Amplification-free Library Preparation Improves Quality of Hi-C Analysis

Longjian Niu^{1,4,5}, Wei Shen^{2,3,5}, Yingzhang Huang^{1,5}, Na He¹, Yuedong Zhang¹, Jialei Sun¹, Jing Wan^{2,3}, Daxin Jiang¹, Manyun Yang¹, Yu Chung Tse¹, Li Li^{2,3,*}, Chunhui Hou^{1,6,*}

¹Department of Biology, Southern University of Science and Technology, Shenzhen 518055, China

²Department of Bioinformatics, Huazhong Agricultural University, Wuhan 430070, China

³Hubei Key Laboratory of Agricultural Bioinformatics, Huazhong Agricultural University, Wuhan 430070, China

⁴Department of Biology, Nankai University, Tianjin 300071, China

⁵These authors contributed equally to this work

⁶Leading correspondence

*Correspondence: houch@sustc.edu.cn; li.li@mail.hzau.edu.cn

Abstract

PCR amplification of Hi-C libraries introduces unusable duplicates and results in a biased representation of chromatin interactions. We present a simplified, fast, and economically efficient Hi-C library preparation procedure that generates sufficient non-amplified ligation products for deep sequencing from 30 million *Drosophila* cells. Comprehensive analysis of the resulting data indicates that amplification-free Hi-C preserves higher complexity of chromatin interaction and lowers sequencing depth dramatically for the same number of unique paired reads. For human cells which has a large genome, this method recovers an amount of ligated fragments enough for direct high-throughput sequencing without amplification on as low as 250 thousand of cells. Comparison with published *in situ* Hi-C on millions of human cells reveals that amplification introduces distance-dependent amplification bias, which results in increasing background noise level against genomic distance. With amplification bias avoided, our method may produce a chromatin interaction network more faithfully reflecting the real three-dimensional genomic architecture.

45 Introduction

46

47 Hi-C is a powerful tool for mapping interaction frequencies between
 48 chromatin fragments in a genome-wide and quantitative manner¹. It compares
 49 the number of ligation events between each fragment pair in a large population
 50 of cells and thereby allows the identification of various genome structural
 51 features including compartments, topologically associating domains (TADs),
 52 and loops¹⁻⁵. The main prerequisite for high-quality Hi-C analysis is the
 53 accurate quantification of chromatin interaction frequencies. The amount of
 54 DNA obtained in a typical Hi-C experiment is assumed to be insufficient for
 55 direct high-throughput sequencing. Thus, PCR amplification is a default step in
 56 Hi-C-related experiments⁶ to guarantee sequencing primer addition and to
 57 produce a sufficient amount of DNA for sequencing, especially for Hi-C
 58 experiments involving single or low cell numbers⁷⁻¹⁸.

59

60 The 3D nature of Hi-C deems sequencing depth and library complexity are
 61 two critical variables in evaluating the achievable resolution from Hi-C
 62 experiments, given a range of fragment sizes determined by the choice of
 63 restriction enzyme. Currently, many biological replicates and multiple rounds of
 64 PCR amplifications are required for high resolution genome architecture
 65 analysis in order to generate sufficient DNA with high enough complexity to
 66 represent the global chromatin interaction diversity within a cell population.

67 Though universal primers are used, PCR amplification introduces duplicates
68 and may skew Hi-C library composition, which may not be fully corrected by
69 normalization methods¹⁹⁻²¹.

70

71 However, whether PCR amplification is really necessary and to what
72 extent it changes the composition of Hi-C libraries has not been systematically
73 examined. Here we present SAFE Hi-C, a simplified, amplification-free, and
74 economically efficient process, in which paired reads generated by
75 independent ligation events are saved. We tested this method on 30 million
76 *Drosophila* S2 and 250 thousand human K562 cells. Comparison to traditional
77 *in situ* Hi-C reveals amplification introduces distance-dependent bias in
78 chromatin interaction frequency, which lowers the ratio between intra-TAD and
79 inter-TAD interaction frequency especially for large genomes such as human
80 cells. Efficient recovery of enough ligated fragments for direct high-throughput
81 sequencing is important for accurate depicting the three dimensional genome
82 architecture and the understanding of its functional role in transcription
83 regulation, replication, genome stability and other critical biological activities
84 related to chromatin.

85

86 Results

87

88 SAFE Hi-C Library Preparation and Sequencing

89 To determine how many biotin-labelled ligation events can be captured by
 90 streptavidin-conjugated beads, we carried out two SAFE Hi-C experiments as
 91 biological replicates on 30 million *Drosophila* S2 cells using DpnII (Fig. 1). We
 92 stripped off Hi-C ligation products from beads after addition of sequencing
 93 primers²² (Fig. 1b and Supplementary Table 1). Qubit quantification showed
 94 each SAFE Hi-C experiment recovered around 100 ng of ligated DNA, which
 95 roughly equals to 50 µl of 12 nM single strand DNA with a size around 500
 96 bases, enough for at least 10 lanes of sequencing on the Illumina HiSeq X10
 97 platform. We also conducted traditional *in situ* Hi-C experiments in replicates
 98 for comparison and amplified Hi-C ligation products from diluted beads using
 99 different numbers of PCR amplification cycles (4, 8, 12, 16 and 20) to produce
 100 similar amounts of DNA as in the SAFE Hi-C experiments (Supplementary
 101 Table 1).

