full-length gene retrocopies are common in canids. *Genome Res*.
717 32:1602-1611.

718 Bertero A. 2021. RNA Biogenesis instructs functional inter-chromosomal genome architecture.
719 *Front Genet*. 12:645863.

720 Bonev B, Cavalli G. 2016. Organization and function of the 3D genome. *Nat Rev Genet*.
721 17:661-678.
722 Bouwman BAM, Crosetto N, Bienko M. 2022. RNA gradients: Shapers of 3D genome
723 architecture. *Curr Opin Cell Biol*. 74:7-12.
724 Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A,
725 Morales J, Mountjoy E, Sollis E, et al. 2019. The NHGRI-EBI GWAS Catalog of
726 published genome-wide association studies, targeted arrays and summary statistics
727 2019. *Nucleic Acids Res*. 47:D1005-D1012.
728 Cardoso-Moreira M, Arguello JR, Gottipati S, Harshman LG, Grenier JK, Clark AG. 2016.
729 Evidence for the fixation of gene duplications by positive selection in *Drosophila*.
730 *Genome Res*. 26:787-798.
731 Carelli FN, Hayakawa T, Go Y, Imai H, Warnefors M, Kaessmann H. 2016. The life history
732 of retrocopies illuminates the evolution of new mammalian genes. *Genome Res*.
733 26:301-314.
734 Casola C, Betran E. 2017. The Genomic Impact of Gene Retrocopies: What have we learned
735 from comparative genomics, population genomics, and transcriptomic analyses?
736 *Genome Biol Evol*. 9:1351-1373.
737 Castro-Mondragon JA, Riudavets-Puig R, Rauluseviciute I, Lemma RB, Turchi L, Blanc-
738 Mathieu R, Lucas J, Boddie P, Khan A, Manosalva Perez N, et al. 2022. JASPAR 2022:
739 the 9th release of the open-access database of transcription factor binding profiles.
740 *Nucleic Acids Res*. 50:D165-D173.
741 Chen S, Krinsky BH, Long M. 2013. New genes as drivers of phenotypic evolution. *Nat Rev*
742 *Genet*. 14:645-660.
743 Chen Y, Belmont AS. 2019. Genome organization around nuclear speckles. *Curr Opin Genet*
744 *Dev*. 55:91-99.
745 Chen Y, Zhang Y, Wang Y, Zhang L, Brinkman EK, Adam SA, Goldman R, van Steensel B,
746 Ma J, Belmont AS. 2018. Mapping 3D genome organization relative to nuclear
747 compartments using TSA-Seq as a cytological ruler. *J Cell Biol*. 217:4025-4048.
748 Ciomborowska-Basheer J, Staszak K, Kubiak MR, Makalowska I. 2021. Not so dead genes-
749 Retrocopies as regulators of their disease-related progenitors and hosts. *Cells-Basel*. 10.
750 Dai Z, Xiong Y, Dai X. 2014. Neighboring genes show interchromosomal colocalization after
751 their separation. *Mol Biol Evol*. 31:1166-1172.

752 Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin
753 W, Schlesinger F, et al. 2012. Landscape of transcription in human cells. *Nature*.
754 489:101-108.
755 Duan Z, Andronescu M, Schutz K, McIlwain S, Kim YJ, Lee C, Shendure J, Fields S, Blau
756 CA, Noble WS. 2010. A three-dimensional model of the yeast genome. *Nature*.
757 465:363-367.
758 Esnault C, Maestre J, Heidmann T. 2000. Human LINE retrotransposons generate processed
759 pseudogenes. *Nat Genet*. 24:363-367.
760 Ewing AD, Ballinger TJ, Earl D, Harris CC, Ding L, Wilson RK, Haussler D, Broad Institute
761 Genome S, Analysis P, Platform. 2013. Retrotransposition of gene transcripts leads to
762 structural variation in mammalian genomes. *Genome Biol*. 14:R22.
763 Feng X, Li H. 2021. Higher rates of processed pseudogene acquisition in humans and three
764 great apes revealed by long read assemblies. *Mol Biol Evol*. 38: 2958-2966.
765 Flasch DA, Macia A, Sanchez L, Ljungman M, Heras SR, Garcia-Perez JL, Wilson TE, Moran
766 JV. 2019. Genome-wide de novo L1 retrotransposition connects endonuclease activity
767 with replication. *Cell*. 177:837-851 e828.
768 Galganski L, Urbanek MO, Krzyzosiak WJ. 2017. Nuclear speckles: molecular organization,
769 biological function and role in disease. *Nucleic Acids Res*. 45:10350-10368.
770 Gossmann TI, Woolfit M, Eyre-Walker A. 2011. Quantifying the variation in the effective
771 population size within a genome. *Genetics*. 189:1389-1402.
772 Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif.
773 *Bioinform*. 27:1017-1018.
774 Gu Z, Gu L, Eils R, Schlesner M, Brors B. 2014. circlize implements and enhances circular
775 visualization in R. *Bioinform*. 30:2811-2812.
776 Harmston N, Ing-Simmons E, Perry M, Baresic A, Lenhard B. 2015. GenomicInteractions: An
777 R/Bioconductor package for manipulating and investigating chromatin interaction data.
778 *BMC Genom*. 16:963.
779 Heger A, Webber C, Goodson M, Ponting CP, Lunter G. 2013. GAT: a simulation framework
780 for testing the association of genomic intervals. *Bioinform*. 29:2046-2048.
781 Huang YF, Gulko B, Siepel A. 2017. Fast, scalable prediction of deleterious noncoding variants
782 from functional and population genomic data. *Nat Genet*. 49: 618-624.
783 Ibn-Salem J, Muro EM, Andrade-Navarro MA. 2017. Co-regulation of paralog genes in the
784 three-dimensional chromatin architecture. *Nucleic Acids Res*. 45:81-91.

