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 sequencing-based and imaging-based chromatin contact maps in human and mouse cell lines and onto Hi-C interaction maps in five other mammals, 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 of 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 frequencies, 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 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

50 and promoters, rendering them to be "dead on arrival", and were frequently dismissed as
51 "processed pseudogenes" (Mighell et al. 2000; Zhang et al. 2003). On the other hand,
52 increasing number of protein-coding retrocopies (retrogenes) with explicit functions have been
53 reported in various organisms (Casola and Betran 2017). For example, a recent study reported
54 that a mammalian retrogene *HAPSTR2* encodes a protein that retains the biochemical features
55 of its parental gene *HAPSTR1* and acts as a safeguard in stress signaling and resilience (Amici
56 et al. 2023). Moreover, accumulating evidence indicates that many retrocopies are robustly
57 expressed, often in a tissue-specific manner (Carelli et al. 2016; Qian et al. 2022). These
58 retrocopy-derived RNAs may function as natural antisense transcripts (NATs), long noncoding
59 RNAs (lncRNAs), or sponging microRNAs to regulate their progenitors or host genes (Kubiak
60 and Makałowska 2017; Ciomborowska-Basheer et al. 2021). Ribosome profiling data further
61 suggest that a sizable proportion of retrocopies may be actively translated (Kim et al. 2014; Ji
62 et al. 2015; Qian et al. 2022). However, the specific proteins they produce and their functions
63 remain subjects for further investigation.

64 How some retrocopies gain their expression capability and "awake" as functional paralogs
65 continually attracts evolutionary biologists (Vinckenbosch et al. 2006). Previous research
66 discovered that retrocopies may exploit preexisting promoters, evolve new promoters from
67 scratch, or recruit proto-promoters from their genomic vicinity (Carelli et al. 2016). For
68 example, the promoter of the aforementioned retrogene *HAPSTR2* originated from a nearby
69 proto-promoter (Amici et al. 2023). However, current studies mostly focused on the role of
70 local genomic context in the expression of retrocopies (Carelli et al. 2016; Machado and
71 Antunes 2020), while the influences of high-order chromatin organization on retrocopy
72 expression and functionality are largely unexplored.

73 In eukaryotes, high-order chromatin organization and transcription are tightly linked, with
74 reciprocal influence on each other (Bonev and Cavalli 2016; van Steensel and Furlong 2019).

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

Here, we used chromatin contact maps obtained by orthogonal techniques in human and mouse cell lines and Hi-C interaction matrices in chimpanzees, rhesus macaques, marmosets, cows, and dogs to detected significant interchromosomal interactions between retrocopies and their parental genes. Through extensive simulations, we demonstrated that mammalian retrocopy-parent pairs exhibit a higher colocalization frequency than random chromatin fragment pairs under various null models. Compared to noncolocalized retrocopies, colocalized retrocopies are enriched in active subcompartments and speckles, and are more actively transcribed and translated, supporting a connection between spatial proximity and functionality of retrocopies. Such differences might be resulted from the higher number of shared regulatory elements with parental genes and the more active epigenetic landscape of colocalized retrocopies. Furthermore, by investigating retroposed gene copy number variants (retroCNVs) in human populations, we discovered that insertion bias and nonadaptive

100 evolution shaped the colocalization between retrocopies and parental genes. Our results shed
101 lights on the function and evolution of retrocopies in the context of 3D genome organization.
102

103 Results

104 Human Interchromosomal Retrocopy-Parent Pairs Exhibit Widespread Nuclear 105 Colocalization

106 To determine the extent of spatial colocalization between retrocopies and their parental genes,
107 we retrieved 5,159 human retrocopy-parent pairs from Carelli et al. (2016).After filtering steps
108 (Materials and Methods), 4,427 interchromosomal pairs were retained for subsequent analyses.
109 Using Hi-C contact maps at 250-kb resolution of seven human cell lines (GM12878, HMEC,
110 HUVEC, IMR90, K562, KBM7, and NEHK) (Rao et al. 2014), we found that between 738
111 (16.7%; in IMR90) and 1,019 (23.0%; in GM12878) retrocopy-parent pairs exhibit significant
112 interchromosomal interactions, i.e., spatial colocalization (**fig. 1A**, **supplementary fig. S1** and
113 **table S1**). More than half (2,511; 56.7%) retrocopy-parent pairs are colocalized in at least one
114 cell lines, and 1,465 (33.1%) pairs exhibit spatial proximity in two or more cell lines (**fig. 1B**).
115 We reasoned that if the spatial proximity between retrocopies and parental genes had
116 biological significance, then a higher-than-expected colocalization frequency should be
117 observed. Therefore, we performed various simulations to determine the statistical significance
118 of retrocopy-parent colocalization (**supplementary fig. S2**). First, we simulated 1,000 datasets,
119 in each of which 4,427 random fragment pairs with matching chromosomes and sequence
120 lengths as the retrocopy-parent pairs were sampled. Notably, we found that the colocalization
121 frequency of true retrocopy-parent pairs is significantly higher than simulated chromatin pairs
122 ($p < 0.001$; **fig. 1C** and **supplementary fig. S3**), suggesting nonrandom colocalization between
123 retrocopies and parental genes. Second, we kept the information of retrocopies unchanged
124 while sampling random coding and noncoding genes with matching chromosomes as parental

125 genes. Third, we kept the information of parental genes unchanged while sampling random
126 fragments with matching chromosomes and sequence lengths as retrocopies. Forth, similar to
127 the second simulation, except that we constrained the randomly sampled genes to exclusively
128 protein-coding genes. For these simulations, we also observed a significant higher
129 colocalization frequency in true retrocopy-parent pairs ($p < 0.05$, except for simulation 4 of
130 NHEK; **fig. 1C** and **supplementary fig. S3**). However, compared to the first null model, the
131 colocalization frequencies in these simulated data are higher, indicating that some properties
132 (e.g., local genomic and epigenomic environments) of retrocopies or parental genes may
133 contribute to interchromosomal colocalization.

134 Lastly, to further investigate the influence of retrocopies and parental genes on spatial
135 proximity, we randomly shuffled the relationship between retrocopies and cognate genes for
136 1,000 times. Particularly, the observed colocalization frequencies in GM12878, HMEC,
137 HUVEC, K562, and NHEK are still significantly higher than expected ($p < 0.05$), while the
138 significance is marginal in IMR90 and KBM7 ($p = 0.082$ and 0.078, respectively; **fig. 1C** and
139 **supplementary fig. S3**). To further evaluate the significance of interchromosomal
140 colocalization between retrocopies and their parental genes, we analyzed contact maps derived
141 from two orthogonal methods: 1) split-pool recognition of interactions by tag extension
142 (SPRITE), which is a sequencing-based but ligation-free technique that enables detection of
143 multiway genome-wide chromatin interactions; and 2) multiplexed error-robust fluorescence
144 in situ hybridization (MERFISH), which is an imaging-based approach to map chromatin
145 structure at genome scale. In GM12878, we observed a significant overlap in the colocalization
146 status of retrocopy-parent pairs derived from Hi-C (Rao et al. 2014) and SPRITE (Quinodoz et
147 al. 2018) when assessed at the same resolution (1 Mb) and significance level ($q < 0.05$).
148 (**supplementary fig. S4A**). Furthermore, when employing significant contacts identified by
149 SPRITE, retrocopies and their parental genes exhibit a colocalization frequency higher than

150 expected ($p < 0.05$ in all simulations; **supplementary fig. S4B**). Notably, in IMR90, retrocopy-
151 parent pairs identified as colocalized by Hi-C display shorter median spatial distances and
152 higher proximity frequencies, as measured by MERFISH (Su et al. 2020), compared to
153 noncolocalized pairs. ($p = 3.5 \times 10^{-6}$ and 2.0×10^{-13}, respectively; **supplementary fig. S5**).
154

155 **Colocalized Retrocopies Are Enriched in Active Subcompartments and Are Close to**
156 **Nuclear Speckles**

157 We next asked where did retrocopies and parental genes colocalize with respect to the 3D
158 genome organization in the nucleus. By clustering interchromosomal contact matrix of the
159 ultra-deep sequenced GM12878 cell line, Rao et al. (2014) divided the active A and inactive B
160 compartments into five primary subcompartments (A1, A2, B1, B2, and B3), which were
161 associated with distinct genomic and epigenomic features. We found that in GM12878,
162 retrocopies are slightly enriched in A2 subcompartment while depleted in B2 and B3
163 subcompartments (**fig. 2A**). Such enrichment of active subcompartments and depletion of
164 inactive subcompartments is more apparent for parental genes. In addition, compared to
165 protein-coding genes as a whole, retrocopies' parental genes show stronger enrichment of
166 active subcompartments and depletion of inactive subcompartments, which is consistent with
167 previous findings that highly expressed genes are more prone to generate retrocopies (Zhang
168 et al. 2003; Podlaha and Zhang 2009; Sisu et al. 2014).

169 Remarkably, colocalized retrocopies show strong enrichment of A1 and A2
170 subcompartments and strong depletion of B2 and B3 subcompartments, resembling the pattern
171 of parental genes (**fig. 2A**). In contrast, noncolocalized retrocopies are depleted in active A1
172 subcompartment and show little depletion of inactive subcompartments. Dividing colocalized
173 and noncolocalized retrocopies into intergenic and intragenic subclasses do not affect the
174 general pattern (**fig. 2A**), indicating that the subcompartment association is not affected by host

175 genes. Besides, parental genes and colocalized retrocopies are slightly enriched in TADs.
176 Contrary to retrocopies, unprocessed pseudogenes are depleted in A1 and A2 subcompartments
177 and TAD while slightly enriched in B1 and B2 subcompartments (**fig. 2A**), reflecting different
178 3D chromatin environments between processed and unprocessed "pseudogenes". Using
179 subcompartment annotations inferred by deep learning on relatively low resolution Hi-C maps
180 (Xiong and Ma 2019), we obtained similar results in HMEC, HUVEC, IMR90, and K562
181 (**supplementary fig. S6**).

182 Of note, for colocalized retrocopy-parent pairs, both ends tend to be overlapped with A1 or
183 A2 subcompartments, whereas for noncolocalized pairs, a substantial proportion of retrocopies
184 are overlapped with B2 or B3 subcompartments, despite the fact that the majority of parental
185 genes still overlap with active subcompartments (**fig. 2B** and **supplementary fig. S7**).

186 Subcompartments are strongly associated with subnuclear structures, such as nuclear lamina,
187 nucleolus, and speckles, which are associated with genome regulation (Chen et al. 2018). Thus,
188 we compared the spatial positioning relative to subnuclear structures of colocalized and
189 noncolocalized retrocopies using TSA-seq data in K562 (Chen et al. 2018). TSA-seq is a
190 technique that combines tyramide signal amplification (TSA) and high-throughput sequencing
191 to map chromosome organization relative to subnuclear structures genome-wide. We found
192 that the $\log_2$(pull-down/input) signals of SON and pSC35 TSA-Seq of colocalized retrocopies
193 or parental genes are significantly higher than their noncolocalized counterparts, whereas an
194 opposite trend is observed for LaminAC and LaminB TSA-Seq signals ($p < 0.001$; **fig. 2C** and
195 **supplementary figs. S8A-C**). SON and pSC35 are two marker proteins for nuclear speckles
196 while LaminAC and LaminB are markers for nuclear lamina, and high signal means shorter
197 cytological distance to the respective subnuclear structure. Thus, our results indicate that
198 colocalized retrocopies and parental genes are closer to speckles, while their noncolocalized
199 counterparts are closer to lamina.