102

103 All libraries were sequenced on an Illumina HiSeq X10 instrument and
 104 aligned to the reference genome using bowtie 2.0. The length of sequenced
 105 fragments in all Hi-C libraries ranged from 200-750 bp, and peaked around
 106 370 bp (Supplementary Fig. 1), consistent with the fact that the median length
 107 of DpnII fragments is 194 bp. Global chromatin interaction frequencies were
 108 highly correlated between biological replicates and between different pairs of
 109 libraries, with the lowest Pearson's correlation value of 0.994 (Supplementary
 110 Fig. 2). We combined biological replicates and obtained 338, 246, 220, 232,

111 238 and 248 million aligned paired reads (Supplementary Table 2). Datasets
112 were normalized as described^{5,23} for further analysis. The ratio of cis- and
113 trans-unique paired reads, generally considered as a proxy indicator of the
114 quality of a Hi-C library, was 13.5 for SAFE Hi-C and approximately the same
115 for amplified libraries (Fig. 2a). After PCR amplification, fragment pairs with
116 lower (<~42%) or higher (>~42%) GC content were significantly under- or
117 over-represented compared to SAFE Hi-C, respectively ($p < 10^{-40}$,
118 Mann–Whitney U test, Supplementary Fig. 3).

119
120 After successfully applying SAFE Hi-C on 30 million *Drosophila* cells
121 whose chromatin content is roughly equal to 1 million human cells, next we
122 determined the lowest number of human cells needed for SAFE Hi-C. 250
123 thousand and 100 thousand human K562 cells were tested. SAFE Hi-C
124 worked only with 250 thousand cells from which we recovered 15 μ l of 4 nM
125 single strand DNA with a size around 575 bases. This is about 1/10 of the DNA
126 recovered from 30 million *Drosophila* S2 cells. We think this may be due to the
127 higher chance of DNA loss frequently experienced with experiment using small
128 amount of starting material. One quarter of a lane on the Illumina HiSeq X10
129 was used for sequencing the human K562 SAFE Hi-C library generating 121
130 million of paired reads, out of which 106 million (87.4%) aligned successfully to
131 the human reference genome (Supplementary Table 3). A similar number of
132 chromatin interactions from *in situ* Hi-C on K562 cells previously published by

133 Rao et al. was downloaded and used for comparison.

134

135 **SAFE Hi-C Avoids Removal of PCR Duplicates**

136 PCR amplification introduces extra duplicates to the Hi-C library, which
 137 lowers the percentage of unique paired reads. For SAFE Hi-C, we define
 138 “suspected PCR duplicates” as those sharing identical sequences at both ends.
 139 These duplicates are most likely derived from ligation events of fragment pairs
 140 with same sequences in millions of cells, which could happen at a high rate
 141 especially when the size of most fragments is short. PCR duplicates can also
 142 be generated from DNA cluster generation on the Illumina sequencing
 143 machine and are considered as “optical duplicates”. This accounts for less
 144 than 1% of the total sequenced paired reads (Supplementary Table 2 and 3)
 145 and were excluded in further analysis.

146 About 8% of total mapped paired reads were “suspected PCR duplicates”
 147 for *Drosophila* S2 SAFE Hi-C libraries (Fig. 2b, light blue bar) and were kept for
 148 Hi-C analysis because no amplification was involved. However, for amplified
 149 libraries, duplicates from independent ligations cannot be distinguished from
 150 those introduced by PCR amplification, thus all were considered arising from
 151 single ligation events. As expected, the proportion of PCR duplicates positively
 152 correlates with the number of amplification cycles (Fig. 2b, red bar). The
 153 percentage of duplicates increased to 14%, 15%, 21%, 49% and 64% after 4,
 154 8, 12, 16 and 20 cycles of amplification, respectively (Fig. 2b, red bar).

Correspondingly, the percentage of non-duplicate paired reads decreased dramatically as amplification cycles increased, especially after 16 and 20 cycles (Fig. 2b, dark blue bar). We also calculated duplicates and valid paired read percentages for all mappable paired reads (Supplementary Fig. 4). PCR duplicate depth analysis shows that most duplicates have two copies for all libraries (Fig. 2c), and the percentage of amplified ligation products of higher duplicate depth increased considerably after 16 and 20 cycles of amplification (Fig. 2c).

163

SAFE Hi-C Increases Library Complexity

For *Drosophila* S2 SAFE Hi-C libraries, the percentage of unique pair reads correlates almost linearly with sequencing depth (Fig. 2d), suggesting that the library complexity was far from exhausted at current sequencing depth.

After amplification, PCR duplicates cannot be kept and their percentage increases dramatically as the cycle of amplification increases (Fig. 2b).

Consistently, the estimated library complexity dramatically dropped from 1.5 billion for SAFE Hi-C libraries to 0.58 billion for *in situ* Hi-C libraries after just 4 cycles of amplification (Fig. 2e).

173

Amplification Bias Is Genomic Distance-Dependent

Chromatin interaction frequency is inversely correlated with genomic distance. Amplification resulted in a moderate change on the decaying pattern

177 of chromatin interactions at any genomic distance for traditional *in situ* Hi-Cs
 178 on *Drosophila* S2 (Fig. 3a, b). However, for K562, the changes were obvious
 179 (4a, b). After normalization against SAFE Hi-C, we found that the relative
 180 chromatin interaction frequency starts lower (~0.9) at 1kb and becomes higher
 181 beyond 3 kb for most amplified libraries in *Drosophila* S2 cells, except for
 182 amplification of 20 cycles (Fig. 3b), suggesting short distance ligations are
 183 generally underrepresented after amplification even after limited number of
 184 amplification cycles.