785 Ji Z, Song R, Regev A, Struhl K. 2015. Many lncRNAs, 5'UTRs, and pseudogenes are
786 translated and some are likely to express functional proteins. *Elife*. 4:e08890.
787 Joo J, Cho S, Hong S, Min S, Kim K, Kumar R, Choi JM, Shin Y, Jung I. 2023. Probabilistic
788 establishment of speckle-associated inter-chromosomal interactions. *Nucleic Acids Res*.
789 Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. *Genome Res*.
790 20:1313-1326.
791 Kaessmann H, Vinckenbosch N, Long M. 2009. RNA-based gene duplication: mechanistic and
792 evolutionary insights. *Nat Rev Genet*. 10:19-31.
793 Kaul A, Bhattacharyya S, Ay F. 2020. Identifying statistically significant chromatin contacts
794 from Hi-C data with FitHiC2. *Nat Protoc*. 15:991-1012.
795 Kim J, Venkata NC, Hernandez Gonzalez GA, Khanna N, Belmont AS. 2020. Gene expression
796 amplification by nuclear speckle association. *J Cell Biol*. 219.
797 Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, Madugundu AK, Kelkar
798 DS, Isserlin R, Jain S, et al. 2014. A draft map of the human proteome. *Nature*. 509:575-
799 581.
800 Kim S, Dunham MJ, Shendure J. 2019. A combination of transcription factors mediates
801 inducible interchromosomal contacts. *Elife*. 8.
802 Kim S, Shendure J. 2019. Mechanisms of interplay between transcription factors and the 3D
803 genome. *Mol Cell*. 76:306-319.
804 Knight PA, Ruiz D. 2013. A fast algorithm for matrix balancing. *Ima J Numer Anal*. 33:1029-
805 1047.
806 Kubiak MR, Makałowska I. 2017. Protein-doding genes' retrocopies and their functions.
807 *Viruses*. 9:80.
808 Kursa MB, Rudnicki WR. 2010. Feature selection with the Boruta package. *J Stat Softw*. 36:1-
809 13.
810 Lei X, Tian X, Wang H, Xu X, Li G, Liu W, Wang D, Xiao Z, Zhang M, Li MJ, et al. 2023.
811 Noncoding SNP at rs1663689 represses ADGRG6 via interchromosomal interaction
812 and reduces lung cancer progression. *EMBO Rep*.e56212.
813 Liu Z, Chen Y, Xia Q, Liu M, Xu H, Chi Y, Deng Y, Xing D. 2023. Linking genome structures
814 to functions by simultaneous single-cell Hi-C and RNA-seq. *Science*. 380:1070-1076.
815 Llopart A, Comeron JM, Brunet FG, Lachaise D, Long M. 2002. Intron presence-absence
816 polymorphism in *Drosophila* driven by positive Darwinian selection. *Proc Natl Acad*
817 *Sci U S A*. 99:8121-8126.

818 Maass PG, Barutcu AR, Rinn JL. 2019. Interchromosomal interactions: A genomic love story
819 of kissing chromosomes. *J Cell Biol*. 218:27-38.
820 Maass PG, Barutcu AR, Weiner CL, Rinn JL. 2018. Inter-chromosomal contact properties in
821 live-cell imaging and in Hi-C. *Mol Cell*. 69:1039-1045.e1033.
822 Machado JP, Antunes A. 2020. The genomic context of retrocopies increases their chance of
823 functional relevancy in mammals. *Genomics*. 112:2410-2417.
824 Mele M, Rinn JL. 2016. "Cat's cradling" the 3D genome by the act of LncRNA transcription.
825 *Mol Cell*. 62:657-664.
826 Mighell AJ, Smith NR, Robinson PA, Markham AF. 2000. Vertebrate pseudogenes. *Febs Lett*.
827 468:109-114.
828 Miller TLA, Orpinelli Rego F, Buzzo JLL, Galante PAF. 2021. sideRETRO: a pipeline for
829 identifying somatic and polymorphic insertions of processed pseudogenes or
830 retrocopies. *Bioinform*. 37:419-421.
831 Moon BS, Huang D, Gao F, Cai M, Lyu G, Zhang L, Chen J, Lu W. 2023. Long range inter-
832 chromosomal interaction of Oct4 distal enhancer loci regulates ESCs pluripotency. *Cell*
833 *Death Discov*. 9:61.
834 Navarro FCP, Galante PAF. 2015. A Genome-wide landscape of retrocopies in primate
835 genomes. *Genome Biol Evol*. 7:2265-2275.
836 Ohno S. 1970. Evolution by gene duplication. Berlin, New York,: Springer-Verlag.
837 Patel B, Kang Y, Cui K, Litt M, Riberio MS, Deng C, Salz T, Casada S, Fu X, Qiu Y, et al.
838 2014. Aberrant TAL1 activation is mediated by an interchromosomal interaction in
839 human T-cell acute lymphoblastic leukemia. *Leukemia*. 28:349-361.
840 Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. 2016. Transcript-level expression analysis
841 of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc*. 11:1650-
842 1667.
843 Podlaha O, Zhang J. 2009. Processed pseudogenes: the 'fossilized footprints' of past gene
844 expression. *Trends Genet*. 25:429-434.
845 Qian SH, Chen L, Xiong YL, Chen ZX. 2022. Evolution and function of developmentally
846 dynamic pseudogenes in mammals. *Genome Biol*. 23:235.
847 Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic
848 features. *Bioinform*. 26:841-842.
849 Quinodoz SA, Guttman M. 2021. Essential Roles for RNA in Shaping Nuclear Organization.
850 *Cold Spring Harb Perspect Biol*. 14.