200 SON TSA-Seq signals can be further converted to actual distance (in □m) between chromatin
201 and nuclear speckles and loci can be assigned to different TSA deciles based on their distance
202 to speckles (Chen et al. 2018). Leveraging these data, we found that colocalized retrocopies
203 have an average distance of 0.48□m to speckles, longer than that of colocalized parental genes
204 but significantly shorter than that of noncolocalized retrocopies ($p < 0.001$; **supplementary**
205 **fig. S8D**). Consistently, colocalized parental genes have the highest proportion of fragments
206 falling in the 9th and 10th SON TSA deciles (chromatin that is mostly close to speckle), followed
207 by colocalized retrocopies and noncolocalized parental genes, while only a small proportion of
208 noncolocalized retrocopies fall in the 9th and 10th deciles (**fig. 2D**). Furthermore, by integrating
209 TSA-Seq, DamID and Hi-C data in the K562 cell line, Wang et al. (2021) partitioned the
210 genome into 10 different spatial compartmentalization states relative to nuclear speckles,
211 lamina, and nucleolus. We found that colocalized retrocopies are enriched in interior active
212 regions and speckle-associated chromatin, while depleted in interior repressive regions and
213 lamina-associated chromatin, similar to the pattern of parental genes (**supplementary fig. S9**).
214 Since both ends of colocalized retrocopy-parent pairs tend to have shorter spatial distances
215 to speckles (**supplementary fig. S10A**) in comparison with noncolocalized pairs, we randomly
216 sampled protein-coding genes with matching speckle distance (± 0.01 μm away) as true
217 parental genes to form simulated chromatin pairs with retrocopies for 1,000 times. We found
218 that after controlling speckle distances, true retrocopy-parent pairs still exhibit higher-than-
219 expected colocalization frequency ($p < 0.001$; **supplementary fig. S10B**), suggesting that
220 speckle proximity cannot fully explaining retrocopy-parent colocalization. However, because
221 there are 10-50 speckles in a nucleus (Spector and Lamond, 2011), we cannot exclude the
222 possibility that retrocopies and their parental genes tend to be close to the same nuclear speckle.
223

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

226 Although retrocopies are often thought to be processed pseudogenes, most of them show
227 evidence of transcription (Carelli et al. 2016). The distinct positioning of colocalized and
228 noncolocalized retrocopies in the nucleus suggests they may have different transcription profile,
229 we then asked whether the spatial colocalization with parental genes is associated with the
230 expression of retrocopies. Using RNA-seq data of GM12878, HUVEC, K562, and NHEK from
231 the ENCODE project (Djebali et al. 2012), we have also detected evidence of transcription for
232 most retrocopies (**supplementary table S1**), with a substantial proportion of them showing
233 robust expression. For instance, when assessing polyadenylated (polyA+) RNAs in whole cells,
234 there are 772 to 1,350 (17.4% to 30.5%) retrocopies exhibiting an FPKM (Fragments Per
235 Kilobase per Million mapped fragments) ⩾ 1 (**supplementary fig. S11A**). The proportion of
236 robustly expressed (FPKM ≥ 1) retrocopies for polyA+ RNAs is larger than 15% and 20% in
237 the nucleus and cytosol, respectively (**supplementary figs. S11B and C**). There are also
238 substantial non-polyadenylated (polyA-) RNAs transcribed from retrocopies in the nucleus and
239 whole cell, but relatively few in the cytosol (**supplementary figs. S11D-F**). More importantly,
240 colocalized retrocopies and parental genes have significantly higher expression level than their
241 noncolocalized counterparts in most cell lines, for RNAs in different compartments (whole cell,
242 nucleus, and cytosol), and for both ployA+ and polyA- RNAs ($p < 0.01$; **fig. 3A** and
243 **supplementary figs. S12 and S13**). In addition, the GRO-Seq and PolII TSA-Seq data in K562
244 (Chen et al. 2018) shows that colocalized retrocopies and parental genes generate more nascent
245 RNAs and are closer to RNA polymerase II than their noncolocalized counterparts ($p < 0.001$;
246 **fig. 3B**), suggesting that RNA biogenesis is more active at colocalized loci.

247 Enhancer sharing has been hypothesized to account for the spatial colocalization and
248 coregulation of paralogous genes located on the same chromosome (Ibn-Salem et al. 2017).

249 Also, transcription factors (TFs) play a part in interchromosomal interactions (Kim and
250 Shendure 2019). We thus interrogated the relationship between shared regulatory elements and
251 spatial colocalization. As expected, we found that the flanking regions (putative transcription
252 start site, i.e., TSS ± 3 kb) of colocalized retrocopies and parental genes harbor more
253 transcription factor binding sites (TFBSs) than noncolocalized ones in most cell lines ($p < 0.05$;
254 **supplementary fig. S14**). In particular, the number of shared TFBSs between colocalized
255 retrocopies and parental genes is significantly greater than that of noncolocalized retrocopy-
256 parent pairs ($p < 0.001$; **fig. 3C**). It should be noted that the choice of flaking regions (e.g., TSS
257 ± 1 kb or upstream 3 kb) does not affect our results (**supplementary fig. S15**). Significantly,
258 colocalized retrocopies show a higher correlation coefficient of expression with their parental
259 genes across 17,282 GTEx samples in comparison with noncolocalized pairs (**supplementary**
260 **fig. S16**), supporting the involvement of shared TFBSs in retrocopy-parent coexpression.
261 Interestingly, the corresponding TFs of shared motifs are mostly zinc finger proteins, such as
262 ZNF263, ZNF384, ZNF460, and ZNF135, consistent with a recent report in which zinc finger
263 factors are found to play a role in speckle-associated interchromosomal interactions (Joo et al.
264 2023).

265 Colocalized and noncolocalized retrocopies are enriched in different subcompartments,
266 which are associated with distinct epigenomic features and may affect the expression of
267 retrocopies. We then explicitly examined the signals of 12 epigenomic features (including
268 DNase I sensitivity, the histone variant H2A.Z, and 10 histone modifications) for retrocopies.
269 Generally, colocalized retrocopies and their flanking regions have stronger signals of activating
270 histone modifications (e.g., H3K4me1, H3K27ac, H3K36me3, and H3K79me2) in comparison
271 with noncolocalized retrocopies (**fig. 4** and **supplementary figs. S17-S21**), indicating an
272 association between epigenetic landscape and interchromosomal colocalization. Notably,
273 retrocopies, even those noncolocalized ones, harbor stronger activating histone modifications

274 than unprocessed pseudogenes (**fig. 4** and **supplementary figs. S17-S21**), further supporting
275 the delimitation between processed and unprocessed "pseudogenes" in the context of 3D
276 genome environment.

277 To evaluate the relative importance of different features in the colocalization between
278 retrocopies and their parental genes, we trained a logistic regression model using 35 genetic
279 and epigenetic features of retrocopies and parental genes as predictors (**supplementary table**
280 **S2**) in each cell line where matched data were available. For GM12878, the machine learning
281 model is able to predict retrocopy-parent colocalization with an accuracy of 0.798 and an area
282 under the receiver operating characteristic curve (AUC) of 0.785 (**supplementary fig. S22A**).
283 Among the features included, the number of shared motifs between retrocopies and parental
284 genes is the most important, with various histone modifications of retrocopies also contribute
285 significantly (**supplementary fig. S22B**). The number of shared motifs and epigenetic features
286 of retrocopies are also top predictors in models of HUVEC, K562, and NHEK, but the
287 performance is worse for these cell lines (**supplementary figs. S23-25**).

288

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

290 Retrocopies may preserved the coding potentials of their parental genes before degenerative
291 mutations disrupting the open reading frames (ORFs), therefore, gaining expression is the first
292 and most important step of being functional for a retrocopy. The resulting RNAs may serve as
293 functional units *per se*, such as NATs, lncRNAs, or microRNAs (Kubiak and Makałowska
294 2017). There are also sporadic cases that retrocopies act as important protein-coding genes in
295 development and disease (Casola and Betran 2017). To further assess the functional relevance
296 of retrocopies at the translation level, we leveraged the ribosome profiling (i.e., ribo-seq) data
297 across 101 studies curated on RPFdb (Wang et al. 2019), and found that colocalized retrocopies
298 have significantly higher RPKM (Reads Per Kilobase per Million mapped reads) values than

299 noncolocalized retrocopies in five of the seven cell lines ($p < 0.001$; **fig. S26**). The difference
300 in ribo-seq signals is statistically marginal between retrocopies that are colocalized in zero and
301 one cell line ($p = 7.8 \times 10^{-2}$), while retrocopies with recurrent colocalized in two or more cell
302 lines are covered by significantly more reads than the other two categories of retrocopies ($p =$
303 5.5×10^{-13} and 3.7×10^{-6}, respectively; **fig. 5A**). Because ribo-seq reads could be derived from
304 genomic fragments randomly co-purified with ribosomes, we further screened retrocopy
305 transcripts with actively translated ORFs. These ORFs were detected using RibORF (Ji et al.
306 2015), which makes use of the 3-nt periodicity and uniformity of read distribution across
307 codons of translated regions, thus reducing spurious signals of translation. When only
308 transcripts with at least one detected ORFs were considered, colocalized retrocopies still
309 present higher ribo-seq signals than noncolocalized ones (**supplementary fig. S27**). If the
310 higher levels in transcription and translation of colocalized retrocopies did confer any
311 functional significance, they would be more evolutionarily conserved than noncolocalized ones.
312 As expected, we observed that the average phastCons conservation scores of colocalized
313 retrocopies are significantly higher than noncolocalized ones (**fig 5B** and **supplementary fig.**
314 **S28**). This pattern is retained when we use LINSIGHT scores (Huang et al. 2017) to measure
315 the conservation of retrocopies (**supplementary fig. S29**). In addition, colocalized retrocopies
316 and their putative promoters (1 kb upstream of TSSs) are more likely to overlap with NHGRI-
317 EBI GWAS single nucleotide polymorphisms (SNPs) (Buniello et al. 2019) than
318 noncolocalized ones ($p < 0.01$ in six out of the seven cell lines, Fisher exact test; **fig. 5C**),
319 further supporting their higher potential of functionality. We found that the phastCons scores
320 of retrocopies are positively correlated with their levels of RNA-seq ($R = 0.45$, $p < 2.2 \times 10^{-16}$)
321 and ribo-seq ($R = 0.28$, $p < 2.2 \times 10^{-16}$; **fig. 5D**). However, after accounting for the variation
322 of RNA expression level, the partial correlation coefficient between phastCons score and ribo-
323 seq signal is only 0.088 ($p = 5.5 \times 10^{-8}$), while the partial correlation coefficient between

324 phastCons score and RNA expression level is 0.374 ($p = 1.8 \times 10^{-12}$) when controlling for ribo-
325 seq signal, suggesting that retrocopies may be of more functional relevance at the RNA level.
326 Of particular note is that for two sister retrocopies derived from the same parental gene, the
327 colocalized one generally has higher levels of expression, motif sharedness, and conservation
328 than the noncolocalized one in GM12878 ($p = 1.4 \times 10^{-2}$, 1.5×10^{-5}, and 3.0×10^{-2},
329 respectively; **supplementary fig. S30**). But the trends are not apparent in other cell lines (data
330 not shown), possibly due to the small sample size of desired retrocopies and the nosier nature
331 of colocalization in cell lines other than GM12878.
332 In addition, to ask whether the colocalization pattern of retrocopy-parent pairs and its
333 functional significance is conserved across mammalian species, we performed similar analyses
334 with retrocopy-parent pairs of mice, chimpanzees, rhesus macaques, marmosets, cows, and
335 dogs (**supplementary table S3**) (Carelli et al. 2016). Using Hi-C maps of corresponding
336 species at 250-kb resolution (Rao et al. 2014; Zhang et al. 2019; Luo et al. 2021; Li et al. 2022),
337 we found that retrocopy-parent pairs in five out of these six species exhibit significantly higher
338 colocalization frequencies than randomly chromatin pairs under various null models
339 (**supplementary fig. S31**). The result in dogs is an exception ($p = 0.132$ for simulation 5),
340 possibly due to the large number and short length of dog's chromosomes (Li et al. 2022). In
341 mouse embryonic stem cells (mESCs), we found that colocalized retrocopy-parent pairs
342 determined by Hi-C (Bonev et al. 2017) and SPRITE (Quinodoz et al. 2018) have a great
343 proportion of overlap (**supplementary fig. S32A**). Leveraging high-resolution genome-wide
344 seqFISH+ imaging data in mESCs (Takei et al. 2023), we also observed a significant shorter
345 mean spatial distance for retrocopy-parent pairs identified as colocalized by Hi-C in
346 comparison with noncolocalized pairs ($p = 6.1 \times 10^{-94}$, **supplementary fig. S33A**). More
347 importantly, when significant interactions ($q < 0.05$) identified by SPRITE or the top 10%
348 interchromosomal chromatin pairs with the shortest spatial distances detected by seqFISH+

349 were used to determine interchromosomal colocalization in mESCs, we found a higher-than-
350 expected colocalization frequency for true retrocopy-parent pairs (**supplementary figs. S32B**
351 **and S33B**).