185 In comparison to *Drosophila* S2, amplification introduced more obvious
 186 biases for the human genome. Compared to published *in situ* Hi-C on human
 187 K562 cells, chromatin interaction frequency from SAFE Hi-C decays at a rate
 188 much closer to the predicted fractal globular model (s^{-1} , Fig. 4a), while *in situ*
 189 Hi-C decays at a rate closer to $s^{-0.5}$ within the genomic distance of 1 Mb (Fig.
 190 4a). When normalized against SAFE Hi-C, the relative chromatin interaction
 191 frequency from *in situ* Hi-C starts even lower (~0.5) at 10kb and continuously
 192 increases and crosses with SAFE Hi-C at a genomic distance of 140kb (Fig.
 193 4b).

194 This comparison reveals an unexpected bias in Hi-C library amplification,
 195 which could be due to the competition of complementary fragment
 196 hybridization and primer annealing. It is reasonable to expect that pairs of
 197 ligated fragments of high concentration tend to hybridize to themselves instead
 198 of primers. The increasing bias over distance also suggests that background

199 noise could be more significant over a longer distance.

200

201 **SAFE Hi-C Maintains a Higher Contrast Ratio of the Intra-TAD and** 202 **Inter-TADs Chromatin Interaction Frequency in the Human Genome**

203 So far, all characterizations of TADs, compartments and loops have been
204 based on amplified Hi-Cs. However, the amplification effects on such analysis
205 have not been evaluated experimentally. With the development of SAFE Hi-C,
206 we are able to show if and to what extent amplification affects TADs,
207 compartments and loops analysis.

208 We first plotted heatmaps for *Drosophila* S2 cells with normalized datasets
209 and identified TADs at 5kb resolution (Fig. 3c and Supplementary Fig. 5).
210 Overall, we observed minimal variation in border strength across the genome
211 after amplification (Fig. 3c). Consistently, the number of identified TADs was
212 similar and most of them were conserved (Fig. 3d). These results together
213 suggest at least SAFE Hi-C is as reliable as traditional in situ Hi-C for TAD
214 characterization.

215 We further calculated the interaction frequency vs distance within TADs
216 (Fig. 3e). Similar to a global decaying pattern, normalization against SAFE
217 Hi-C within TADs showed that chromatin interaction frequency was
218 underrepresented within and overrepresented beyond 3 kb for most amplified
219 libraries (Fig. 3f). Recently, sub-kb resolution Hi-C identified 4,123 TADs in the
220 *Drosophila* genome, with TADs as small as 3 kb²⁴. Our SAFE Hi-C results also

221 indicate that high frequency of interactions in the 3 kb range may be an
222 important feature of the *Drosophila* genome.

223 Next, we identified TADs for SAFE Hi-C and *in situ* Hi-C on human K562
224 cells. Most of the TADs identified overlapped (2946), however, we found many
225 more specific TADs for SAFE Hi-C (1668) as compared with *in situ* Hi-C (556)
226 (Fig. 4c). Visual inspection shows a very similar pattern of border index for
227 both experiments (Fig. 4d). However, for *in situ* Hi-C border index peaks are
228 frequently less sharp (arrowhead, Fig. 4d) and the shape of possible TADs is
229 generally less clean compared with SAFE Hi-C (heatmap, Fig. 4d). The border
230 index values are lower for *in situ* Hi-C as well (Fig. 4d). We aligned the borders
231 of all identified TADs, shared TADs and specific TADs for SAFE Hi-C and *in*
232 *situ* Hi-C and calculated the ratio of chromatin interaction density of intra-TAD
233 over inter-TAD (Fig. 4e). The ratios of SAFE Hi-C (15.55, 15.07 and 15.58) are
234 consistently higher than the values for *in situ* Hi-C (9.26, 8.49 and 8.81)
235 (bottom left, Fig. 4e). These results together show that amplification weakens
236 the contrast between TADs and inter-TAD regions, which could be caused by
237 biased amplification over genomic distance as shown in Fig. 4a and 4b, and
238 thus underlines the importance of avoiding amplification in order to further
239 improve the quality of analysis.

240

241 **Amplification Effects on Compartment Analysis and Chromatin Loop**
242 **Identification**

243 We characterized compartments for each chromosome arm of the
244 *Drosophila* genome (Supplementary Fig. 6). The span of absolute eigenvalues
245 decreased for amplified Hi-C (Supplementary Fig. 7), suggesting SAFE Hi-C
246 may somehow improve compartments analysis. For human K562 cells, the
247 eigenvalues correlated highly for SAFE Hi-C and *in situ* Hi-C (Supplementary
248 Fig. 8). Together, these results show that the effect of amplification on
249 compartment analysis is less obvious than for TADs identification.