851 Quinodoz SA, Jachowicz JW, Bhat P, Ollikainen N, Banerjee AK, Goronzy IN, Blanco MR,
852 Chovanec P, Chow A, Markaki Y, et al. 2021. RNA promotes the formation of spatial
853 compartments in the nucleus. *Cell*. 184:5775-5790 e5730.
854 Quinodoz SA, Ollikainen N, Tabak B, Palla A, Schmidt JM, Detmar E, Lai MM, Shishkin AA,
855 Bhat P, Takei Y, et al. 2018. Higher-order inter-chromosomal hubs shape 3D genome
856 organization in the nucleus. *Cell*. 174:744-757 e724.
857 Ramirez F, Ryan DP, Gruning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dundar F,
858 Manke T. 2016. deepTools2: a next generation web server for deep-sequencing data
859 analysis. *Nucleic Acids Res*. 44:W160-165.
860 Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL,
861 Machol I, Omer AD, Lander ES, et al. 2014. A 3D map of the human genome at
862 kilobase resolution reveals principles of chromatin looping. *Cell*. 159:1665-1680.
863 Rosikiewicz W, Kabza M, Kosinski JG, Ciomborowska-Basheer J, Kubiak MR, Makalowska
864 I. 2017. RetrogeneDB-a database of plant and animal retrocopies. *Database (Oxford)*.
865 2017.
866 Schrider DR, Navarro FC, Galante PA, Parmigiani RB, Camargo AA, Hahn MW, de Souza SJ.
867 2013. Gene copy-number polymorphism caused by retrotransposition in humans. *PLoS*
868 *Genet*. 9:e1003242.
869 Sisu C, Pei B, Leng J, Frankish A, Zhang Y, Balasubramanian S, Harte R, Wang D, Rutenberg-
870 Schoenberg M, Clark W, et al. 2014. Comparative analysis of pseudogenes across three
871 phyla. *Proc Natl Acad Sci U S A*. 111:13361-13366.
872 Spector DL, Lamond AI. 2011. Nuclear speckles. *Cold Spring Harb Perspect Biol*. 3.
873 Sultana T, van Essen D, Siol O, Bailly-Bechet M, Philippe C, Zine El Aabidine A, Pioger L,
874 Nigumann P, Saccani S, Andrau JC, et al. 2019. The landscape of L1 retrotransposons
875 in the human genome is shaped by pre-insertion sequence biases and post-insertion
876 selection. *Mol Cell*. 74:555-570 e557.
877 Tan S, Cardoso-Moreira M, Shi W, Zhang D, Huang J, Mao Y, Jia H, Zhang Y, Chen C, Shao
878 Y, et al. 2016. LTR-mediated retroposition as a mechanism of RNA-based duplication
879 in metazoans. *Genome Res*. 26:1663-1675.
880 Troskie RL, Jafrani Y, Mercer TR, Ewing AD, Faulkner GJ, Cheetham SW. 2021. Long-read
881 cDNA sequencing identifies functional pseudogenes in the human transcriptome.
882 *Genome Biol*. 22:146.
883 van Steensel B, Furlong EEM. 2019. The role of transcription in shaping the spatial
884 organization of the genome. *Nat Rev Mol Cell Biol*. 20:327-337.

885 Vinckenbosch N, Dupanloup I, Kaessmann H. 2006. Evolutionary fate of retroposed gene
886 copies in the human genome. *Proc Natl Acad Sci U S A*. 103:3220-3225.
887 Vouzas AE, Gilbert DM. 2021. Mammalian DNA replication timing. *Cold Spring Harb*
888 *Perspect Biol*. 13.
889 Wang H, Yang L, Wang Y, Chen L, Li H, Xie Z. 2019. RPFdb v2.0: an updated database for
890 genome-wide information of translated mRNA generated from ribosome profiling.
891 *Nucleic Acids Res*. 47:D230-D234.
892 Wang Y, Zhang Y, Zhang R, van Schaik T, Zhang L, Sasaki T, Peric-Hupkes D, Chen Y,
893 Gilbert DM, van Steensel B, et al. 2021. SPIN reveals genome-wide landscape of
894 nuclear compartmentalization. *Genome Biol*. 22:36.
895 Xie T, Yang QY, Wang XT, McLysaght A, Zhang HY. 2016. Spatial colocalization of human
896 ohnolog pairs acts to maintain dosage-balance. *Mol Biol Evol*. 33:2368-2375.
897 Xiong K, Ma J. 2019. Revealing Hi-C subcompartments by imputing inter-chromosomal
898 chromatin interactions. *Nat Commun*. 10:5069.
899 Zhang J. 2003. Evolution by gene duplication: an update. *Trends Ecol Evol*. 18:292-298.
900 Zhang W, Tautz D. 2022. Tracing the origin and evolutionary fate of recent gene retrocopies
901 in natural populations of the house mouse. *Mol Biol Evol*. 39:msab360.
902 Zhang W, Xie C, Ullrich K, Zhang YE, Tautz D. 2021. The mutational load in natural
903 populations is significantly affected by high primary rates of retroposition. *Proc Natl*
904 *Acad Sci U S A*. 118: e2013043118.
905 Zhang Y, Li S, Abyzov A, Gerstein MB. 2017. Landscape and variation of novel
906 retroduplications in 26 human populations. *PLoS Comput Biol*. 13:e1005567.
907 Zhang Z, Harrison PM, Liu Y, Gerstein M. 2003. Millions of years of evolution preserved: a
908 comprehensive catalog of the processed pseudogenes in the human genome. *Genome*
909 *Res*. 13:2541-2558.
910

911 ## Figure Legends

912 **Figure 1.** Spatial colocalization between retrocopies and their parental genes. (*A*) Circos plot
913 highlighting the chromatin interactions between retrocopies and parental genes in the
914 GM12878 cell line. From outer to inner are the ideogram of human chromosomes, the density
915 of parental genes, the density of retrocopies, and significant chromatin interactions between
916 retrocopies and parental genes. (*B*) Bar plot showing the number of human interchromosomal
917 retrocopy-parent pairs as a function of the number of cell lines wherein the spatial
918 colocalization is detected. (*C*) Distribution of colocalization frequency between chromatin
919 pairs under different null models, while the red arrow denotes the colocalization frequency of
920 true retrocopy-parent pairs. Simulations and true frequency are based on Hi-C data of
921 GM12878. *P*-value < 0.001 in all cases.