352 In addition, we observed that colocalized retrocopies and parental genes also expressed at a
353 higher level than noncolocalized ones in mouse CH12-LX cells (**supplementary fig. S34**).
354 Comparison of ribo-seq signals and phastCons scores indicate mouse colocalized retrocopies
355 are more actively translated and tend to be more conserved than noncolocalized ones ($p = 0.002$
356 and 0.012, respectively; **supplementary fig. S35**). Besides, the conservation of retrocopies is
357 positively correlated with their levels of transcription and translation in mouse, with a greater
358 correlation coefficient between conservation and RNA expression (**supplementary fig. S36**).
359

360 **Biased Insertion and Natural Selection Shaped the Interchromosomal Colocalization**
361 **Between Retrocopies and Parental Genes**

362 Lastly, we explored what molecular and evolutionary mechanisms might have shaped the
363 spatial colocalization between retrocopies and parental genes. To this end, we identified 625
364 polymorphic retrocopies (retroCNVs) in five human populations (CEU, CHS, LWK, PEL, and
365 YRI; 491 individuals) using sideRETRO (Miller et al. 2021) (**supplementary table S4**;
366 Materials and Methods). Similar to Zhang et al. (2021), we did not distinguish homozygotes
367 and heterozygotes, only accounted for the presence or absence of each allele in different
368 individuals. Our list captures most retroCNVs identified in other studies (Abyzov et al. 2013;
369 Ewing et al. 2013; Schrider et al. 2013; Zhang et al. 2017; Feng and Li 2021; Batcher et al.
370 2022) (**supplementary table S5**) and the neighbor-joining tree based on retroCNV presence
371 frequency is topologically the same as the tree built from genome-wide SNPs (**supplementary**
372 **fig. S37**), indicating the overall accuracy of retroCNV identification. The presence frequency
373 spectrum of retroCNVs is more skewed toward low-frequency in comparison with SNPs

374 (**supplementary fig. S38**), congruent with the highly deleterious effect of retroCNV insertions.
375 Ninety-eight retroCNVs are present in only one individual, which may be enriched in *de novo*
376 retrocopy insertions. We then performed the same simulations for these presumably *de novo*
377 retrocopies and their parental genes as did for putative "fixed" retrocopy-parent pairs
378 (**supplementary fig. S2**), and found that the colocalization frequency between *de novo*
379 retrocopy insertion sites and their parental genes is only significantly higher than expected in
380 some simulations (**fig. 6A** and **supplementary fig. S39**). This result suggests that the insertion
381 of retrocopies is not random, but nonrandom insertion cannot fully explain the current
382 colocalization pattern between retrocopies and parental genes.

383 Then we investigated the potential role of post-insertion selection in determining retrocopy-
384 parent colocalization. In contrast to *de novo* retrocopy insertions, high-frequency (occurs in ≥
385 28 individuals; n = 96, a sample size that is comparable to singletons) retroCNVs and their
386 parental genes exhibit a higher-than-expected colocalization frequency in all simulations ($p <$
387 0.01 except for simulations 4 and 5 in HUVEC. **fig. 6A** and **supplementary fig. S40**),
388 indicating the action of post-insertion selection on retrocopy survival in the context of
389 interchromosomal interactions. Furthermore, we found that the presence frequency spectrum
390 of retroCNVs exhibiting colocalization in two or more cell lines is less skewed toward low
391 frequency compared to retroCNVs showing colocalization in zero or one cell line (**fig. 6B**),
392 suggesting relaxed negative selection on colocalized retroCNVs. The difference in frequency
393 spectrum holds true for retroCNV subclasses base on colocalization status in individuals cell
394 lines (**supplementary fig. S41**). With respect to the 3D genome organization, recurrently
395 colocalized retroCNVs are enriched in the active A1 subcompartment (**fig. 2A** and
396 **supplementary fig. S4**) and interior active regions and speckles (**supplementary fig. S6**).
397 These active chromatin regions may have higher density of selected sites and/or more intense
398 selection, which are negatively correlated with local effective population size (Gossmann et al.

399 2011), leading to reduced selection efficacy on recurrently colocalized retroCNVs. Indeed, the
400 density of conserved sites (phastCons score > 0.5) flanking colocalized retroCNVs (insertion
401 site ± 50 kb) is significantly higher than noncolocalized retroCNVs (**supplementary fig. S42**).
402 Also, the unfolded site frequency spectrum of SNPs in the flanking regions (insertion site ± 50
403 kb) of recurrently colocalized retroCNVs is more skewed toward low frequency
404 (**supplementary fig. S43**).

405 Because there is also a higher proportion of intermediate-to-high-frequency alleles for
406 recurrently colocalized retroCNVs, we sought to detect whether the elevated frequency of
407 colocalized retroCNVs is driven by positive selection. If a retroCNV insertion was under strong
408 positive selection, selective sweep would: 1) reduce the nucleotide diversity (π) among
409 sequences near the insertion for retroCNV carriers compared to noncarriers; 2) reduce the
410 Tajima's D among sequences near the insertion for retroCNV carriers compared to genomic
411 background; 3) increase the linkage disequilibrium (LD) between SNPs near the insertion for
412 retroCNV carriers compared to genomic background (Zhang and Tautz 2022). Based on these
413 principles, we scanned signals of positive selection for each retroCNV with allele frequency ≥
414 0.5 in individual populations using SNPs from the 1000 Genomes Project (Altshuler et al. 2015)
415 (Materials and Methods). We found that neither colocalized nor noncolocalized high-frequency
416 retroCNVs are under positive selection using the above criteria (**supplementary table S6**).
417 Therefore, our results indicate that the intermediate-to-high-frequency of both colocalized and
418 noncolocalized retroCNVs is probably the result of nonadaptive process or weak positive
419 selection that left untraceable signal.

420

421 ## Discussion

422 In the present study, we revealed that retrocopy-parent pairs exhibit higher frequency of
423 interchromosomal colocalization than expected under various null models and that colocalized

424 retrocopies are more active in terms of transcription and translation. We further explored the
425 potential mechanisms, functional consequences, and evolutionary implications of the
426 nonrandom colocalization between retrocopies and their parental genes.

427 We focused exclusively on interactions between retrocopies and parental genes that are
428 located on different chromosomes, i.e., interchromosomal contacts. Compared to
429 intrachromosomal structures (e.g., TADs and chromatin loops), the mechanisms and functional
430 consequences of interchromosomal interactions are less well understood. Although often being
431 thought as noise or technical artifacts, emerging evidence suggests that interchromosomal
432 interactions are relevant to fundamental cellular activity, such as transcription and splicing
433 (Maass et al. 2019). Using CRISPR/Cas9 live-cell imaging, Maass et al. (2018) showed that
434 interchromosomal contacts are as stable and frequent as intrachromosomal interactions. In
435 another study, Quinodoz et al. (2018) detected two hubs of interchromosomal interactions that
436 shape 3D genome organization in the nucleus using SPRITE. Along with accumulating
437 evidences of *trans* gene regulation in health and in disease (Patel et al. 2014; Bashkirova and
438 Lomvardas 2019; Maass et al. 2019; Lei et al. 2023; Moon et al. 2023), recent progress
439 indicates that interchromosomal interactions play structural and functional roles in cells. Our
440 finding that retrocopy function is linked to *trans* interactions with parental genes also hints
441 nonrandom organization of interchromosomal contacts.

442 Interchromosomal contacts often occur at farther spatial distances than *cis* interactions
443 (Maass et al. 2018), rendering their detection difficult for regular ligation-based methods (e.g.,
444 Hi-C). In this study, our main analyses of interchromosomal interactions rely on Hi-C contact
445 maps at 250-kb resolution, given the constraints of sequencing depth. Leveraging the ultra-
446 high-depth Hi-C data in GM12878, we further detected spatial colocalization between
447 retrocopies and parental genes at 100-kb and 500kb resolutions. Under the same criterion (q-
448 value < 0.05 in FitHiC2), the number of colocalized pairs of retrocopy-parent pairs decreases

449 with the increase of resolution. Nevertheless, the colocalization pairs based on high-resolution
450 data is almost a subset of pairs detected using low-resolution contact maps (**supplementary**
451 **fig. S44**). Furthermore, retrocopies and their parental genes always possess a higher-than-
452 expected colocalization frequencies at different resolutions (**supplementary fig. S45**). More
453 importantly, chromatin contact maps derived from orthogonal methods (SPRITE, MERFISH,
454 and seqFISH+) in human and mouse also support the significant colocalization between
455 retrocopy-parent pairs. However, a limitation in our assessment of interchromosomal
456 interactions is the low resolution of certain contact maps and the scarcity of cell lines with
457 complementary data simultaneously available.

458 Subsequently, we established that colocalized and noncolocalized retrocopies exhibit distinct
459 properties in terms of subcompartment annotations and cytological distance to subnuclear
460 structures (**fig. 2 and supplementary figs. S4-7**). For subcompartments, both A1 and A2
461 harbor highly expressed genes and activating chromatin marks, and are depleted in nucleolus-
462 associated domains (NADs), but they show difference in replication timing and GC content.
463 B1 represents facultative heterochromatin that exhibits intermediate level of transcriptional
464 activity and high level of H3K27me3. B2 and B3 are associated with nuclear lamina, but B2 is
465 enriched in NADs whereas B3 is strongly depleted in NADs (Rao et al. 2014; Xiong and Ma
466 2019). Using TSA-seq data, Chen et al. (2018) further revealed that chromatin in A1
467 subcompartment is close to nuclear speckles while A2 subcompartment corresponds to region
468 with non-speckle associated transcription. Therefore, our observations (**fig. 2 and**
469 **supplementary figs. S4-7**) imply that many interchromosomal interactions between
470 retrocopies and their parental genes may be mediated by nuclear speckles. Nuclear speckles
471 are domains enriched in pre-mRNA splicing factors (Spector and Lamond 2011) and are also
472 hubs of interchromosomal interactions where active transcription takes place nearby
473 (Galganski et al. 2017; Quinodoz et al. 2018; Chen and Belmont 2019; Kim et al. 2020).

474 Remarkably, zinc finger proteins are found to facilitate the establishment of speckle-associated
475 interchromosomal contacts (Joo et al. 2023), in accordance with our findings that the number
476 of shared regulatory motifs is the most importance feature in predicting retrocopy-parent
477 colocalization and that most shared motifs are bind by TFs in the zinc finger family. Thus, our
478 findings further support a role of TFs in interchromosomal interactions (Kim et al. 2019; Kim
479 and Shendure 2019).

480 Speckle-associated interchromosomal interactions, along with more active chromatin marks
481 and higher number of shared motifs associated with colocalized retrocopies, signify colocalized
482 retrocopies should be more actively transcribed. The implications of higher expression of
483 colocalized retrocopies could be twofold. First, retrocopy-derived RNAs could function as
484 NATs, lncRNAs, or microRNAs in diverse cellular activities, including regulation of their
485 parental genes (Kubiak and Makałowska 2017). Besides, RNAs transcribed from retrocopies
486 may be processed and translated, and function at the protein level (Qian et al. 2022). Second,
487 the action of transcription and the resulting RNAs can orchestrate both intra- and inter-
488 chromosomal chromatin organization (van Steensel and Furlong 2019; Bertero 2021;
489 Quinodoz and Guttman 2021). In the nucleus, 3D genome organization and transcription are
490 intermingled, with reciprocal influence on each other. Recent single-cell Hi-C and RNA-seq
491 data discovered widespread chromatin interaction rewiring before transcription activation,
492 strongly supporting a regulatory role of chromatin organization on gene expression (Liu et al.
493 2023). Meanwhile, TFs can bring DNA fragments on the same or different chromosomes to
494 spatial proximity through direct oligomerization or cofactor oligomerization, leading to
495 transcription-mediated interactions in *cis* or *trans* (Kim et al. 2019; Kim and Shendure 2019).
496 In addition, non-coding RNAs in the nucleus may serve as chromosome scaffolds or seeds to
497 guide chromatin organization (Mele and Rinn 2016; Quinodoz et al. 2021; Bouwman et al.

498 2022). However, the actual impact of retrocopy transcription and retrocopy-derived RNAs on
499 interchromosomal interactions needs additional investigations.

500 We further showed that both biased insertion and natural selection contribute to the
501 nonrandom colocalization between retrocopies and parental genes (**fig. 6**). Mechanistically,
502 new L1 insertions are linked to DNA replication, with a preference in early-replicating regions
503 of the genome (Flasch et al. 2019; Sultana et al. 2019). Because early-replicating regions are
504 enriched with active subcompartments and speckles and are more like to be involved in
505 interchromosomal interactions (Vouzas and Gilbert 2021), if retrocopy insertions have the
506 same property as L1s, this insertion bias would partly explain the higher-than-expected
507 colocalization frequency between *de novo* retroCNVs and parental genes.