250 Finally, we identified long-range chromatin interactions at 5 kb resolution
251 for the *Drosophila* genome (q value < 0.1) but not for human genome because
252 the resolution is too low for meaningful and reliable chromatin loop
253 identification. Signal-to-noise ratio was calculated as described⁵. The number
254 of identified loops negatively correlated with the number of amplification cycles
255 (Supplementary Fig. 9 and 10a). Approximately 43% of loops from each
256 amplified Hi-C library (Supplementary Fig. 10a, dark blue bar) overlapped with
257 those identified using SAFE Hi-C. The aggregate P2LL values and ZscoreLL
258 for all loops identified for both SAFE Hi-C and amplified libraries changed little
259 (Supplementary Fig. 10b, c). For shared loops, the P2LL values rose as the
260 number of PCR cycles increased (Supplementary Fig. 10d, upper panel),
261 suggesting amplification favored these interactions. However, for loops lost
262 after amplification, the P2LL values decreased (Supplementary Fig. 10d, lower
263 panel), suggesting under-amplification of these interactions.

264

265 Discussion

266

267 In addition to the introduction of PCR duplicates and a significant reduction
268 in library complexity, our results show that amplification has a different effect
269 on the three-dimensional genome architecture analysis for a small (*Drosophila*)
270 and big (human) genomes. For small genomes like *Drosophila*, amplification
271 affects the characterization of loops and to lesser extent compartments, but
272 shows little effect on accurate TAD identification when the same number of raw
273 sequencing paired reads were used.

274 Though SAFE Hi-C was initially tested on *Drosophila* genome which is
275 much smaller than typical mammalian genomes, it is also suitable for large
276 genomes due to its ability to maintain the original complexity and integrity of
277 chromatin interaction diversity, effectively lowering sequencing depth and
278 saving labor and cost. We estimate that 1-2 million mammalian cells digested
279 with DpnII will generate the same amount of Hi-C ligation products as 30
280 million *Drosophila* cells, which will be enough for at least ten lanes of
281 sequencing on the Illumina X10 platform. Theoretically, the complexity of a
282 single SAFE Hi-C library will be in the hundreds of billions if the number of
283 single molecular DNA ligation products was calculated in a library made from
284 1-2 million mouse or human cells.

285 We tested the low limit of human cell number for SAFE Hi-C. With 250
286 thousand of K562 cells, we successfully recovered ligates enough for at least

287 one lane of sequencing on the Illumina X10 platform. With a lower number of
288 cells, the chance of DNA loss increased during library preparation. Our data
289 analysis reveals that avoiding amplification significantly improves TAD
290 identification for the human genome. It is reasonable to conclude that the
291 genomic distance dependent bias introduced by amplification may very likely
292 compromise the quality of TAD analysis.

293 In sum, by avoiding amplification, SAFE Hi-C can be used to improve the
294 quality of Hi-C analysis as well as to save time, reagents and reduce cost. In
295 cases when the available cell number is too low, SAFE Hi-C is also compatible
296 with PCR amplification to ensure enough sequencing material can be
297 generated. Finally, other Hi-C methods using enzymes like DNase or MNase
298 may also be simplified and be amplification-free.

299

300

301 **Methods**

302

303 **Cell culture**

304 S2 cells were cultured in Schneider's Medium (Gibco, 21720024)
305 supplemented with 10% heat inactivated FBS (Sigma, F7524) and 1%
306 Penicillin/Streptomycin P/S (Sigma, P0781) at 27°C.

307 K562 cells were incubated in 1× RPMI1640 media supplemented with 10%
308 FBS at 37°C with 5% CO₂.

309

310 ***In situ* Hi-C**

311 *In situ* Hi-C was carried out as described. Cells were crosslinked with 1%
 312 formaldehyde then lysed to collect nuclei. Pelleted nuclei were digested with
 313 DpnII restriction enzyme (NEB, R0147). The restriction fragment overhangs
 314 were filled and marked with biotin-labelled dATP (Thermo Fisher, 19524016)
 315 and dCTP, dTTP and dGTP before ligation. DNA was reverse crosslinked,
 316 purified and fragmented by sonication on a Covaris sonicator. Biotin labelled
 317 DNA was pulled down on Streptavidin Dynabeads (NEB, S1420S). After DNA
 318 repair and 3' A addition, SHORT Y-Adaptor (Supplementary Table 1) was
 319 added. Diluted DNA on Dynabeads was used for PCR amplification (4, 8,
 320 12,16 and 20 cycles) to produce similar amounts of DNA for sequencing on the
 321 Illumina HiSeq X10 platform (PE 2x150 bp reads).

322

323 **SAFE Hi-C**

324 SAFE Hi-C is a modification of *in situ* Hi-C^{5,25}. Cells were crosslinked with 1%
 325 formaldehyde for 10 min at room temperature (RT). The reaction was stopped
 326 by adding 1/10 volume of 2.5 M glycine. Up to 30 million crosslinked cells were
 327 resuspended in 500 µL of ice-cold Hi-C lysis buffer and rotated at 4 °C for 30
 328 min. Nuclei were pelleted at 4 °C for 5 min at 2,500 relative centrifugal force
 329 (RCF), and the supernatant was discarded. Pelleted nuclei were washed once
 330 with 500 µL of ice-cold Hi-C lysis buffer. The supernatant was removed again,