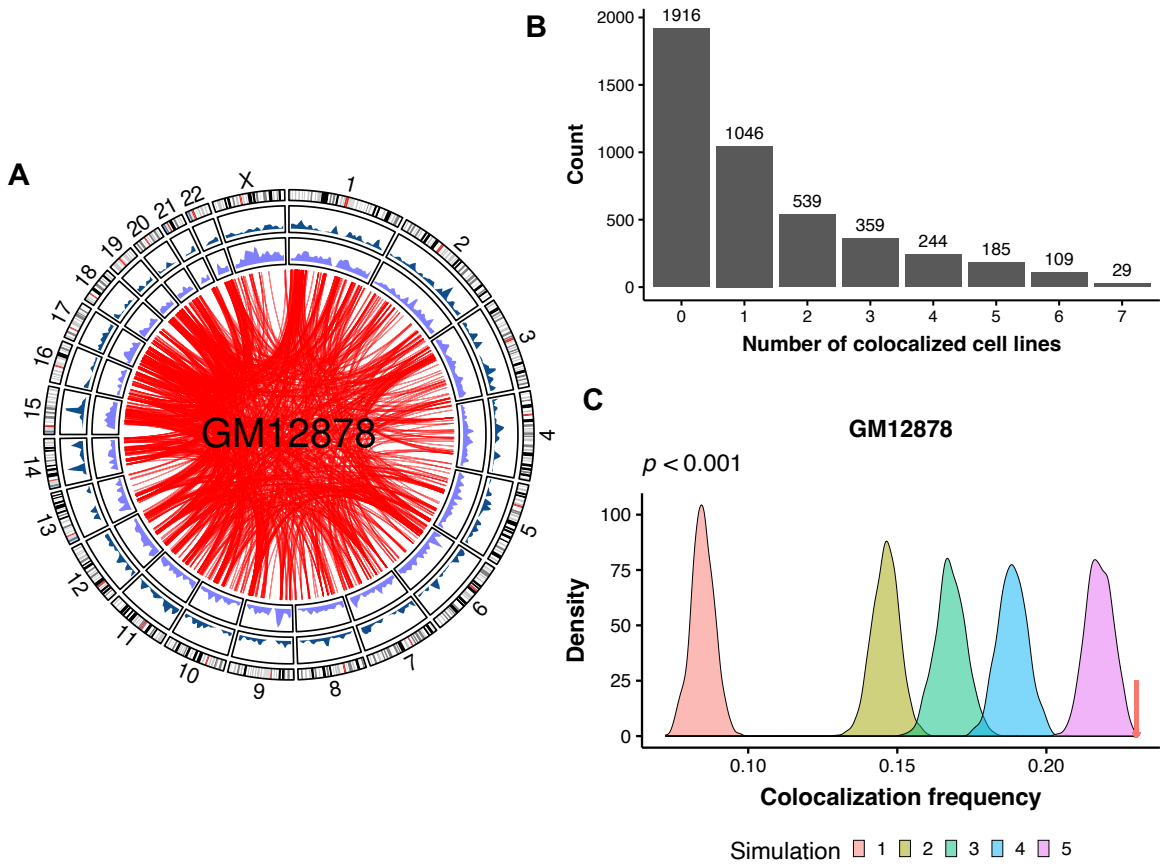
922

923 **Figure 2.** Chromatin positioning of retrocopies and parental genes in the 3D genome. (*A*)
924 Enrichments of retrocopies, parental genes, unprocessed pseudogenes, and retroCNVs in five
925 primary subcompartments and TAD in the GM12878 cell line. (*B*) Sankey diagram showing
926 the connection of colocalized (left panel) and noncolocalized (right panel) retrocopy-parent
927 pairs in the context of subcompartment annotations in GM12878. The height of bar indicates
928 the proportion of loci that fall in respective categories. Lines connect retrocopies and their
929 corresponding parent genes. (*C*) Comparison of SON TSA-seq signal between colocalized and
930 noncolocalized retrocopies/parental genes in K562. Two-tailed Wilcoxon *p*-values are
931 displayed over the bars. (*D*) Proportion of colocalized and noncolocalized retrocopies/parental
932 genes that overlapped with different SON TSA-seq deciles in K562. Decile 1 (estimated
933 distance: 0.868-1.020 μm) and decile 10 (estimated distance: 0.054-0.399 μm) represent the
934 top 10% of chromatin that are farthest and closest to SON, respectively, and other deciles are
935 those that fall between. Coloc_retro: colocalized retrocopies; noncoloc_retro: noncolocalized

936 retrocopies; coloc_parent: colocalized parental genes; noncoloc_parent: noncolocalized
937 parental genes.
938
939 **Figure 3.** Expression and regulatory element sharing of retrocopies and parental genes. (*A*)
940 Plots of cumulative fraction of expression of retrocopies (left panels) and parental genes (right
941 panels) in GM12878. The expression levels of polyadenylated (ployA+; top panels) and non-
942 polyadenylated (ployA-; bottom panels) RNAs are plotted separately. (*B*) Comparison of GRO-
943 seq signal (left panel) and Pol II TSA-seq signal (right panel) between colocalized and
944 noncolocalized retrocopies/parental genes in K562. (*C*) Comparison of the number of shared
945 TFBSs between colocalized and noncolocalized retrocopy-parent pairs. Two-tailed Wilcoxon
946 *p*-values are given in each plot to indicate the statistical significance of difference between
947 colocalized retrocopies/parental genes and their noncolocalized counterparts.
948
949 **Figure 4.** Epigenetic signals for DNase I hypersensitivity, histone variant H2A.Z, and 10
950 histone modifications along colocalized retrocopies (red lines), noncolocalized retrocopies
951 (blue lines), and unprocessed pseudogenes (gray lines) and their flanking regions (± 3 kb) in
952 the GM12878 cell line. Solid lines represent the mean fold change signal, shadowed areas
953 denote the standard error. TSS: putative transcription start site; TES: putative transcription end
954 site.
955
956 **Figure 5.** Coding potential and functional importance of colocalized and noncolocalized
957 retrocopies. (*A*-*B*) Comparison of ribo-seq RPKM values (*A*) and phastCons scores (*B*) among
958 retrocopies that are colocalized in 0, 1, and ≥ 2 cell lines. For ribo-seq data, only retrocopies
959 with RPKM ≥ 1 were included. Two-tailed Wilcoxon *p*-values are given over the bars. (*C*)
960 Percentage of colocalized and noncolocalized retrocopies in the seven cell lines that are