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

519 However, we are aware of some potential caveats should be considered in interpreting the
520 population genetic analysis of retroCNVs. First, our search for retroCNVs is not exhaustive.
521 Only five representative human populations were used to identify retroCNVs, and only 98
522 singletons were identified to represent potential *de novo* retroCNVs, which may reduce the

523 statistical power of our analysis. Second, our approach ignored retroCNVs that are present in
524 the reference genomes, but absent in some other individuals. However, such retroCNVs may
525 represent a relatively small proportion of total retroduplication variants (Zhang et al. 2017).
526 Third, we were unable to derive the ancestral state of retroCNVs. Some low-frequency
527 retroCNVs may arise from losses of ancestral retrocopy insertions. Nevertheless, our
528 investigation exemplifies a first yet valuable attempt to understand the evolutionary dynamics
529 of retrocopies in the context of 3D genome organization. In the future, more thorough survey
530 of retroCNVs at large scale, coupled with long-read sequencing (e.g., Feng and Li 2021;
531 Troskie et al. 2021), should give us a deeper understanding of the link between 3D genome
532 organization and the function and evolution of retrocopies.
533 In conclusion, we demonstrated that retrocopies and their parental genes displayed
534 significant colocalization in the nucleus, and characterized the spatial positioning and
535 functional profiling of colocalized and noncolocalized retrocopies in great detail. Our results
536 highlight the importance of 3D genome organization in the function and evolution of stripped-
537 down retrocopies.
538

539 ## Materials and Methods

540 **Retrocopy-Parent Pair Datasets**

541 The lists of retrocopies and their cognate parental genes of human (GRCh37), mouse
542 (NCBIM37), chimpanzee (CHIMP2.1.4), rhesus macaque (MMUL_1) were retrieved from the
543 Supplementary Table S1 of Carelli et al. (2016). The retrocopy information of marmoset
544 (C_jacchus3.2.1), cow (UMD3.1), and dog (CanFam3.1) were downloaded from
545 RetroGeneDB2 (Rosikiewicz et al. 2017). Retrocopy-parent pairs fit any of the following
546 criteria were removed from further analyses: 1) a retrocopy or its parental gene was not located
547 on autosomes or the X chromosome; 2) a retrocopy and its parental gene were located on the

548 same chromosome; 3) the information of parental gene of a retrocopy was inferred from other
549 species (Carelli et al. 2016).
550
551 **Identification of Colocalized Retrocopy-Parent Pairs**
552 The raw interchromosomal contact matrices at 250-kb resolution for seven human cell lines
553 and one mouse cell line were downloaded from NCBI (GEO accession: GSE63525) (Rao et al.
554 2014). These contact matrices were constructed using Hi-C read pairs that were uniquely
555 mapped to the respective reference genomes (mapping quality ≥ 30). The contact matrices in
556 hic format of mESC (4DNESDXUWBD9; mm10), chimpanzee (panTro6) and marmoset
557 (calJac3) induced Pluripotent Stem Cell (iPSC) (GSE116862), and macaque cortex plate
558 (GSE163177; rheMac8) were downloaded from 4DN portal or NCBI and were converted to
559 FitHiC2 format using HiCExplorer v3.7.2 (Wolff et al. 2020). For cow and dog, we
560 downloaded the raw reads of ear skin Hi-C sequencing from NCBI (GSE167581) and
561 constructed contact matrices using HiC-Pro v3.1.0 (Servant et al. 2015). Only contacts that
562 were supported by high-mapping-quality read pairs (mapping quality ≥ 30) were retained for
563 downstream analyses.
564 Significant interchromosomal interactions at 250-kb were identified using FitHiC2 v2.0.7
565 (Ay et al. 2014; Kaul et al. 2020). The Knight-Ruiz matrix balancing algorithm (Knight and
566 Ruiz 2013) was applied to normalized Hi-C data during identification. Unlike
567 intrachromosomal interactions where the observed contacts can be compared with the null
568 distribution of contacts based on the linear distances between fragments to determine statistical
569 significance, *p*-values of interchromosomal interactions in FitHiC2 were calculated using a
570 uniform probability model (Duan et al. 2010). At the level of *q*-value (Benjamini-Hochberg
571 adjusted *p*-value) < 0.05, 5,152,588 significant interchromosomal interactions at 250-kb were
572 identified. Due to the much shallow depth of Hi-C sequencing in the other cell lines and species,

573 far fewer significant interactions at q-value < 0.05 were detected. In order to perform
574 comparable analyses, we selected the top-5 million fragment pairs with lowest p-values as
575 significant interactions for all samples, as suggested by the authors of FitHiC2 (Kaul et al.
576 2020).

577 The colocalization status between retrocopies and their parental genes were determined from
578 the interaction of their residing genomic fragments using the findOverlaps function of the R
579 package GenomicInteractions v1.30.0 (Harmston et al. 2015). We converted the genomic
580 coordinates of retrocopies to that of Hi-C matrices using the UCSC liftover tool when there
581 were genome assembly mismatches. If a retrocopy and its parental gene were overlapped
582 respectively with two ends of an interchromosomal fragment pair that was identified as
583 significant interaction, then they were denoted as "colocalized", otherwise they were
584 "noncolocalized". Visualization of significant interaction between retrocopies and parental
585 genes was plotted with the R package circlize v0.4.14 (Gu et al. 2014).

586

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

588 To assess the statistical significance of interchromosomal interactions between retrocopies
589 and parental genes, we performed five different simulations. 1) Random fragment pairs with
590 matching chromosomes and sequence lengths as the retrocopy-parent pairs were generated to
591 represent simulated chromatin pairs. 2) The information of retrocopies was kept unchanged
592 while sampling random genes (including coding and noncoding ones) with matching
593 chromosomes as parental genes to form simulated chromatin pairs. 3) The information of
594 parental genes was kept unchanged while sampling random fragments with matching
595 chromosomes and sequence lengths as retrocopies to be simulated chromatin pairs. 4) Similar
596 to the second simulation, but the randomly sampled genes were restricted to protein-coding
597 genes only. 5) The pairing relationship between retrocopies and parental genes was randomly

598 shuffled. For each sampling, the same number of chromatin pairs as true retrocopy-parent pairs
599 was generated, and 1,000 samplings were carried out for each simulation. Colocalization of
600 simulated chromatin pairs was determined using Hi-C data of respective cell lines, and the *p*-
601 value was calculated as the fraction of simulated data sets that exhibited a colocalization
602 frequency higher than that of true retrocopy-parent pairs. See also **supplementary fig. S2** for
603 graphic illustrations.
604

605 **Analysis of SPRITE Data**

606 The SPRITE interchromosomal interaction matrices of GM12878 and mESC were
607 downloaded from NCBI (GSE114242). Similar to the original paper (Quinodoz et al. 2018),
608 we applied FitHiC2 v2.0.7 (Ay et al. 2014; Kaul et al. 2020) to detect significant interactions
609 using contact maps binned at 1Mb resolution based on SPRITE clusters containing 2 to 1000
610 reads without down-weighting for cluster size.
611

612 **Analysis of MERFISH and Two-layer seqFISH+ Imaging-based Data**

613 The physical coordinates (x, y, and z) of 1,041 uniformly distributed genomic loci detected
614 by MERFISH in IMR90 cells (Su et al. 2020) were downloaded from Zenodo (doi:
615 10.5281/zenodo.3928890). The spatial distance between any pair of loci was calculated as the
616 Euclidean distance between their fitted 3D Gaussian centers using the scripts provided by the
617 authors (https://github.com/ZhuangLab/Chromatin_ Analysis_2020_cell). The median
618 distance of a given pair of loci was calculated from ~5400 cells where both ends were detected.
619 The proximity frequency between two loci was computed as the number of cells in which their
620 spatial distance was less than 1 μm divided by the total number of detected cells. The
621 coordinates of DNA spots were extended to 1.5 Mb before overlapping with retrocopies and
622 parental genes.

623 For two-layer seqFISH+ data in mESC (Takei et al. 2023), we downloaded the physical
624 coordinates (x, y, and z) of 100,049 genomic loci at 25-kb resolution in 1,076 cells from
625 Zenodo (doi: 10.5281/zenodo.7693825). The spatial distance between any pair of loci binned
626 at 200-kb resolution was calculated using the scripts provided by the authors
627 (https://github.com/CaiGroup/dna-seqfish-plus-multi-omics). The mean distance between any
628 retrocopy-parent pair was obtained by overlapping the mean distance matrix at 200-kb
629 resolution calculated across detected cells.
630

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

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

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

642 Enrichment or depletion of retrocopies, parental genes, unprocessed pseudogenes, and
643 retroCNV insertion sites in subcompartments, TAD, and SPIN states were computed using the
644 Genomic Association Tester (GAT) (Heger et al. 2013). Only mappable regions of the hg19
645 genome were included for the sampling of random regions to determine the enrichment or
646 depletion of test genomic elements. Retrocopies and parental genes were divided into
647 "colocalized" and "noncolocalized" ones based on Hi-C data of individual cell lines. To assess

648 the impact of host genes, retrocopies were further divided into “intragenic” and “intergenic”
649 subcategories, based on whether they overlap with protein-coding genes. Protein-coding genes
650 and unprocessed pseudogenes (GENCODE release 39; lifted to GRCh37) were included for
651 comparison. Note that “transcribed unprocessed pseudogenes” annotated by GENCODE were
652 not included. For retroCNVs, we used the estimated coordinates of insertion sites for
653 enrichment analysis.
654

655 **TSA-seq and Related Data**

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

661 To assess the influence of speckle distance on colocalization, we carried out an additional
662 simulation, in which the information of retrocopies were kept unchanged, while sampling
663 random protein-coding genes with matching speckle distance (± 0.01 μm away) as true parental
664 genes to form simulated chromatin pairs. The *p*-value was calculated as the number of
665 occurrences where the colocalization frequency of simulated pairs is higher than true
666 retrocopy-parent pairs over 1,000 samplings.
667

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

669 The genomic alignment files in bam format of RNA-seq for GM12878, HUVEC, K562, and
670 NHEK were downloaded from the ENCODE RNA Dashboard website (https://public-
671 docs.crg.eu/rguigo/Data/jlagarde/encode_RNA_dashboard//hg19/). RNA-seq data of mouse
672 CH12-LX cell line was downloaded from ENCODE (ENCSR000AJV). Only uniquely mapped

673 reads were kept for downstream analysis. The FPKM values of retrocopies and other genes
674 were estimated using StringTie v2.2.1 (Pertea et al. 2016) and are averaged across replicates.
675 The position frequency matrices (PFMs) of 949 human TFs were downloaded from the
676 JASPAR2022 database (https://jaspar.genereg.net) (Castro-Mondragon et al. 2022), and FIMO
677 v5.4.1 (Grant et al. 2011) was used to scan the occurrences of TFBSs in the flanking sequences
678 (± 3 kb, ± 1 kb, or upstream 3 kb of putative TSSs) of retrocopies and parental genes. Only
679 TFBSs with q-value < 0.01 were retained for further analysis. The TPM (Transcripts Per
680 Million) values of human genes and transcripts across 17,382 samples were downloaded from
681 GTEx portal (release V8).
682 Genome-wide fold change signals of DNase I sensitivity, histone variant H2A.Z and 10
683 histone modifications (H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K9me3, H3K27ac,
684 H3K27me3, H3K36me3, H3K9me2, and H4K20me1) in bigwig format for GM12878, HMEC,
685 HUVEC, IMR90, K562, and NHEK were obtained from the Roadmap Epigenomics Project
686 (https://egg2.wustl.edu/roadmap/web_portal/). The signals of retrocopies and unprocessed
687 pseudogenes and their flanking regions (± 3 kb) were computed and visualized with DeepTools
688 v3.5.1 (Ramirez et al. 2016).
689
690 **Logistic Regression and Evaluation of Feature Importance**
691 For GM12878, HUVEC, K562, and NHEK, we built a logistic regression model to predict
692 the colocalization status of retrocopy-parent pairs using 35 genetic and epigenetic features
693 related to retrocopies and parental genes as predictors (**supplementary table S2**). The model
694 was trained using 70% randomly selected retrocopy-parent pairs, and the remaining 30% of
695 data was used to evaluate the performance of trained model. The R package Boruta v7.0.0
696 (Kursa and Rudnicki 2010) was applied to assess the importance of individual features.
697

698 **Coding Potential and Evolutionary Importance of Retrocopies**

699 The RPKM values and ORF information of human and mouse genes inferred from ribosome
700 profiling data were downloaded from RPFdb v2.0 (Wang et al. 2019). For human and mouse,
701 101 and 52 ribo-seq studies were included in the RPFdb database. The maximum RPKM of a
702 given gene among all ribo-seq samples was used to compare the coding potential between
703 colocalized and noncolocalized retrocopies, similar to Qian et al. (2022). The ORFs were
704 detected with RibORF (Ji et al. 2015), which combines the 3-nt periodicity and uniformity of
705 read distribution across codons of translated regions to identify translated ORFs from ribo-seq
706 data.

707 The human 46-way phastCons scores and the mouse 60-way phastCons scores were
708 downloaded from UCSC Genome Browser. For both species, the phastCons score of the
709 placental mammal subset was used. For human, we also downloaded the precomputed
710 LINSIGHT score from http://compgen.cshl.edu/LINSIGHT/LINSIGHT.bw. LINSIGHT
711 leverages both functional and population genomic data to detect deleterious noncoding variants.
712 The average phastCons/LINSIGHT scores of retrocopies were computed with the
713 bigWigAverageOverBed tool of UCSC.