331 and the pellet was resuspended in 100 µL of 0.5% SDS and incubated at 62°C
332 for 10 min with no shaking or rotation. 285 µL of water and 50 µL of 10% Triton
333 X-100 were added, and samples were rotated at 37°C for 15 min to quench the
334 SDS. 50 µL of NEB Buffer 3.1 and 20 µL of 10 U/µL DpnII restriction enzyme
335 (NEB, R0147) were then added, and the sample was rotated at 37°C for 4 h.
336 DpnII was then heat inactivated at 62°C for 20 min with no shaking or rotation.
337 To fill-in the restriction fragment overhangs and mark the DNA ends with biotin,
338 52 µL of incorporation master mix was then added: 37.5 µL of 0.4 mM
339 biotin–dATP (Thermo Fisher, 19524016); 4.5 µL of dCTP, dGTP, and dTTP mix
340 at 10 mM each; and 10 µL of 5 U/µL DNA Polymerase I Large (Klenow)
341 Fragment (NEB, M0210). The reactions were then rotated at 37°C for 45 min.
342 948 µL of ligation master mix was then added: 150 µL of 10× NEB T4 DNA
343 ligase buffer with 10 mM ATP (NEB, B0202), 125 µL of 10% Triton X-100, 3 µL
344 of 50 mg/mL BSA (Thermo Fisher, AM2616), 10 µL of 400 U/µL T4 DNA Ligase
345 (NEB, M0202), and 660 µL of water. The reactions were then rotated at 16 °C
346 for 4 h and room temperature for 1 h. 45 µL of 10% SDS and 55 µL of 20 mg/ml
347 Proteinase K were added for crosslinking reversal. Incubate at 55°C for at
348 least 2 hours (overnight recommended). DNA was purified by
349 phenol:chloroform:isoamyl alcohol (25:24:1) extraction. Purified DNA in
350 solution was transferred into a 1.5 ml tube and sonicated to 400 bp on a
351 Covaris sonicator. Biotin labelled DNA was pulled-down on Streptavidin
352 Dynabeads (NEB, S1420S). After DNA repair and 3' A addition, Full Y-Adaptor

(Supplementary Table 1) was added. DNA -on Dynabeads was resuspended in 100 µl of 0.8x PCR Buffer and incubated at 98°C for 10 min before being cooled off in ice water. The supernatant was recovered, quantified, and used for direct sequencing on the Illumina HiSeq X10 platform (PE 2x150 bp reads). SAFE Hi-C on human K562 cells was carried out similarly with the reagents reduced in proportion to the estimated chromatin contents, not the cell number.

359

360 Data processing

We chose dm3 and hg19 as the reference genomes to align sequenced reads. Mapping, filtration, duplication removal, construction and normalization of contact matrices and basic library statistics of reads from all experiments were processed using the Juicer pipeline²⁶. For SAFE Hi-C, the analysis should only remove optical duplicates, which is caused by sequencing when a single cluster of reads is part of two adjacent tiles' on the same slide and used to compute two read calls separately. We used a modified AWK script derived from juicer's dups.awk script which removes duplicates and judges the source of duplicates (PCR or optical) to remove only optical duplicates. We set 1 as map quality threshold and all downstream analyses were based on KR normalized matrices²³ which ensures that each row and column of the contact matrix sums to the same value.

373

374 Analysis of PCR duplication rates for *Drosophila* S2 libraries

375 For duplication depth analysis, we used a modified AWK script derived from
376 juicer's dups.awk script to count the duplicate number of each duplicated
377 contact. We also tested different wobble number (0, 1, 2, and 3) to process the
378 deduplication step for each library. We show that SAFE Hi-C libraries have
379 high library complexity in standard deduplication (wobble=4) process and even
380 higher when set wobble to 0.

381

382 **Library complexity estimation for *Drosophila* S2 libraries**

383 Estimation of library complexity has been described before^{5,27}. For SAFE Hi-C
384 libraries, we computed the "PCR duplication rate" (although this library does
385 not contain real PCR duplicates) to estimate library complexity.

386

387 **Topologically associating domain identification**

388 Identification of TADs in *Drosophila* S2 cells was processed by Juicer
389 Arrowhead algorithm at 5kb resolution. For the identification of TADs in human
390 K562 cells, border strength index was calculated at 25kb resolution with a
391 moving block size of 8 bins. TAD borders of human K562 dataset were defined
392 by calling peaks through R package pracma. We used bedtools²⁸ intersect
393 command to call overlapped TADs and the overlapped region between two
394 overlapped TADs should span at least 90 percent of each TAD range
395 (command: bedtools intersect -f 0.9 -F 0.9 -sortout -a \$tad_1 -b \$tad_2)

396

397 **Compartment analysis**

398 Method of compartment analysis has been described before⁶. We used the
399 Pearson's and eigenvector command of Juicer tools to obtain the Pearson's
400 correlation matrix and eigenvector at 10 kb resolution.

401

402 **Long-range chromatin interaction calling and aggregate peak analysis**

403 Loops were identified at 5 kb and 10 kb resolution respectively using Juicer's
404 HiCCUPs algorithm²⁶ (parameters: -m 2048 -r 5000,10000 -k KR
405 --ignore_sparsity). P2LL (Peak to Lower Left) and ZscoreLL (Zscore Lower Left)
406 are defined to measure the enrichment of HiCCUPs peaks during aggregate
407 peak analysis (APA). P2LL is the ratio of the central pixel to the mean of the
408 pixels in the lower left corner. ZscoreLL stands for the Z-score of the central
409 pixel relative to all of the pixels in the lower left corner. Note that, for P2LL
410 scatter plot (fig*), the P2LL value is the ratio of each peaks' pixel to the expect
411 value of lower left.