961 overlapped with at least one NHGRI-EBI GWAS SNPs. *P*-values of Fisher's exact tests are
962 indicated. (*D*) Correlation between phastCons score and RNA expression level (left panel) and
963 between phastCons score and ribo-seq signal (right panel) of retrocopies. Solid line and
964 shadowed area in each plot denote the linear regression line and its confident interval. Pearson
965 correlation coefficient and its associated *p*-value is given at the top left corner of each plot.
966
967 **Figure 6.** Colocalization and evolution of retroCNVs in humans. (*A*) Density plots showing
968 the distribution of colocalization frequency between chromatin pairs under different null
969 models for singleton retroCNVs (left panel) and high-frequency retroCNVs (right panel). Red
970 arrows denote the colocalization frequency between true retroCNVs and their parental genes.
971 Simulations and true frequencies are based on Hi-C data in GM12878. *P*-values are indicated
972 for simulation 5. (*B*) The presence frequency spectra for retroCNVs that are colocalized in 0,
973 1, and ≥ 2 cell lines.
974

975 Figure 1.



976

977

Figure 2.

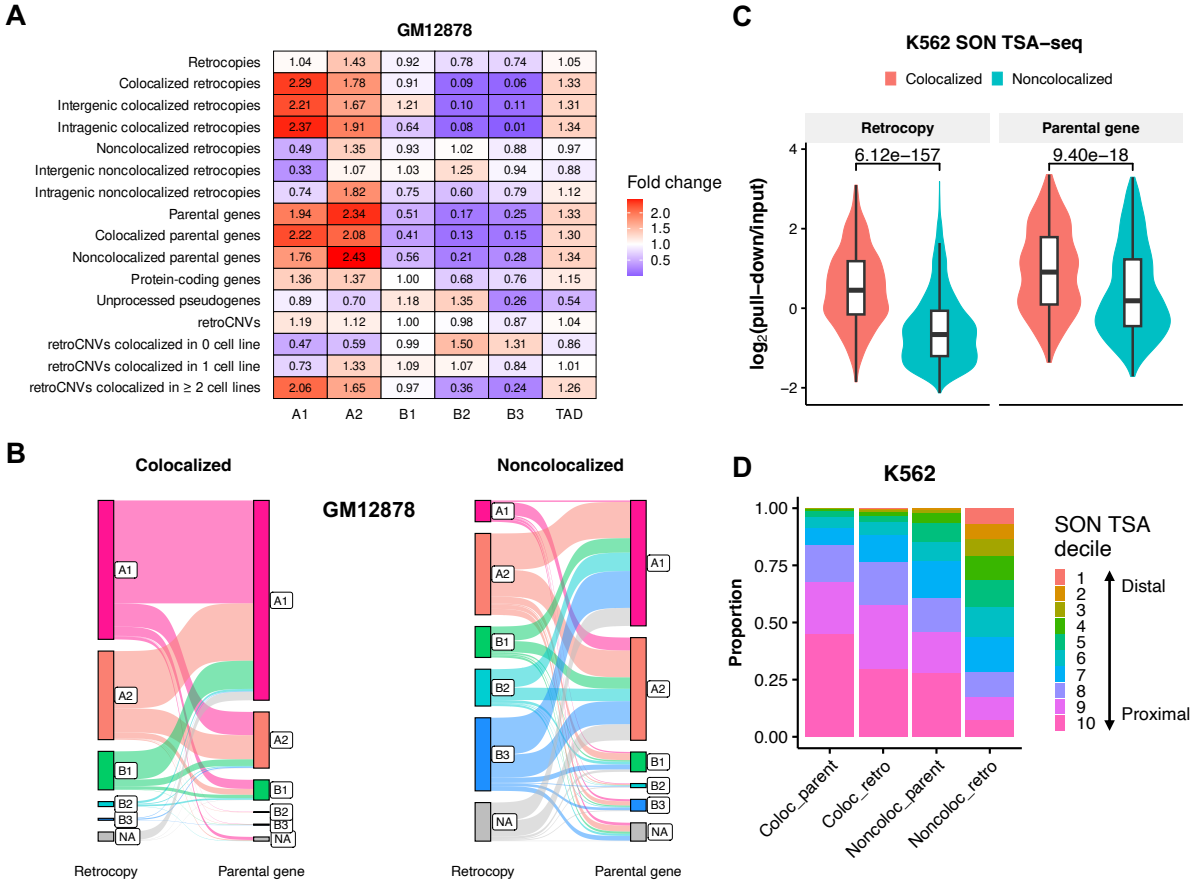

A
GM12878
Retrocopies 1.04 1.43 0.92 0.78 0.74 1.05
Colocalized retrocopies 2.29 1.78 0.91 0.09 0.06 1.33
Intergenic colocalized retrocopies 2.21 1.67 1.21 0.10 0.11 1.31
Intragenic colocalized retrocopies 2.37 1.91 0.64 0.08 0.01 1.34
Noncolocalized retrocopies 0.49 1.35 0.93 1.02 0.88 0.97
Intergenic noncolocalized retrocopies 0.33 1.07 1.03 1.25 0.94 0.88