714 A total of 516,945 GWAS SNPs associating with various diseases were downloaded from
715 the National Human Genome Research Institute's (NHGRI) GWAS catalog (Buniello et al.
716 2019). Duplicated recodes were removed, leaving 254,286 unique SNPs for further analysis.

717

718 **Identification and Population Genetic Analysis of retroCNVs**

719 To investigate the insertion bias and evolutionary forces of newly inserted retrocopies in
720 relation to interchromosomal interactions, we used sideRETRO v1.1.3 (Miller et al. 2021) to
721 identify retroCNVs using whole genome sequencing (WGS) data from the 1000 Genomes
722 Project. sideRETRO utilize abnormal alignments (i.e., discordantly aligned paired-end reads

723 and split reads) in WGS or whole exome sequencing (WES) to identify unfixed retrocopies
724 absent in the reference genome, but present in the sequenced individual. Considering the
725 computation complexity, only the WGS mapping files of 491 individuals from five diverse
726 populations (CEU, CHS, LWK, PEL, and YRI) were used. These individuals were sequenced
727 to a high depth (30×), and were good resources for detecting retroCNVs using sideRETRO.
728 The initial list of retroCNVs was further filtered with the following criteria: 1) retroCNVs
729 that were not located on autosomes and the X chromosome were removed. 2) If the insertion
730 site of a retroCNV and its parental gene were on the same chromosome, they were excluded
731 from further analysis. 3) retroCNVs with a missing rate greater than 20% were discarded.
732 To detect possible positive selection on retroCNVs, we calculated three population genetic
733 statistics for each retroCNV that had a frequency ≥ 50% in any of the five populations: 1) the
734 average nucleotide diversity (π) of 100 kb flanking regions centered at the insertion sites of
735 retroCNV carriers over that of non-carriers (π_{ratio}) (Schrider et al. 2013). 2) average Tajima's
736 D of 100 kb flanking regions centered at the insertion sites of retroCNV carriers (Llopart et al.
737 2002). 3) average LD (r^2) of 10 kb flanking regions centered at the insertion sites of retroCNV
738 carriers (Cardoso-Moreira et al. 2016). These statistics were computed for individual
739 populations using matching polymorphism data from the 1000 Genomes Project. In order to
740 evaluate the statistical significance of the estimates, we randomly generated "pseudo" insertion
741 site in the genome, while keeping the same setup of carriers and noncarriers as the focal
742 retroCNV. A total of 200 samplings were carried out and the Wilcoxon rank-sum test was used
743 to determine how likely to get a smaller π_{ratio}, a smaller Tajima's D, or a greater LD from
744 "pseudo" insertion sites as that from a true retroCNV insertion (Zhang and Tautz 2022).
745

Data Availability

No new data were generated or analysed in support of this research. The custom written scripts used in this study are available at GitHub (https://github.com/yanyubin/Retrocopy_Evolution). The sources of public data and their corresponding genome assembly versions are summarized in supplementary table S7.

Acknowledgements

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

Author Contributions

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

Supplementary Material

Supplementary tables S1-S7 and supplementary figures S1-S45 are available as separate files.

References

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

771 Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Donnelly P,
772 Eichler EE, Flicek P, Gabriel SB, et al. 2015. A global reference for human genetic
773 variation. *Nature*. 526:68-74.
774 Amici DR, Cingoz H, Alasady MJ, Alhayek S, Phoumyvong CM, Sahni N, Yi SS, Mendillo
775 ML. 2023. The HAPSTR2 retrogene buffers stress signaling and resilience in mammals.
776 *Nat Commun*. 14:152.
777 Ay F, Bailey TL, Noble WS. 2014. Statistical confidence estimation for Hi-C data reveals
778 regulatory chromatin contacts. *Genome Res*. 24:999-1011.
779 Bashkirova E, Lomvardas S. 2019. Olfactory receptor genes make the case for inter-
780 chromosomal interactions. *Current Opin Genet Dev*. 55:106-113.
781 Batcher K, Varney S, York D, Blacksmith M, Kidd JM, Rebhun R, Dickinson P, Bannasch D.
782 2022. Recent, full-length gene retrocopies are common in canids. *Genome Res*.
783 32:1602-1611.
784 Bertero A. 2021. RNA Biogenesis instructs functional inter-chromosomal genome architecture.
785 *Front Genet*. 12:645863.
786 Bonev B, Cavalli G. 2016. Organization and function of the 3D genome. *Nat Rev Genet*.
787 17:661-678.
788 Bonev B, Mendelson Cohen N, Szabo Q, Fritsch L, Papadopoulos GL, Lubling Y, Xu X, Lv
789 X, Hugnot JP, Tanay A, et al. 2017. Multiscale 3D genome rewiring during mouse
790 neural development. *Cell*. 171:557-572 e524.
791 Bouwman BAM, Crosetto N, Bienko M. 2022. RNA gradients: Shapers of 3D genome
792 architecture. *Curr Opin Cell Biol*. 74:7-12.
793 Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A,
794 Morales J, Mountjoy E, Sollis E, et al. 2019. The NHGRI-EBI GWAS Catalog of
795 published genome-wide association studies, targeted arrays and summary statistics
796 2019. *Nucleic Acids Res*. 47:D1005-D1012.
797 Cardoso-Moreira M, Arguello JR, Gottipati S, Harshman LG, Grenier JK, Clark AG. 2016.
798 Evidence for the fixation of gene duplications by positive selection in *Drosophila*.
799 *Genome Res*. 26:787-798.
800 Carelli FN, Hayakawa T, Go Y, Imai H, Warnefors M, Kaessmann H. 2016. The life history
801 of retrocopies illuminates the evolution of new mammalian genes. *Genome Res*.
802 26:301-314.

803 Casola C, Betran E. 2017. The Genomic Impact of Gene Retrocopies: What have we learned
804 from comparative genomics, population genomics, and transcriptomic analyses?
805 *Genome Biol Evol*. 9:1351-1373.
806 Castro-Mondragon JA, Riudavets-Puig R, Rauluseviciute I, Lemma RB, Turchi L, Blanc-
807 Mathieu R, Lucas J, Boddie P, Khan A, Manosalva Perez N, et al. 2022. JASPAR 2022:
808 the 9th release of the open-access database of transcription factor binding profiles.
809 *Nucleic Acids Res*. 50:D165-D173.
810 Chen S, Krinsky BH, Long M. 2013. New genes as drivers of phenotypic evolution. *Nat Rev*
811 *Genet*. 14:645-660.
812 Chen Y, Belmont AS. 2019. Genome organization around nuclear speckles. *Curr Opin Genet*
813 *Dev*. 55:91-99.
814 Chen Y, Zhang Y, Wang Y, Zhang L, Brinkman EK, Adam SA, Goldman R, van Steensel B,
815 Ma J, Belmont AS. 2018. Mapping 3D genome organization relative to nuclear
816 compartments using TSA-Seq as a cytological ruler. *J Cell Biol*. 217:4025-4048.
817 Ciomborowska-Basheer J, Staszak K, Kubiak MR, Makalowska I. 2021. Not so dead genes-
818 Retrocopies as regulators of their disease-related progenitors and hosts. *Cells-Basel*. 10.
819 Dai Z, Xiong Y, Dai X. 2014. Neighboring genes show interchromosomal colocalization after
820 their separation. *Mol Biol Evol*. 31:1166-1172.
821 Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin
822 W, Schlesinger F, et al. 2012. Landscape of transcription in human cells. *Nature*.
823 489:101-108.
824 Duan Z, Andronescu M, Schutz K, McIlwain S, Kim YJ, Lee C, Shendure J, Fields S, Blau
825 CA, Noble WS. 2010. A three-dimensional model of the yeast genome. *Nature*.
826 465:363-367.
827 Esnault C, Maestre J, Heidmann T. 2000. Human LINE retrotransposons generate processed
828 pseudogenes. *Nat Genet*. 24:363-367.
829 Ewing AD, Ballinger TJ, Earl D, Harris CC, Ding L, Wilson RK, Haussler D, Broad Institute
830 Genome S, Analysis P, Platform. 2013. Retrotransposition of gene transcripts leads to
831 structural variation in mammalian genomes. *Genome Biol*. 14:R22.
832 Feng X, Li H. 2021. Higher rates of processed pseudogene acquisition in humans and three
833 great apes revealed by long read assemblies. *Mol Biol Evol*. 38: 2958-2966.
834 Flasch DA, Macia A, Sanchez L, Ljungman M, Heras SR, Garcia-Perez JL, Wilson TE, Moran
835 JV. 2019. Genome-wide de novo L1 retrotransposition connects endonuclease activity
836 with replication. *Cell*. 177:837-851 e828.

837 Galganski L, Urbanek MO, Krzyzosiak WJ. 2017. Nuclear speckles: molecular organization,
838 biological function and role in disease. *Nucleic Acids Res*. 45:10350-10368.
839 Gossmann TI, Woolfit M, Eyre-Walker A. 2011. Quantifying the variation in the effective
840 population size within a genome. *Genetics*. 189:1389-1402.
841 Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif.
842 *Bioinform*. 27:1017-1018.
843 Gu Z, Gu L, Eils R, Schlesner M, Brors B. 2014. circlize implements and enhances circular
844 visualization in R. *Bioinform*. 30:2811-2812.
845 Harmston N, Ing-Simmons E, Perry M, Baresic A, Lenhard B. 2015. GenomicInteractions: An
846 R/Bioconductor package for manipulating and investigating chromatin interaction data.
847 *BMC Genom*. 16:963.
848 Heger A, Webber C, Goodson M, Ponting CP, Lunter G. 2013. GAT: a simulation framework
849 for testing the association of genomic intervals. *Bioinform*. 29:2046-2048.
850 Huang YF, Gulko B, Siepel A. 2017. Fast, scalable prediction of deleterious noncoding variants
851 from functional and population genomic data. *Nat Genet*. 49: 618-624.
852 Ibn-Salem J, Muro EM, Andrade-Navarro MA. 2017. Co-regulation of paralog genes in the
853 three-dimensional chromatin architecture. *Nucleic Acids Res*. 45:81-91.
854 Ji Z, Song R, Regev A, Struhl K. 2015. Many lncRNAs, 5'UTRs, and pseudogenes are
855 translated and some are likely to express functional proteins. *Elife*. 4:e08890.
856 Joo J, Cho S, Hong S, Min S, Kim K, Kumar R, Choi JM, Shin Y, Jung I. 2023. Probabilistic
857 establishment of speckle-associated inter-chromosomal interactions. *Nucleic Acids Res*.
858 Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. *Genome Res*.
859 20:1313-1326.
860 Kaessmann H, Vinckenbosch N, Long M. 2009. RNA-based gene duplication: mechanistic and
861 evolutionary insights. *Nat Rev Genet*. 10:19-31.
862 Kaul A, Bhattacharyya S, Ay F. 2020. Identifying statistically significant chromatin contacts
863 from Hi-C data with FitHiC2. *Nat Protoc*. 15:991-1012.
864 Kim J, Venkata NC, Hernandez Gonzalez GA, Khanna N, Belmont AS. 2020. Gene expression
865 amplification by nuclear speckle association. *J Cell Biol*. 219.
866 Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, Madugundu AK, Kelkar
867 DS, Isserlin R, Jain S, et al. 2014. A draft map of the human proteome. *Nature*. 509:575-
868 581.
869 Kim S, Dunham MJ, Shendure J. 2019. A combination of transcription factors mediates
870 inducible interchromosomal contacts. *Elife*. 8.