412

413 **Data availability**

414 *Drosophila* S2 raw Illumina reads are available at
415 <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA470784>.

416 Accession code: PRJNA470784

417 Human K562 raw Illumina reads are available at

418 <https://www.ncbi.nlm.nih.gov/bioproject/524051>

419 Accession code: PRJNA524051

420

421 Code availability

422 All custom codes are available at

423 https://github.com/shenscore/Safe_Hi-C_script.

424

425 Acknowledgments

426 We gratefully acknowledge financial support from the National Key R&D
427 Program of China (2018YFC1004500), National Natural Science Foundation
428 of China (31571347 to C.H., 31771430 to L. L. and 31671409 to Y.T.),
429 Guangdong Science and Technology Department (2016A030313642 to C.H.),
430 Shenzhen Science and Technology Innovation Commission
431 (JCYJ20150529152146478 to C.H., JCYJ20170307105005654 to Y.T.),
432 Huazhong Agricultural University Scientific & Technological Self-innovation
433 Foundation (to L. L.) and the Thousand Talent Youth Program (to C.H.). We
434 thank Dr. Victor G. Corces for helpful advices on manuscript preparation and
435 Dr. Edwin Cheung for text editing.

436

437 Author contributions

438 C.H. and L.N. conceived the study and designed the experiments; L.N.
439 performed the experiments with help from Y.Z., Y.H., J.S., D.J., M. Y. and Y.T.;
440 W.S., N.H. and J. W. carried out the data analysis; L. L. supervised the data

441 analysis; C.H. supervised the study; C.H. wrote the manuscript with input from
442 all authors.

443

444 **Competing interests**

445 The authors declare no competing financial interests.

446

447 **References**

448

449 1 Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range
450 interactions reveals folding principles of the human genome. *Science*
451 **326**, 289-293, doi:10.1126/science.1181369 (2009).

452 2 Sexton, T. *et al.* Three-dimensional folding and functional organization
453 principles of the Drosophila genome. *Cell* **148**, 458-472,
454 doi:10.1016/j.cell.2012.01.010 (2012).

455 3 Hou, C., Li, L., Qin, Z. S. & Corces, V. G. Gene density, transcription,
456 and insulators contribute to the partition of the Drosophila genome into
457 physical domains. *Mol Cell* **48**, 471-484,
458 doi:10.1016/j.molcel.2012.08.031 (2012).

459 4 Dixon, J. R. *et al.* Topological domains in mammalian genomes
460 identified by analysis of chromatin interactions. *Nature* **485**, 376-380,
461 doi:10.1038/nature11082 (2012).

462 5 Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution

- 463 reveals principles of chromatin looping. *Cell* **159**, 1665-1680,
464 doi:10.1016/j.cell.2014.11.021 (2014).
- 465 6 Sati, S. & Cavalli, G. Chromosome conformation capture technologies
466 and their impact in understanding genome function. *Chromosoma* **126**,
467 33-44, doi:10.1007/s00412-016-0593-6 (2017).
- 468 7 Lin, D. *et al.* Digestion-ligation-only Hi-C is an efficient and
469 cost-effective method for chromosome conformation capture. *Nat Genet*,
470 doi:10.1038/s41588-018-0111-2 (2018).
- 471 8 Li, T., Jia, L., Cao, Y., Chen, Q. & Li, C. OCEAN-C: mapping hubs of
472 open chromatin interactions across the genome reveals gene regulatory
473 networks. *Genome Biol* **19**, 54, doi:10.1186/s13059-018-1430-4 (2018).
- 474 9 Ramani, V. *et al.* Massively multiplex single-cell Hi-C. *Nat Methods* **14**,
475 263-266, doi:10.1038/nmeth.4155 (2017).
- 476 10 Liang, Z. *et al.* BL-Hi-C is an efficient and sensitive approach for
477 capturing structural and regulatory chromatin interactions. *Nat Commun*
478 **8**, 1622, doi:10.1038/s41467-017-01754-3 (2017).
- 479 11 Flyamer, I. M. *et al.* Single-nucleus Hi-C reveals unique chromatin
480 reorganization at oocyte-to-zygote transition. *Nature* **544**, 110-114,
481 doi:10.1038/nature21711 (2017).
- 482 12 Belaghzal, H., Dekker, J. & Gibcus, J. H. Hi-C 2.0: An optimized Hi-C
483 procedure for high-resolution genome-wide mapping of chromosome
484 conformation. *Methods* **123**, 56-65, doi:10.1016/j.ymeth.2017.04.004