Intragenic noncolocalized retrocopies 0.74 1.82 0.75 0.60 0.79 1.12
Parental genes 1.94 2.34 0.51 0.17 0.25 1.33
Colocalized parental genes 2.22 2.08 0.41 0.13 0.15 1.30
Noncolocalized parental genes 1.76 2.43 0.56 0.21 0.28 1.34
Protein-coding genes 1.36 1.37 1.00 0.68 0.76 1.15
Unprocessed pseudogenes 0.89 0.70 1.18 1.35 0.26 0.54
retroCNVs 1.19 1.12 1.00 0.98 0.87 1.04
retroCNVs colocalized in 0 cell line 0.47 0.59 0.99 1.50 1.31 0.86
retroCNVs colocalized in 1 cell line 0.73 1.33 0.09 1.07 0.84 1.01
retroCNVs colocalized in ≥ 2 cell lines 2.06 1.65 0.97 0.36 0.24 1.26
A1 A2 B1 B2 B3 TAD
Fold change
2.0
1.5
1.0
0.5
C
K562 SON TSA−seq
Colocalized
Noncolocalized
Retrocopy
Parental gene
6.12e−157
9.40e−18
log2(pull−down/input)
4
2
0
−2
B
Colocalized
GM12878
Noncolocalized
A1
A2
B1
B2
B3
NA
Retrocopy
Parental gene
D
K562
Proportion
1.00
0.75
0.50
0.25
0.00
Coloc_parent
Coloc_retro
Noncoloc_parent
Noncoloc_retro
SON TSA decile
1
2
3
4
5
6
7
8
9
10
Distal
Proximal

Figure 3.

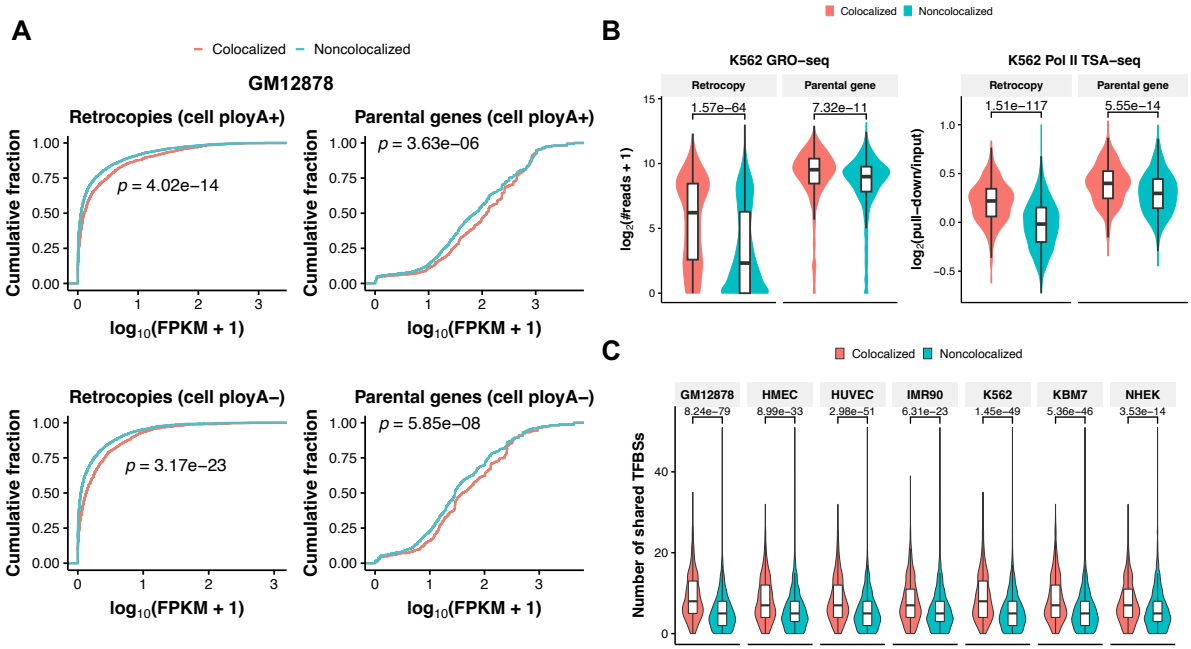

A
Colocalized
Noncolocalized
GM12878
Retrocopies (cell ployA+)
Parental genes (cell ployA+)
Retrocopies (cell ployA−)
Parental genes (cell ployA−)
Cumulative fraction
log10(FPKM + 1)
p = 4.02e−14
p = 3.63e−06
p = 3.17e−23
p = 5.85e−08
B
K562 GRO−seq
K562 Pol II TSA−seq
Retrocopy
Parental gene
1.57e−64
7.32e−11
1.51e−117
5.55e−14
log2(#reads + 1)
log2(pull−down/input)
C
GM12878
HMEC
HUVEC
IMR90
K562
KBM7
NHEK
8.24e−79
8.99e−33
2.98e−51
6.31e−23
1.45e−49
5.36e−46
3.53e−14
Number of shared TFBSs

Figure 4.