871 Kim S, Shendure J. 2019. Mechanisms of interplay between transcription factors and the 3D
872 genome. *Mol Cell*. 76:306-319.
873 Knight PA, Ruiz D. 2013. A fast algorithm for matrix balancing. *Ima J Numer Anal*. 33:1029-
874 1047.
875 Kubiak MR, Makałowska I. 2017. Protein-doding genes' retrocopies and their functions.
876 *Viruses*. 9:80.
877 Kursa MB, Rudnicki WR. 2010. Feature selection with the Boruta package. *J Stat Softw*. 36:1-
878 13.
879 Lei X, Tian X, Wang H, Xu X, Li G, Liu W, Wang D, Xiao Z, Zhang M, Li MJ, et al. 2023.
880 Noncoding SNP at rs1663689 represses ADGRG6 via interchromosomal interaction
881 and reduces lung cancer progression. *EMBO Rep*.e56212.
882 Li D, He M, Tang Q, Tian S, Zhang J, Li Y, Wang D, Jin L, Ning C, Zhu W, et al. 2022.
883 Comparative 3D genome architecture in vertebrates. *BMC Biol*. 20:99.
884 Liu Z, Chen Y, Xia Q, Liu M, Xu H, Chi Y, Deng Y, Xing D. 2023. Linking genome structures
885 to functions by simultaneous single-cell Hi-C and RNA-seq. *Science*. 380:1070-1076.
886 Llopart A, Comeron JM, Brunet FG, Lachaise D, Long M. 2002. Intron presence-absence
887 polymorphism in *Drosophila* driven by positive Darwinian selection. *Proc Natl Acad*
888 *Sci U S A*. 99:8121-8126.
889 Luo X, Liu Y, Dang D, Hu T, Hou Y, Meng X, Zhang F, Li T, Wang C, Li M, et al. 2021. 3D
890 Genome of macaque fetal brain reveals evolutionary innovations during primate
891 corticogenesis. *Cell*. 184:723-740 e721.
892 Maass PG, Barutcu AR, Rinn JL. 2019. Interchromosomal interactions: A genomic love story
893 of kissing chromosomes. *J Cell Biol*. 218:27-38.
894 Maass PG, Barutcu AR, Weiner CL, Rinn JL. 2018. Inter-chromosomal contact properties in
895 live-cell imaging and in Hi-C. *Mol Cell*. 69:1039-1045.e1033.
896 Machado JP, Antunes A. 2020. The genomic context of retrocopies increases their chance of
897 functional relevancy in mammals. *Genomics*. 112:2410-2417.
898 Mele M, Rinn JL. 2016. "Cat's cradling" the 3D genome by the act of LncRNA transcription.
899 *Mol Cell*. 62:657-664.
900 Mighell AJ, Smith NR, Robinson PA, Markham AF. 2000. Vertebrate pseudogenes. *Febs Lett*.
901 468:109-114.
902 Miller TLA, Orpinelli Rego F, Buzzo JLL, Galante PAF. 2021. sideRETRO: a pipeline for
903 identifying somatic and polymorphic insertions of processed pseudogenes or
904 retrocopies. *Bioinform*. 37:419-421.

905 Moon BS, Huang D, Gao F, Cai M, Lyu G, Zhang L, Chen J, Lu W. 2023. Long range inter-
906 chromosomal interaction of Oct4 distal enhancer loci regulates ESCs pluripotency. *Cell*
907 *Death Discov*. 9:61.
908 Navarro FCP, Galante PAF. 2015. A Genome-wide landscape of retrocopies in primate
909 genomes. *Genome Biol Evol*. 7:2265-2275.
910 Ohno S. 1970. Evolution by gene duplication. Berlin, New York,: Springer-Verlag.
911 Patel B, Kang Y, Cui K, Litt M, Riberio MS, Deng C, Salz T, Casada S, Fu X, Qiu Y, et al.
912 2014. Aberrant TAL1 activation is mediated by an interchromosomal interaction in
913 human T-cell acute lymphoblastic leukemia. *Leukemia*. 28:349-361.
914 Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. 2016. Transcript-level expression analysis
915 of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc*. 11:1650-
916 1667.
917 Podlaha O, Zhang J. 2009. Processed pseudogenes: the 'fossilized footprints' of past gene
918 expression. *Trends Genet*. 25:429-434.
919 Qian SH, Chen L, Xiong YL, Chen ZX. 2022. Evolution and function of developmentally
920 dynamic pseudogenes in mammals. *Genome Biol*. 23:235.
921 Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic
922 features. *Bioinform*. 26:841-842.
923 Quinodoz SA, Guttman M. 2021. Essential Roles for RNA in Shaping Nuclear Organization.
924 *Cold Spring Harb Perspect Biol*. 14.
925 Quinodoz SA, Jachowicz JW, Bhat P, Ollikainen N, Banerjee AK, Goronzy IN, Blanco MR,
926 Chovanec P, Chow A, Markaki Y, et al. 2021. RNA promotes the formation of spatial
927 compartments in the nucleus. *Cell*. 184:5775-5790 e5730.
928 Quinodoz SA, Ollikainen N, Tabak B, Palla A, Schmidt JM, Detmar E, Lai MM, Shishkin AA,
929 Bhat P, Takei Y, et al. 2018. Higher-order inter-chromosomal hubs shape 3D genome
930 organization in the nucleus. *Cell*. 174:744-757 e724.
931 Ramirez F, Ryan DP, Gruning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dundar F,
932 Manke T. 2016. deepTools2: a next generation web server for deep-sequencing data
933 analysis. *Nucleic Acids Res*. 44:W160-165.
934 Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL,
935 Machol I, Omer AD, Lander ES, et al. 2014. A 3D map of the human genome at
936 kilobase resolution reveals principles of chromatin looping. *Cell*. 159:1665-1680.

937 Rosikiewicz W, Kabza M, Kosinski JG, Ciomborowska-Basheer J, Kubiak MR, Makalowska
938 I. 2017. RetrogeneDB-a database of plant and animal retrocopies. *Database (Oxford)*.
939 2017.
940 Schrider DR, Navarro FC, Galante PA, Parmigiani RB, Camargo AA, Hahn MW, de Souza SJ.
941 2013. Gene copy-number polymorphism caused by retrotransposition in humans. *PLoS*
942 *Genet*. 9:e1003242.
943 Servant N, Varoquaux N, Lajoie BR, Viara E, Chen CJ, Vert JP, Heard E, Dekker J, Barillot
944 E. 2015. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome*
945 *Biol*. 16:259.
946 Sisu C, Pei B, Leng J, Frankish A, Zhang Y, Balasubramanian S, Harte R, Wang D, Rutenberg-
947 Schoenberg M, Clark W, et al. 2014. Comparative analysis of pseudogenes across three
948 phyla. *Proc Natl Acad Sci U S A*. 111:13361-13366.
949 Spector DL, Lamond AI. 2011. Nuclear speckles. *Cold Spring Harb Perspect Biol*. 3: a000646.
950 Su JH, Zheng P, Kinrot SS, Bintu B, Zhuang X. 2020. Genome-scale imaging of the 3D
951 organization and transcriptional activity of chromatin. *Cell*. 182:1641-1659.
952 Sultana T, van Essen D, Siol O, Bailly-Bechet M, Philippe C, Zine El Aabidine A, Pioger L,
953 Nigumann P, Saccani S, Andrau JC, et al. 2019. The landscape of L1 retrotransposons
954 in the human genome is shaped by pre-insertion sequence biases and post-insertion
955 selection. *Mol Cell*. 74:555-570 e557.
956 Tan S, Cardoso-Moreira M, Shi W, Zhang D, Huang J, Mao Y, Jia H, Zhang Y, Chen C, Shao
957 Y, et al. 2016. LTR-mediated retroposition as a mechanism of RNA-based duplication
958 in metazoans. *Genome Res*. 26:1663-1675.
959 Takei Y, Yang Y, White J, Yun J, Prasad M, Ombelets L, Schindler S, Cai L. 2023. High-
960 resolution spatial multi-omics reveals cell-type specific nuclear compartments. *bioRxiv*.
961 https://doi.org/10.1101/2023.05.07.539762
962 Troskie RL, Jafrani Y, Mercer TR, Ewing AD, Faulkner GJ, Cheetham SW. 2021. Long-read
963 cDNA sequencing identifies functional pseudogenes in the human transcriptome.
964 *Genome Biol*. 22:146.
965 van Steensel B, Furlong EEM. 2019. The role of transcription in shaping the spatial
966 organization of the genome. *Nat Rev Mol Cell Biol*. 20:327-337.
967 Vinckenbosch N, Dupanloup I, Kaessmann H. 2006. Evolutionary fate of retroposed gene
968 copies in the human genome. *Proc Natl Acad Sci U S A*. 103:3220-3225.
969 Vouzas AE, Gilbert DM. 2021. Mammalian DNA replication timing. *Cold Spring Harb*
970 *Perspect Biol*. 13.

971 Wang H, Yang L, Wang Y, Chen L, Li H, Xie Z. 2019. RPFdb v2.0: an updated database for
972 genome-wide information of translated mRNA generated from ribosome profiling.
973 *Nucleic Acids Res*. 47:D230-D234.
974 Wang Y, Zhang Y, Zhang R, van Schaik T, Zhang L, Sasaki T, Peric-Hupkes D, Chen Y,
975 Gilbert DM, van Steensel B, et al. 2021. SPIN reveals genome-wide landscape of
976 nuclear compartmentalization. *Genome Biol*. 22:36.
977 Wolff J, Rabbani L, Gilsbach R, Richard G, Manke T, Backofen R, Gruning BA. 2020. Galaxy
978 HiCExplorer 3: a web server for reproducible Hi-C, capture Hi-C and single-cell Hi-C
979 data analysis, quality control and visualization. *Nucleic Acids Res*. 48:W177-W184.
980 Xie T, Yang QY, Wang XT, McLysaght A, Zhang HY. 2016. Spatial colocalization of human
981 ohnolog pairs acts to maintain dosage-balance. *Mol Biol Evol*. 33:2368-2375.
982 Xiong K, Ma J. 2019. Revealing Hi-C subcompartments by imputing inter-chromosomal
983 chromatin interactions. *Nat Commun*. 10:5069.
984 Zhang J. 2003. Evolution by gene duplication: an update. *Trends Ecol Evol*. 18:292-298.
985 Zhang W, Tautz D. 2022. Tracing the origin and evolutionary fate of recent gene retrocopies
986 in natural populations of the house mouse. *Mol Biol Evol*. 39:msab360.
987 Zhang W, Xie C, Ullrich K, Zhang YE, Tautz D. 2021. The mutational load in natural
988 populations is significantly affected by high primary rates of retroposition. *Proc Natl*
989 *Acad Sci U S A*. 118: e2013043118.
990 Zhang Y, Li S, Abyzov A, Gerstein MB. 2017. Landscape and variation of novel
991 retroduplications in 26 human populations. *PLoS Comput Biol*. 13:e1005567.
992 Zhang Y, Li T, Preissl S, Amaral ML, Grinstein JD, Farah EN, Destici E, Qiu Y, Hu R, Lee
993 AY, et al. 2019. Transcriptionally active HERV-H retrotransposons demarcate
994 topologically associating domains in human pluripotent stem cells. *Nat Genet*. 51:1380-
995 1388.
996 Zhang Z, Harrison PM, Liu Y, Gerstein M. 2003. Millions of years of evolution preserved: a
997 comprehensive catalog of the processed pseudogenes in the human genome. *Genome*
998 *Res*. 13:2541-2558.
999

1000 ## Figure Legends

1001 **Figure 1.** Spatial colocalization between retrocopies and their parental genes. (*A*) Circos plot
1002 highlighting the chromatin interactions between retrocopies and parental genes in the
1003 GM12878 cell line. From outer to inner are the ideogram of human chromosomes, the density
1004 of parental genes, the density of retrocopies, and significant chromatin interactions between
1005 retrocopies and parental genes. (*B*) Bar plot showing the number of human interchromosomal
1006 retrocopy-parent pairs as a function of the number of cell lines wherein the spatial
1007 colocalization is detected. (*C*) Distribution of colocalization frequency between chromatin
1008 pairs under different null models, while the red arrow denotes the colocalization frequency of
1009 true retrocopy-parent pairs. Simulations and true frequency are based on Hi-C data of
1010 GM12878. *P*-value < 0.001 in all cases.

1011

1012 **Figure 2.** Chromatin positioning of retrocopies and parental genes in the 3D genome. (*A*)
1013 Enrichments of retrocopies, parental genes, unprocessed pseudogenes, and retroCNVs in five
1014 primary subcompartments and TAD in the GM12878 cell line. (*B*) Sankey diagram showing
1015 the connection of colocalized (left panel) and noncolocalized (right panel) retrocopy-parent
1016 pairs in the context of subcompartment annotations in GM12878. The height of bar indicates
1017 the proportion of loci that fall in respective categories. Lines connect retrocopies and their
1018 corresponding parent genes. NA denotes a fragment is not annotated as any of the five
1019 subcompartments. (*C*) Comparison of SON TSA-seq signal between colocalized and
1020 noncolocalized retrocopies/parental genes in K562 ($^{***}p < 0.001$; two tailed Wilcoxon test).
1021 (*D*) Proportion of colocalized and noncolocalized retrocopies/parental genes that overlapped
1022 with different SON TSA-seq deciles in K562. Decile 1 (estimated distance: 0.868-1.020 μm)
1023 and decile 10 (estimated distance: 0.054-0.399 μm) represent the top 10% of chromatin that
1024 are farthest and closest to SON, respectively, and other deciles are those that fall between.