- 485 (2017).
- 486 13 Ramani, V. *et al.* Mapping 3D genome architecture through in situ
487 DNase Hi-C. *Nat Protoc* **11**, 2104-2121, doi:10.1038/nprot.2016.126
488 (2016).
- 489 14 Mumbach, M. R. *et al.* HiChIP: efficient and sensitive analysis of
490 protein-directed genome architecture. *Nat Methods* **13**, 919-922,
491 doi:10.1038/nmeth.3999 (2016).
- 492 15 Hsieh, T. S., Fudenberg, G., Goloborodko, A. & Rando, O. J. Micro-C
493 XL: assaying chromosome conformation from the nucleosome to the
494 entire genome. *Nat Methods* **13**, 1009-1011, doi:10.1038/nmeth.4025
495 (2016).
- 496 16 Mifsud, B. *et al.* Mapping long-range promoter contacts in human cells
497 with high-resolution capture Hi-C. *Nat Genet* **47**, 598-606,
498 doi:10.1038/ng.3286 (2015).
- 499 17 Hsieh, T. H. *et al.* Mapping Nucleosome Resolution Chromosome
500 Folding in Yeast by Micro-C. *Cell* **162**, 108-119,
501 doi:10.1016/j.cell.2015.05.048 (2015).
- 502 18 Nagano, T. *et al.* Single-cell Hi-C reveals cell-to-cell variability in
503 chromosome structure. *Nature* **502**, 59-64, doi:10.1038/nature12593
504 (2013).
- 505 19 Yaffe, E. & Tanay, A. Probabilistic modeling of Hi-C contact maps
506 eliminates systematic biases to characterize global chromosomal

- 507 architecture. *Nat Genet* **43**, 1059-1065, doi:10.1038/ng.947 (2011).
- 508 20 Hu, M. *et al.* HiCNorm: removing biases in Hi-C data via Poisson
509 regression. *Bioinformatics* **28**, 3131-3133,
510 doi:10.1093/bioinformatics/bts570 (2012).
- 511 21 Imakaev, M. *et al.* Iterative correction of Hi-C data reveals hallmarks of
512 chromosome organization. *Nat Methods* **9**, 999-1003,
513 doi:10.1038/nmeth.2148 (2012).
- 514 22 Kozarewa, I. *et al.* Amplification-free Illumina sequencing-library
515 preparation facilitates improved mapping and assembly of
516 (G+C)-biased genomes. *Nat Methods* **6**, 291-295,
517 doi:10.1038/nmeth.1311 (2009).
- 518 23 Knight, P. A. & Ruiz, D. A fast algorithm for matrix balancing. *Ima J*
519 *Numer Anal* **33**, 1029-1047, doi:10.1093/imanum/drs019 (2013).
- 520 24 Wang, Q., Sun, Q., Czajkowsky, D. M. & Shao, Z. Sub-kb Hi-C in *D.*
521 *melanogaster* reveals conserved characteristics of TADs between
522 insect and mammalian cells. *Nat Commun* **9**, 188,
523 doi:10.1038/s41467-017-02526-9 (2018).
- 524 25 Nagano, T. *et al.* Comparison of Hi-C results using in-solution versus
525 in-nucleus ligation. *Genome Biol* **16**, 175,
526 doi:10.1186/s13059-015-0753-7 (2015).
- 527 26 Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing
528 Loop-Resolution Hi-C Experiments. *Cell Syst* **3**, 95-98,

529 doi:10.1016/j.cels.2016.07.002 (2016).

530 27 Lander, E. S. & Waterman, M. S. Genomic mapping by fingerprinting

531 random clones: a mathematical analysis. *Genomics* **2**, 231-239 (1988).

532 28 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for

533 comparing genomic features. *Bioinformatics* **26**, 841-842,

534 doi:10.1093/bioinformatics/btq033 (2010).

535

536 **Legends:**

537

538 **Fig. 1** Procedures of SAFE Hi-C and traditional *in situ* Hi-C.

539 **a** Side-by-side comparison of SAFE Hi-C and *in situ* Hi-C procedures. The

540 black text shows shared steps in both methods, blue and red texts correspond

541 steps specific for SAFE Hi-C and *in situ* Hi-C, respectively.

542 **b** In the preparation of both *in situ* Hi-C and SAFE Hi-C libraries, partially

543 complementary adapters with a 3' thymine (T) overhang are ligated onto

544 repaired and 3' adenine (A)-tailed DNA captured on streptavidin beads. The

545 sequence of adaptors is listed in Supplementary Table 1.

546

547 **Fig. 2** Amplification increases PCR duplicates and reduces Hi-C library

548 complexity of the *Drosophila* genome.

549 **a** Cis paired reads are uniquely mapped on the same chromosome, and trans

550 paired reads are mapped on different chromosomes. SAFE Hi-C is referred to

551 as 0 PCR cycle in figures.

552 **b** Percentage of unique paired reads and duplicates that can be aligned out of
553 the total sequenced paired reads. The light blue bar shows the PCR duplicates
554 in SAFE Hi-C libraries, which we kept as unique for no amplification was
555 involved in the library preparation process.

556 **c** Frequencies of duplicate ligates introduced by PCR amplification.

557 **d** Accumulated percentage of unique paired reads against the percentage of
558 sequencing depth.

559 **e** Library complexity estimates. For SAFE Hi-C, the complexity was estimated
560 as described in Methods.

561

562 **Fig. 3** Distance-related amplification bias and TADs identification in *Drosophila*
563 genome.

564 **a** Chromatin interaction frequency as a function of genomic distance averaged
565 across the genome for the *Drosophila* genome.

566 **b** Average chromatin interaction frequency across the genome normalized
567 against SAFE Hi-C for *Drosophila* genome.

568 **c** Hi-C interaction heatmap for the region of chromosome 2L from 6 Mb to 8 Mb
569 is shown. Border index values for SAFE Hi-C and amplified Hi-C libraries are
570 compared.