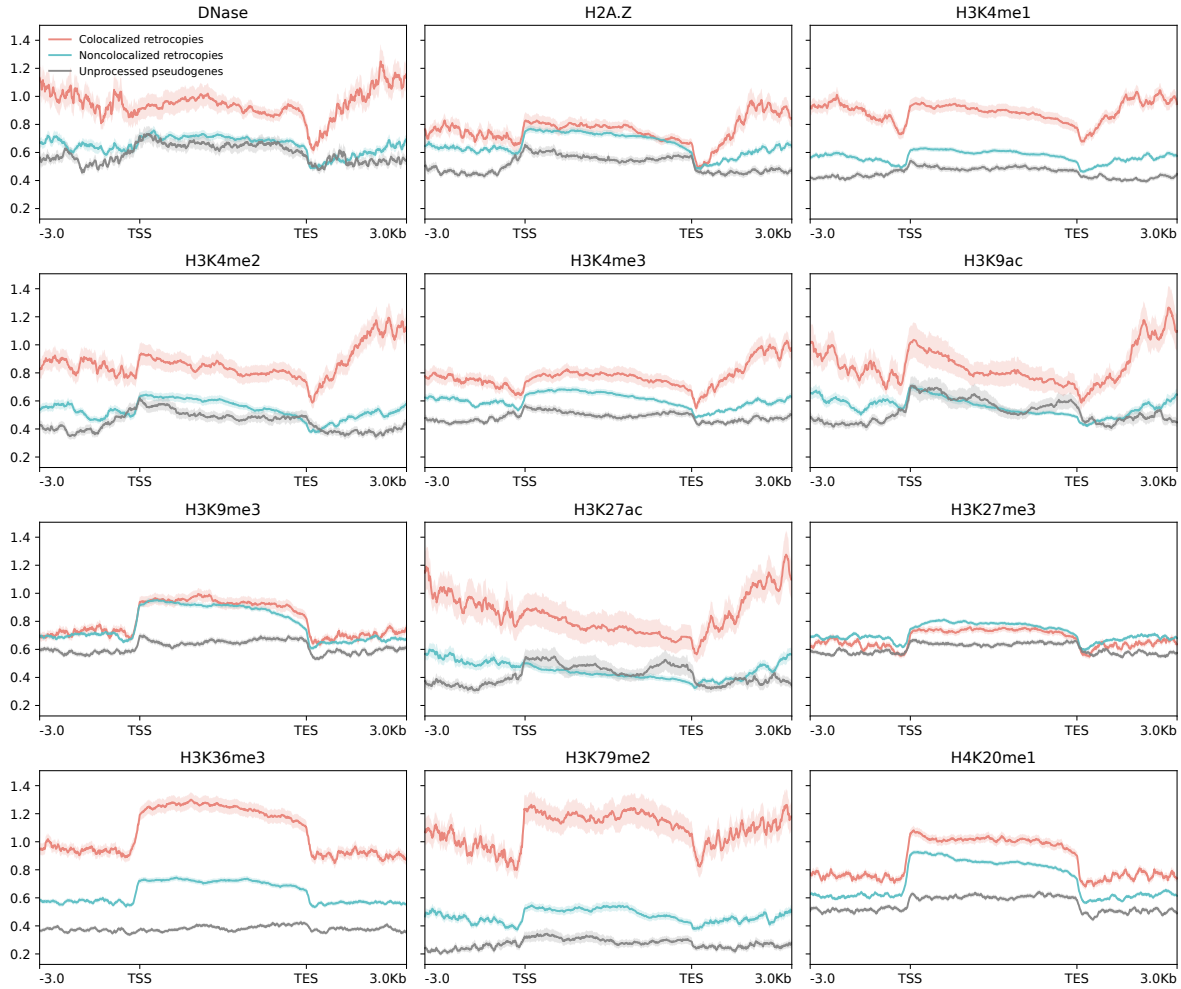
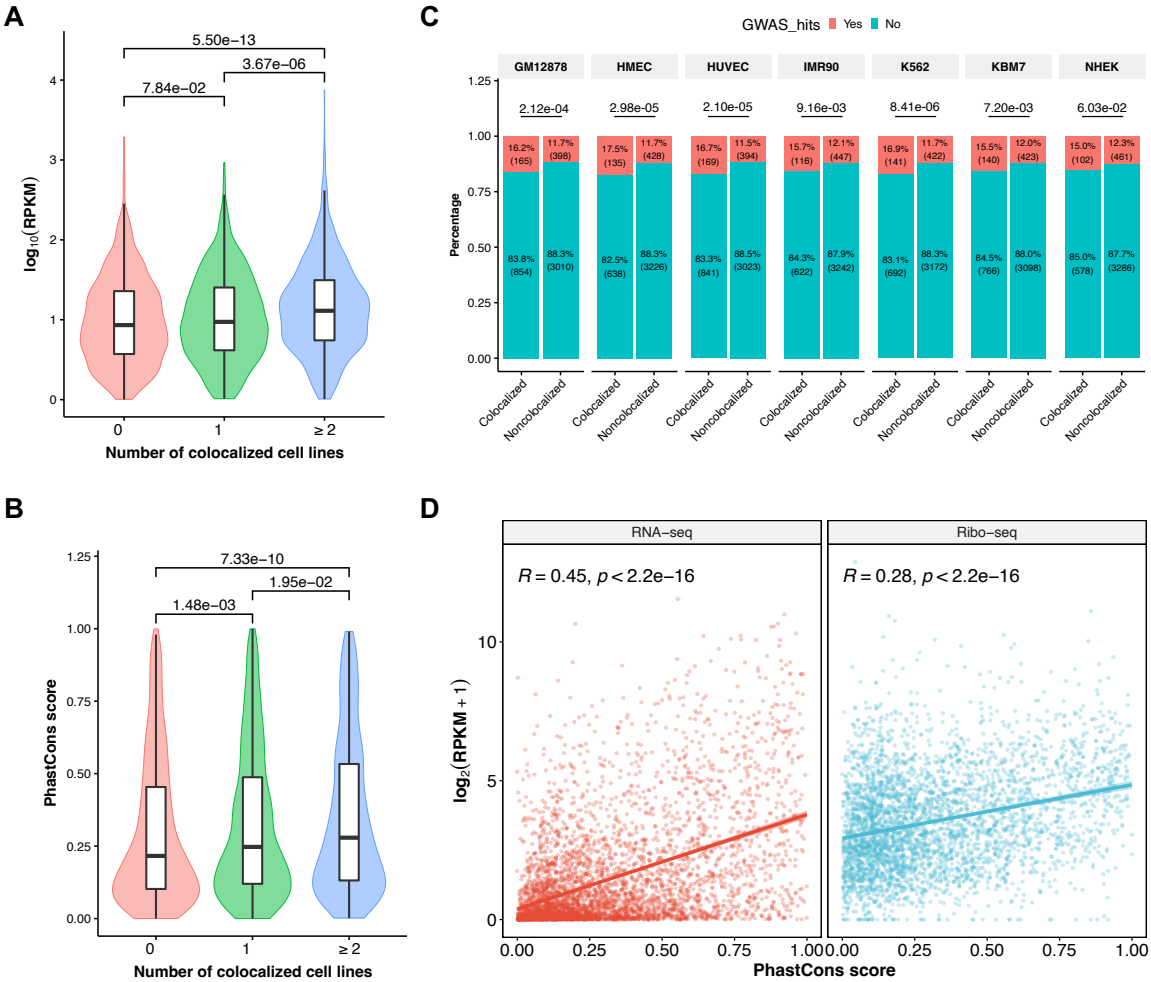
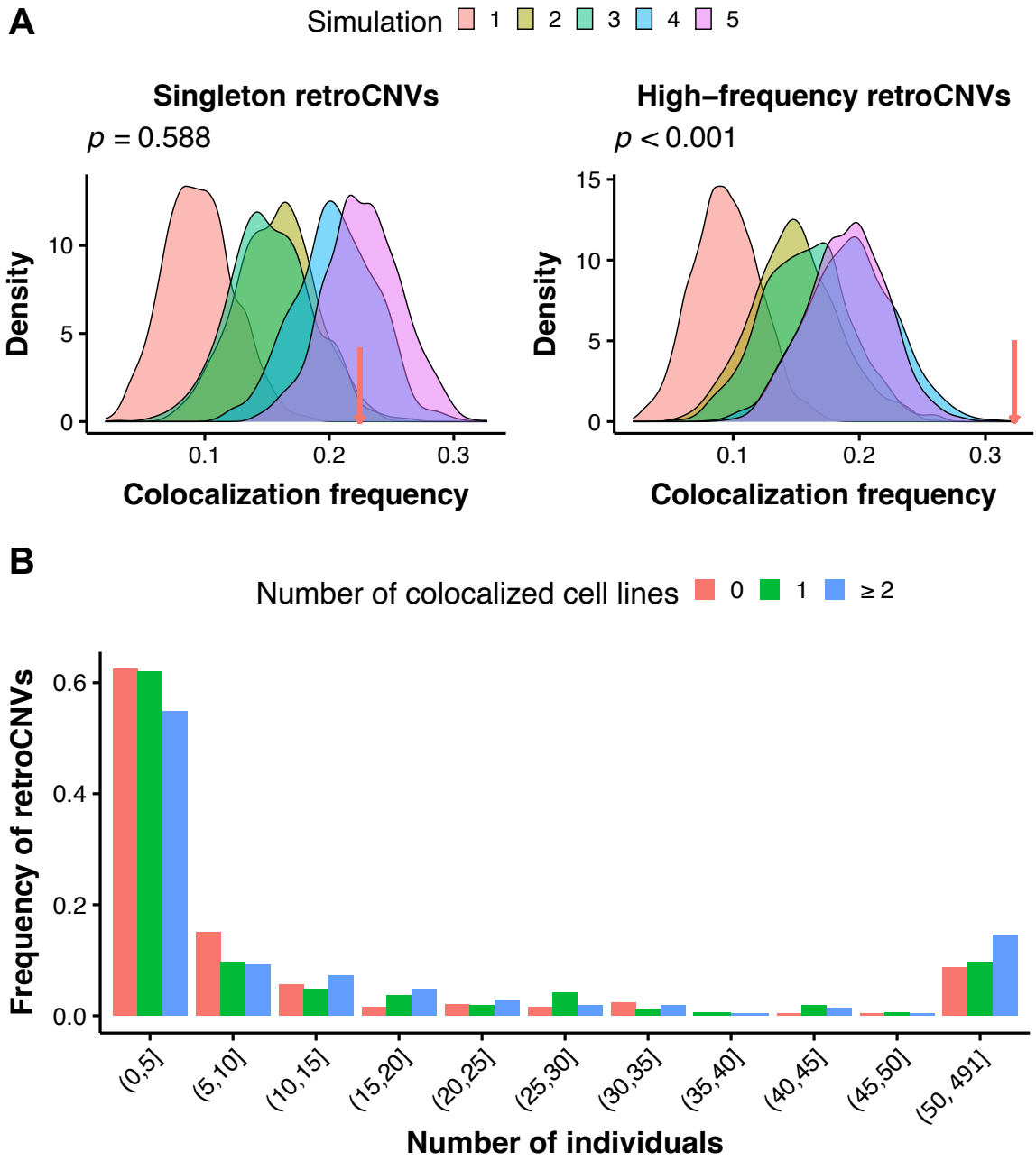

DNase
H2A.Z
H3K4me1
H3K4me2
H3K4me3
H3K9ac
H3K9me3
H3K27ac
H3K27me3
H3K36me3
H3K79me2
H4K20me1
Colocalized retrocopies
Noncolocalized retrocopies
Unprocessed pseudogenes
1.4
1.2
1.0
0.8
0.6
0.4
0.2
-3.0
TSS
TES
3.0Kb

Figure 5.


A
5.50e−13
3.67e−06
7.84e−02
log10(RPKM)
Number of colocalized cell lines
B
7.33e−10
1.95e−02
1.48e−03
PhastCons score
Number of colocalized cell lines
C
GWAS_hits Yes No
GM12878
HMEC
HUVEC
IMR90
K562
KBM7
NHEK
2.12e-04
2.98e-05
2.10e-05
9.16e-03
8.41e-06
7.20e-03
6.03e-02
Percentage
Colocalized
Noncolocalized
D
RNA−seq
Ribo−seq
R = 0.45, p < 2.2e−16
R = 0.28, p < 2.2e−16
log2(RPKM + 1)
PhastCons score

990 Figure 6.



991