1025 Coloc_retro: colocalized retrocopies; noncoloc_retro: noncolocalized retrocopies;
1026 coloc_parent: colocalized parental genes; noncoloc_parent: noncolocalized parental genes.
1027
1028 **Figure 3.** Expression and regulatory element sharing of retrocopies and parental genes. (*A*)
1029 Plots of cumulative fraction of expression of retrocopies (left panels) and parental genes (right
1030 panels) in GM12878. The expression levels of polyadenylated (ployA+; top panels) and non-
1031 polyadenylated (ployA-; bottom panels) RNAs are plotted separately. (*B*) Comparison of GRO-
1032 seq signal (left panel) and Pol II TSA-seq signal (right panel) between colocalized and
1033 noncolocalized retrocopies/parental genes in K562. (*C*) Comparison of the number of shared
1034 TFBSs between colocalized and noncolocalized retrocopy-parent pairs. $^{***}p < 0.001$; two tailed
1035 Wilcoxon test.
1036
1037 **Figure 4.** Epigenetic signals for DNase I hypersensitivity, histone variant H2A.Z, and 10
1038 histone modifications along colocalized retrocopies (red lines), noncolocalized retrocopies
1039 (blue lines), and unprocessed pseudogenes (gray lines) and their flanking regions (± 3 kb) in
1040 the GM12878 cell line. Solid lines represent the mean fold change signal, shadowed areas
1041 denote the standard error. TSS: putative transcription start site; TES: putative transcription end
1042 site.
1043
1044 **Figure 5.** Coding potential and functional importance of colocalized and noncolocalized
1045 retrocopies. (*A*-*B*) Comparison of ribo-seq RPKM values (*A*) and phastCons scores (*B*) among
1046 retrocopies that are colocalized in 0, 1, and ≥ 2 cell lines. For ribo-seq data, only retrocopies
1047 with RPKM ≥ 1 were included. (*C*) Percentage of colocalized and noncolocalized retrocopies
1048 in the seven cell lines that are overlapped with at least one NHGRI-EBI GWAS SNPs. (*D*)
1049 Correlation between phastCons score and RNA expression level (left panel) and between

1050 phastCons score and ribo-seq signal (right panel) of retrocopies. Solid line and shadowed area
1051 in each plot denote the linear regression line and its confident interval. $^{***}p < 0.001$, $^{**}p < 0.01$,
1052 $^{*}p < 0.05$, NS. not significant; two tailed Wilcoxon test in A and B, Fisher's exact test in C.
1053
1054 **Figure 6.** Colocalization and evolution of retroCNVs in humans. (*A*) Density plots showing
1055 the distribution of colocalization frequency between chromatin pairs under different null
1056 models for singleton retroCNVs (left panel) and high-frequency retroCNVs (right panel). Red
1057 arrows denote the colocalization frequency between true retroCNVs and their parental genes.
1058 Simulations and true frequencies are based on Hi-C data in GM12878. *P*-values are indicated
1059 for simulation 5. (*B*) The presence frequency spectra for retroCNVs that are colocalized in 0,
1060 1, and ≥ 2 cell lines.
1061

Figure 1.

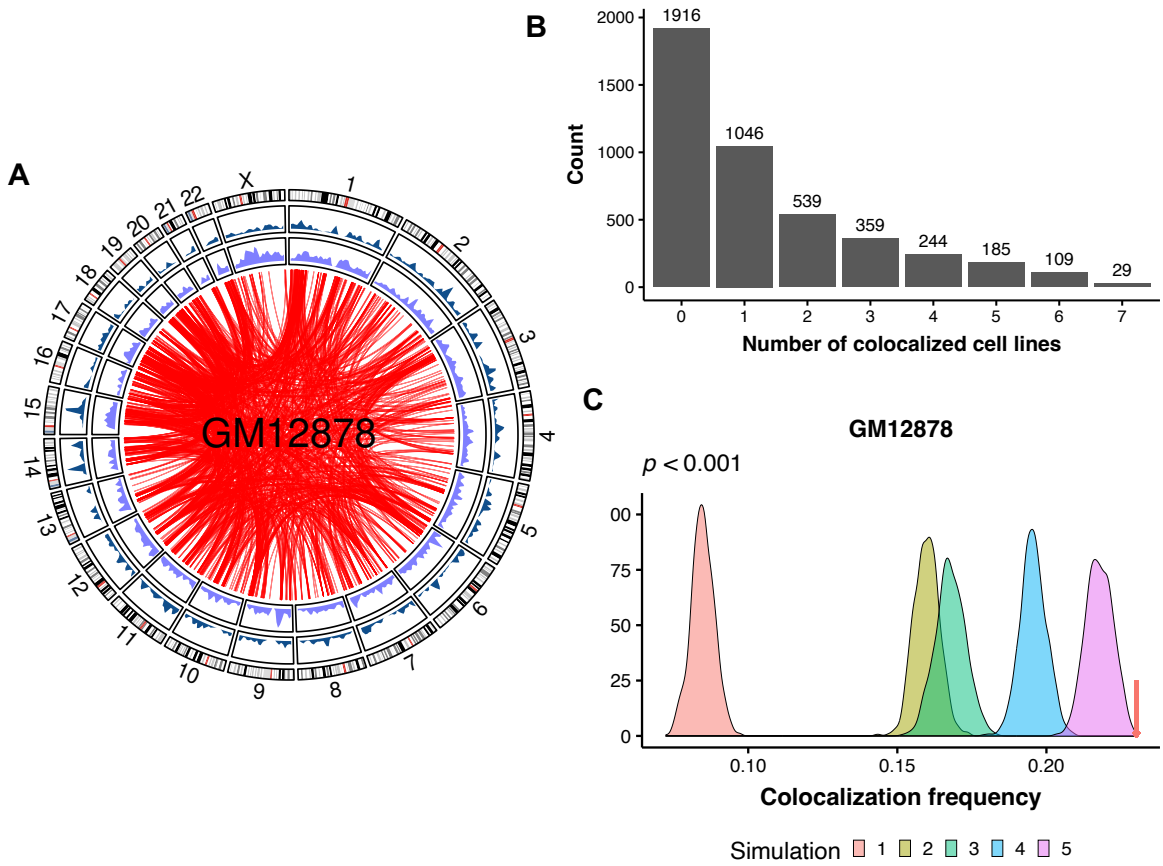

A
GM12878
B
Count
Number of colocalized cell lines
1916
1046
539
359
244
185
109
29
C
GM12878
p < 0.001
Colocalization frequency
Simulation 1 2 3 4 5

Figure 2.

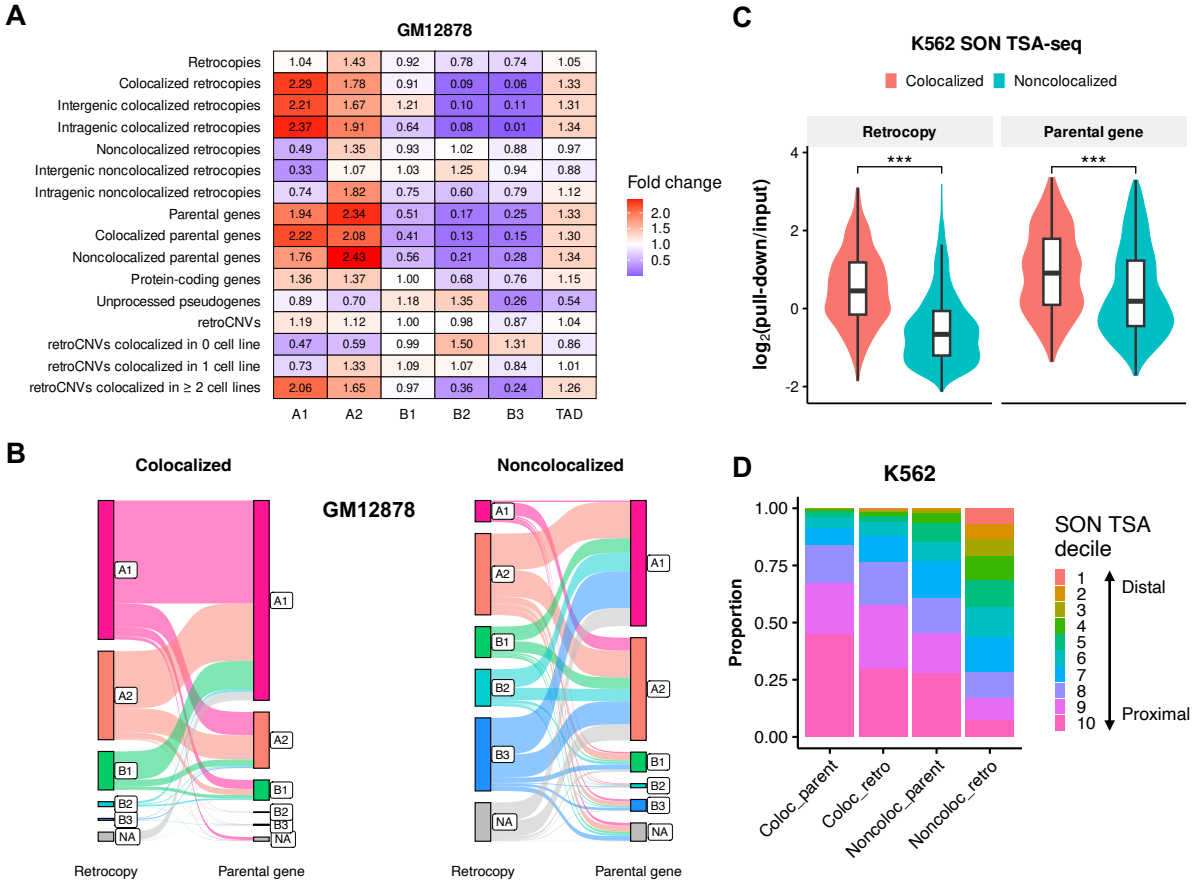

A
GM12878
Retrocopies 1.04 1.43 0.92 0.78 0.74 1.05
Colocalized retrocopies 2.29 1.78 0.91 0.09 0.06 1.33
Intergenic colocalized retrocopies 2.21 1.67 1.21 0.10 0.11 1.31
Intragenic colocalized retrocopies 2.37 1.91 0.64 0.08 0.01 1.34
Noncolocalized retrocopies 0.49 1.35 0.93 1.02 0.88 0.97
Intergenic 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 1.09 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
***
***
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.

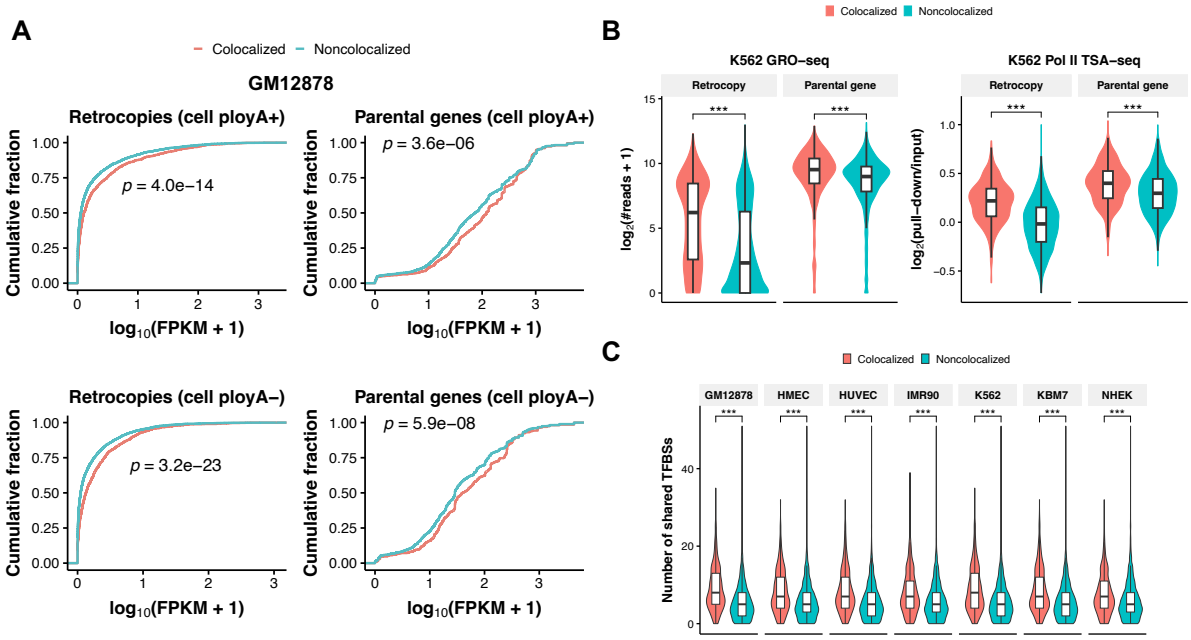

A
Colocalized
Noncolocalized
GM12878
Retrocopies (cell ployA+)
Parental genes (cell ployA+)
Retrocopies (cell ployA−)
Parental genes (cell ployA−)
Cumulative fraction
log10(FPKM + 1)
p = 4.0e−14
p = 3.6e−06
p = 3.2e−23
p = 5.9e−08
B
K562 GRO−seq
K562 Pol II TSA−seq
Retrocopy
Parental gene
log2(#reads + 1)
log2(pull−down/input)
C
GM12878
HMEC
HUVEC
IMR90
K562
KBM7
NHEK
Number of shared TFBSs

Figure 4.