571 **d** Dark blue bars show the number of TADs shared with SAFE Hi-C, red bars
572 show the number of TADs uniquely found for amplified Hi-C libraries.

573 **e** Chromatin interaction frequency as a function of genomic distance averaged
574 within the TADs.

575 **f** Average chromatin interaction frequency within the TADs normalized against
576 SAFE Hi-C.

577

578 **Fig. 4** SAFE Hi-C on 250 thousand human K562 cells.

579 **a** Comparison of chromatin interaction frequency against the genomic distance
580 between SAFE Hi-C (blue line) and *in situ* Hi-C (red line). s^{-1} (black dashed line)
581 and $s^{-0.5}$ (red dotted line) represent the predicted fractal globule and mitotic
582 states, respectively. The turning point of chromatin interactions on the
583 decaying curves for SAFE Hi-C and *in situ* Hi-C is indicated by an arrowhead.

584 **b** Average chromatin interaction frequency of *in situ* Hi-C normalized against
585 SAFE Hi-C across the genome. The crossing point is shown by an arrowhead.

586 **c** Venn diagram shows the overlap between TADs identified for SAFE Hi-C
587 and *in situ* Hi-C.

588 **d** Hi-C interaction heatmap and border index comparison between SAFE Hi-C
589 and *in situ* Hi-C for a region of chromosome 1 from 40 Mb to 45 Mb.

590 **e** Borders are aligned for all, shared and specific TADs. The chromatin contact
591 frequency in flanking TADs ($A+A'$) is divided by that of the inter-TADs (B), the
592 ratios of $(A+A')/B$ for aggregated borders are shown at the bottom left in each
593 heatmap.

594

595

596 **Legends for Supplementary Figures and Tables:**

597

598 **Supplementary Fig. 1** Size distribution of ligation products from Hi-Cs on

599 *Drosophila* S2. Size was calculated based on the distance of each end of a

600 paired read to its closest DpnII sequence at the genomic site.

601

602 **Supplementary Fig. 2** Pearson correlation between biological duplicates and

603 between SAFE Hi-C and *in situ* Hi-C libraries with the different number of

604 amplification cycles for *Drosophila* S2.

605

606 **Supplementary Fig. 3** Amplification effect on paired reads of different GC

607 content. Paired reads with GC content from 20%-80% are equally divided into

608 100 equal bins.

609 **(a)** Ratio of paired reads as a function of GC content for SAFE Hi-C and *in situ*

610 Hi-C libraries with the different number of amplification cycles. Ratio is the

611 number of paired reads of each specific bin over the number of total paired

612 reads.

613 **(b)** Ratio of paired reads was normalized against SAFE Hi-C.

614

615 **Supplementary Fig. 4** Percentage of unique paired reads and duplicates of

616 the total alignable paired reads. The light blue bar shows the “suspected PCR

617 duplicates” that were kept as unique.

618

619 **Supplementary Fig. 5** 5 kb heatmaps for each chromosome arm of

620 *Drosophila* S2 genome.

621

622 **Supplementary Fig. 6** 10 kb resolution compartments for each chromosome

623 arm of *Drosophila* S2 genome.

624

625 **Supplementary Fig. 7** Pearson correlation analysis of compartment

626 eigenvalues from SAFE Hi-C and amplified Hi-Cs for *Drosophila* S2 cells.

627

628 **Supplementary Fig. 8** Pearson correlation analysis of compartment

629 eigenvalues of SAFE Hi-C and *in situ* Hi-C on human K562 cells.

630

631 **Supplementary Fig. 9** Examples of long-range chromatin interaction loops

632 identified on Chr 2L between 10.2 – 11.0 Mb for the *Drosophila* genome by

633 SAFE Hi-C and amplified Hi-Cs. Loops identified for SAFE Hi-C and for

634 amplified Hi-Cs are marked by blue and green circles, respectively. Short

635 vertical lines above heatmaps indicate CTCF binding sites. *Drosophila* S2

636 CTCF binding sites are downloaded from modENCODE.

637

638 **Supplementary Fig. 10** Comparison of long-range chromatin interactions

639 loops identified for the *Drosophila* genome by SAFE Hi-C and amplified
640 libraries.

641 (a) Number of long-range chromatin interaction loops. Dark blue bars show the
642 number of loops shared with SAFE Hi-C, red bars show the number of loops
643 uniquely found for amplified Hi-C libraries.

644 (b) Average signal-to-noise values. P2LL is calculated as described in
645 Methods.

646 (c) Average ZscoreLL for identified chromatin loops. ZscoreLL is calculated as
647 described in Methods.

648 (d) Correlation analysis of P2LL values for loops shared with SAFE Hi-C
649 (upper panel) and for loops found for SAFE Hi-C but lost after amplification
650 (lower panel). Red dots show loops with higher P2LL value and blue ones with
651 lower values after amplification compared to SAFE Hi-C.

652

653 **Supplementary Table 1** List of adaptors and primer sequences.

654

655 **Supplementary Table 2** Libraries sequencing statistics for *Drosophila* S2

656 cells.

657

658 **Supplementary Table 3** SAFE Hi-C and *in situ* Hi-C sequencing statistics for

659 human K562 cells.

Figure 1-Niu et al.

Figure 2-Niu et al.

bioRxiv preprint doi: <https://doi.org/10.1101/562280>; this version posted February 27, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 4-Niu et al.