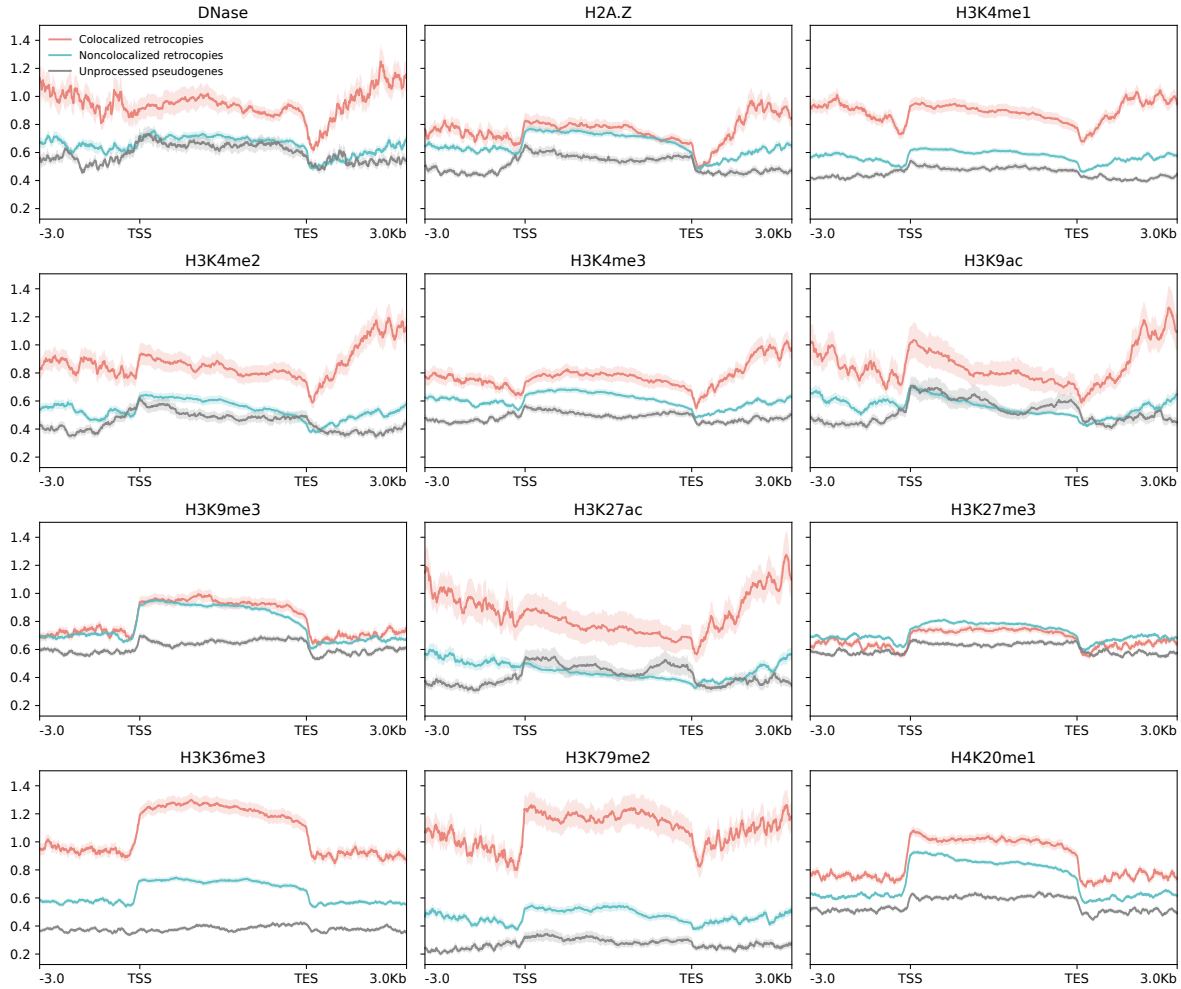

DNase
H2A.Z
H3K4me1
H3K4me2
H3K4me3
H3K9ac
H3K9me3
H3K27ac
H3K27me3
H3K36me3
H3K79me2
H4K20me1
Colocalized retrocopies
Noncolocalized retrocopies
Unprocessed pseudogenes
1.4
1.2
1.0
0.8
0.6
0.4
0.2
-3.0
TSS
TES
3.0Kb

Figure 5.

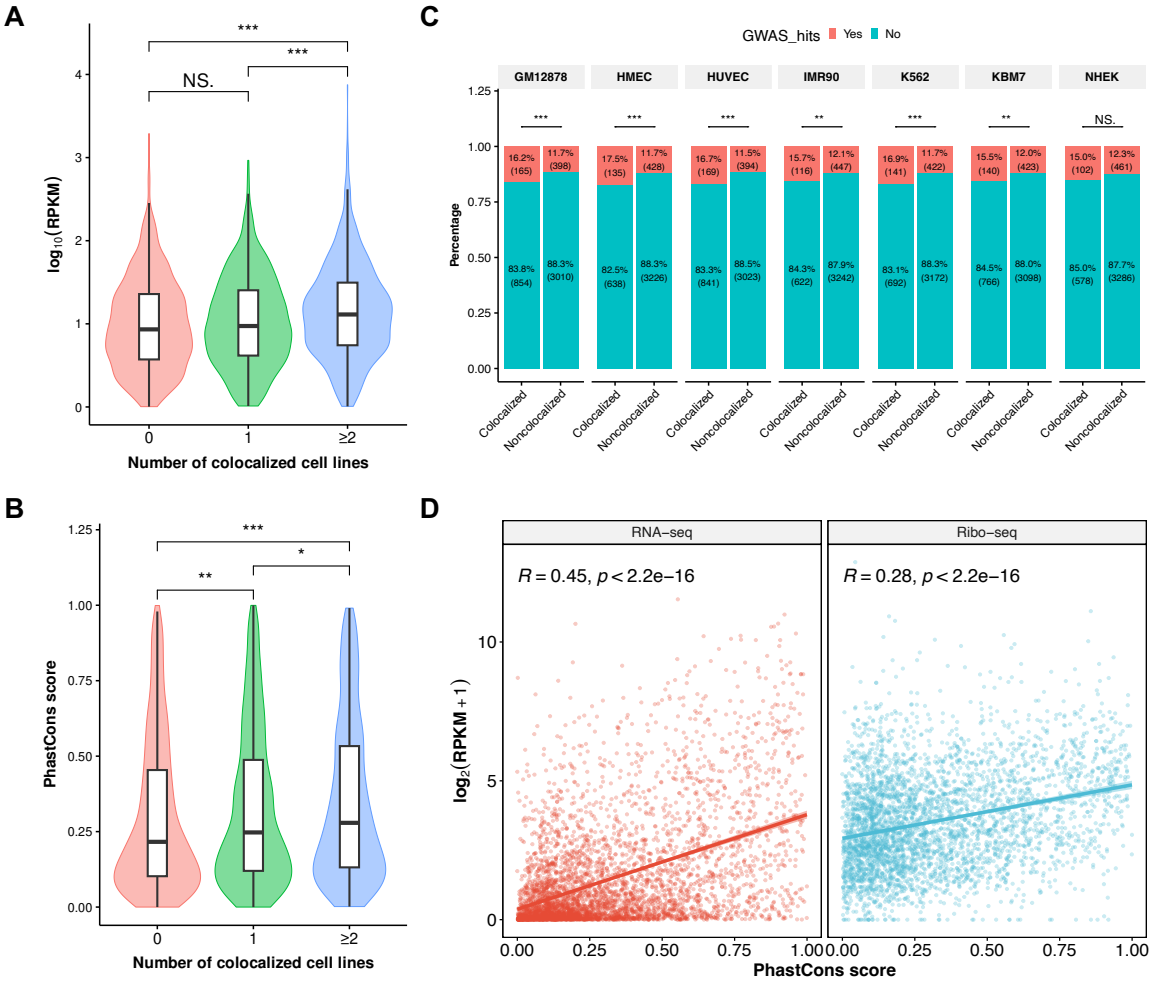
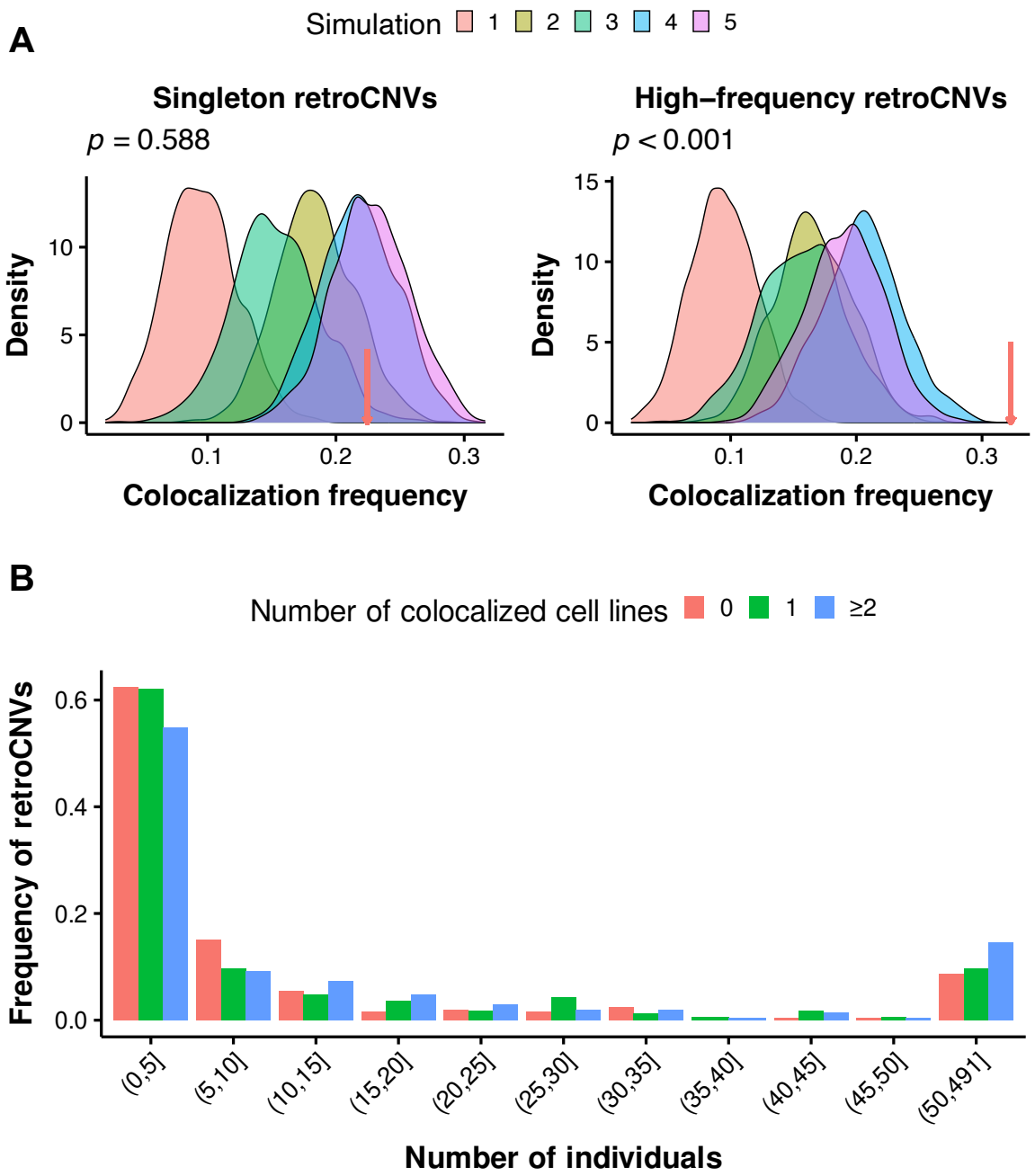

A
***
***
NS.
log10(RPKM)
Number of colocalized cell lines
0
1
≥2
B
***
*
**
PhastCons score
Number of colocalized cell lines
C
GWAS_hits Yes No
GM12878
HMEC
HUVEC
IMR90
K562
KBM7
NHEK
Percentage
Colocalized
Noncolocalized
D
RNA−seq
Ribo−seq
R = 0.45, p < 2.2e−16
R = 0.28, p < 2.2e−16
log2(RPKM + 1)
PhastCons score

Figure 6.


A
Simulation 1 2 3 4 5
Singleton retroCNVs
p = 0.588
High−frequency retroCNVs
p < 0.001
Density
Colocalization frequency
B
Number of colocalized cell lines 0 1 ≥2
Frequency of retroCNVs
(0,5]
(5,10]
(10,15]
(15,20]
(20,25]
(25,30]
(30,35]
(35,40]
(40,45]
(45,50]
(50,491]
Number of individuals